VALIDACIÓN CRUZADA DE PRUEBAS PSICOMÉTRICAS

Nelson Rodríguez Trujillo Ph.D.
www.psicoconsult.com
psicoconsult@cantv.net
Escuela de Psicología, Universidad Central de Venezuela
Presentado ante el Congreso Interamericano de Medición y Evaluación Psicológica.
Caracas, Venezuela, 1999

Introducción.

En la construcción de instrumentos psicométricos es usual que, después de definir la variable y establecer las especificaciones del test, se utilice el siguiente procedimiento:

1. Desarrollo de los reactivos (ítems).
2. Administración del instrumento experimental a una muestra de sujetos.
3. Realización de análisis de ítems y selección de los mejores.
4. Administración del instrumento así obtenido a una nueva muestra de sujetos.
5. Determinación de la confiabilidad (sea de consistencia interna, test retest o formas paralelas).
6. Determinación de la ecuación de regresión y del coeficiente de validez, si el instrumento va a ser utilizado para predecir algún criterio.

Las muestras de sujetos de los pasos 2 y 4 (que de ahora en adelante se llamarán muestras normativas), constituyen en realidad algo circunstancial, puesto que la población a la cual va dirigida el instrumento es usualmente mucho más amplia, por lo que, cuando los indicadores psicométricos del test son aceptables en la muestra normativa, el instrumento se considera listo para su uso en otras muestras de la población, bajo el supuesto de que el comportamiento de los ítems y del test como totalidad serán iguales en esas nuevas muestras. Este supuesto, sin embargo, no siempre está garantizado, y puede haber fluctuaciones en los estadísticos, de magnitud y dirección desconocidas, de una muestra a otra.

Esta situación puede llegar a ser crítica cuando se hace uso de muestras pequeñas, o cuando se tiene una sola muestra normativa para las cuatro actividades: selección de los ítems, cálculo de los estadísticos de consistencia interna, determinación de coeficientes de validez y determinación de las ecuaciones de regresión. Es por ello, que muchos textos de psicometría (Magnusson, 1966, Nunnally, 1987, Anastasi y Urbina, 1998*)* sugieren la realización de estudios de validación cruzada, para conocer la sensibilidad de los indicadores a los errores muestrales y establecer la magnitud de las posibles fluctuaciones. En este artículo se analizan los conceptos en los que se sustenta la validación cruzada y se presenta una estrategia para realizar este tipo de estudio de manera eficiente y económica.

Fundamento teórico de la Validación Cruzada.

Los métodos correlacionales que se utilizan para determinar las características psicométricas de un instrumento, se fundamentan en el principio de los mínimos cuadrados, que consiste en lograr que la suma de los desvíos al cuadrado entre los valores sea un mínimo **en esa muestra normativa.** Esto capitaliza, tanto en las tendencias existentes entre las variables, como **en los errores de muestreo**, por lo que, al utilizar el instrumento en muestras diferentes, se puede esperar una reducción de la magnitud de la correlación, debido a que los errores muestrales y otras variaciones al azar, difieren en la nueva muestra con respecto a la muestra normativa. Esto es cierto para los coeficientes biserial y punto biserial con que se evalúa el funcionamiento de los ítems, para las ecuaciones de regresión y coeficientes de correlación con criterios externos, y para los coeficientes de confiabilidad de consistencia interna.

En la selección de ítems esto se hace más crítico, ya que cada ítem constituye una muestra muy pequeña de conducta, lo que hace que los resultados sean inconfiables, además de que seleccionamos aquellos que muestran una correlación significativamente diferente de cero por encima de un valor especificado previamente. Esos ítems son precisamente los que tienden a tener desviaciones positivas debidas al azar, lo que nos llevará

a contar con ítems que correlacionarán más bajo en muestras sucesivas y habremos ignorado algunos que, también por azar, correlacionaron bajo en la muestra normativa.

Por ello es recomendable someter los indicadores psicométricos a estudios de validación cruzada, es decir, probar su comportamiento en muestras sucesivas e independientes. Es conveniente destacar, que el término Validación Cruzada ha sido utilizado para referirse a aspectos diferentes; las referencias a continuación permiten aclarar algunos de esos usos:

1. Mosier (1951) plantea que existen al menos tres contextos para la validación cruzada: A) Para la determinación de la estabilidad de lo sindicadores psicométricos **en muestras sucesivas de la misma población**. B) Para la "Generalización de la validez", que se utiliza cuando se trata de muestras de diferentes poblaciones. C) Para la "Extensión de la validez" que se refiere del uso de diferentes criterios a predecir con el mismo instrumento. Estos dos últimos casos no se tratan en este artículo.
2. Mosteller y Tukey (1968) hacen la diferencia entre "forma" y "procedimiento" cuando se establece una ecuación de regresión. "Forma" se refiere a la selección de variables que van a entrar en una ecuación, lo que se hace luego de probar muchos predictores y descartar los menos prometedores. "Procedimiento" se reserva para la ecuación de regresión específica, con valores beta (ponderaciones) determinados. Mosteller y Tukey recomiendan utilizar una primera muestra para la selección de la "forma", una segunda para establecer el "procedimiento" y una tercera para evaluar el comportamiento de la ecuación cuando se desea determinar hasta qué punto se pueden esperar fluctuaciones en muestras sucesivas de la misma población.

Para realizar estudios de validación cruzada es necesario contar con varias muestras, pero, cuando se dispone de escasos recursos, o de poco tiempo para desarrollar instrumentos (situación por demás común en investigación psicológica), lo usual es que se utilice una misma muestra para seleccionar los ítems y establecer todos los indicadores del test. Esta solución puede ser riesgosa, como se mencionó anteriormente, ya que hay incertidumbre con relación al comportamiento del instrumento en nuevas muestras de sujetos.

Supongamos que es necesario desarrollar instrumentos para predecir el rendimiento académico pero por limitaciones de tiempo y recursos no es posible realizar varios estudios piloto, ni obtener varios grupos normativos para probar el funcionamiento de los ítems y establecer la estabilidad de los coeficientes de confiabilidad y validez. La estrategia que se propone a continuación está fundamentada en proposiciones de Katzell (1951), Magnusson (1966) y Henrissen (1971), y ampliada para atender a algunos aspectos de la validación cruzada.

Validación cruzada de los ítems.

Katzell (1951) propone utilizar una sola muestra normativa para la selección de los ítems pero tratarla como que si fuesen dos muestras, mientras que Magnusson (1966) y Henrissen (1971) proponen seleccionar los ítems de acuerdo a criterios internos y externos. Nuestra proposición sigue el siguiente proceso:

1. Tome un número relativamente grande de sujetos, preferiblemente 200 o más. A cada sujeto se le aplica la prueba piloto y se obtiene un criterio externo de validación (promedio de notas académicas, por ejemplo).
2. Divida al azar la muestra en dos grupos iguales, llamados ahora A y B.
3. Haga análisis de ítems por separado para cada grupo.
4. Determine el nivel de significación estadística en el cual se considera que la correlación difiere de cero. Katzell recomienda utilizar niveles de significación generosos (p=0,1), permitiendo así la inclusión de ítems con fluctuaciones pequeñas alrededor de cero.
5. Para cada grupo por separado, seleccione los ítems que sobrepasen el nivel de significación establecido.
6. Al final del paso anterior, hay dos claves de corrección, una para A y otra para B. Probablemente los ítems que aparecen en una y otra clave de corrección no sean los mismos.
7. Corrija con cada clave las pruebas **del otro grupo**.
8. Con los puntajes obtenidos, compute las correlaciones de orden cero entre el test y el criterio.
9. Para determinar los ítems que se seleccionarán para el instrumento final, seleccione aquellos ítems cuya probabilidad compuesta sea de 0,05 o menos de ser diferente de cero. El nivel de significación se determina con la fórmula (1) con df = 4.

$$(1)\ X^{2} = -4.605 \log p_1 p_2$$

Las ventajas de este método son varias:

1) Se utilizan todos los casos para el análisis de ítems y para la validación cruzada, y ésta se obtiene inmediatamente y con un solo grupo.
2) Se obtiene un mayor número de ítems, es decir, se rechazan menos ítems debido a las fluctuaciones al azar, lo que resultará en un conjunto de ítems con características más estables de muestra en muestra.
3) Se obtienen dos estimados de la validez del test, que pueden compararse estadísticamente si la corrección de los exámenes se hace usando la clave final que incluirá ítems tomados de ambas claves de corrección, en lugar de la clave de cada grupo.

Existe la desventaja de que los estadísticos en cada grupo probablemente son menos estables que para la muestra total, ya que se fundamentan en grupos más pequeños, pero las ventajas superan con creces esta desventaja, sobre todo si la muestra total es mayor de 200 sujetos. De hecho, aquí se aprovecha uno de los principios en que se fundamenta la validación cruzada y es que la confianza que puede tenerse en un resultado es mayor, cuando se obtiene una confirmación independiente del mismo: Es más confiable obtener resultados congruentes en varias muestras pequeñas e independientes, que en un solo experimento con una muestra grande, aún cuando el número total de sujetos sea igual en ambos casos, ya que en el primer caso, el nivel de significación puede ser dividido para el conjunto de hipótesis.

El segundo elemento que compone esta estrategia, se refiere a que en la construcción de un test que va a ser utilizado para predecir un criterio externo, se puede dar énfasis a maximizar la confiabilidad o la validez (Magnusson, 1966, Henryssen, 1971). En el primer caso, se seleccionarán ítems con altas intercorrelaciones entre sí, es decir, que correlacionen alto con el puntaje total en el test (criterio interno). En el segundo caso, se seleccionarán ítems que correlacionen alto con el criterio externo que se desea predecir.

En ambos casos se corre un riesgo: la maximización de la confiabilidad puede producir un test muy homogéneo, que tiende a evaluar una sola variable, pero que tiene poca relación con un criterio externo complejo o heterogéneo. Por su lado, la selección de ítems en función del criterio externo, puede producir un test que correlaciona alto con el criterio que se desea predecir, pero de baja consistencia interna, lo que lo hace de difícil interpretación. Dado que al escoger ítems utilizamos como indicadores el nivel de dificultad y la correlación con un criterio, es como que si estuviésemos trabajando con “tests” de un solo ítem; aquí se corre el riesgo (sobre todo al correlacionar con el criterio externo) de que en muestras sucesivas, el test como totalidad tenga mucho menos validez que en la muestra inicial.

Una forma de evitar ambos riesgos, es seleccionar los ítems que correlacionan significativamente con los criterios interno y externo. La Figura 1 presenta gráficamente el método. La ordenada representa el índice de confiabilidad, la relación con la puntuación total en el test; la abscisa representa el índice de validez del ítem, la relación con el criterio externo.

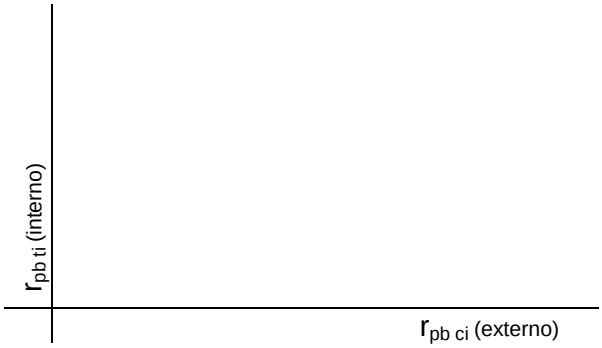

Figura 1. Representación gráfica de la relación entre correlación interna y externa.

Las ecuaciones (2) y (3) permiten determinar ambos índices, en donde $r_{bis(ti)}$ es la correlación biserial y el criterio interno, $r_{bis(ci)}$ es la correlación biserial con el criterio externo y Y_i es la ordenada en la distribución normal en la que se divide el área bajo la curva en las proporciones p y 1-p.

$$(2)\ r_{bis(ti)}\ Y_i$$

$$(3)\ r_{bis(ci)}\ Y_i$$

El valor obtenido en (2) es una medida de la contribución del ítem a la varianza total, y por ello a la confiabilidad; la suma de estos valores para todos los ítems permite establecer la desviación típica del test total. El valor obtenido en (3) representa la contribución del ítem individual a la validez del test total. La ecuación (4) permite deducir, que el valor numérico del coeficiente de validez es la proporción entre la suma de los índices de validez de los ítems y la suma de los coeficientes de confiabilidad.

$$(4)\ r_{tg} = \frac{\sum r_{bi\ (ci)} Y_i}{\sum r_{bis\ (ti)} Y_i}$$

Como afirma Magnusson: “Para obtener, para un criterio determinado, la más alta validez posible en el test, debemos, obviamente, escoger ítems que tengan una alta correlación con el criterio y bajas correlaciones con los puntajes en el test. Un ítem cuyo índice de validez es igual a su índice de confiabilidad, mide en igual medida las variables del criterio y las variables del test, y contribuye así en igual forma a la validez del test y a su confiabilidad” (p. 216).

En el gráfico 1, se deberían seleccionar los ítems del nivel más alto y a la derecha del cuadrante. El punto de corte en ambos dimensiones dependerá del propósito del test y los requerimientos de validez o confiabilidad.

Validación Cruzada de la Ecuación de Regresión.

La ecuación de regresión consiste en las ponderaciones aplicadas a las variables predictoras para estimar el criterio. Para establecer esas ponderaciones es necesario tener, para una muestra particular, medidas tanto en el criterio como en las variables predictoras; una vez establecida la recta de regresión, es puede utilizar en muestras en las que tenemos solamente las variables predictoras. Lo usual es que las medidas del criterio resuman actividades de naturaleza compleja, que exigen procesos psicológicos y conductuales, que no necesariamente están intercorrelacionados, por lo que es recomendable utilizar más de un predictor. Una de las ventajas de utilizar predictores múltiples es, que si se escogen bien, hay mayor probabilidad de contar con dimensiones relevantes de la medida criterio. Una vez que se han seleccionado las variables y establecido la recta de regresión, es necesario realizar validación cruzada, para asegurar que el uso de esa ecuación está garantizada en muestras sucesivas.

La separación de la muestra en dos submuestras es también aplicable para este propósito. Se logra utilizando la ecuación de regresión de una de las muestras para predecir el rendimiento en la otra muestra; dado que se trata de muestras independientes, podemos suponer que los errores de muestreo serán diferentes y debemos esperar una reducción de la correlación.

Una forma de definir el coeficiente de correlación es calculando la correlación entre los puntajes obtenidos en una muestra y los puntajes predichos usando la ecuación de regresión. Si se administra un conjunto de predictores a dos muestras de una misma población, se obtendrán dos coeficientes de correlación y dos ecuaciones de regresión. Usando la ecuación de regresión de cada grupo para predecir los puntajes del otro grupo y calculando la correlación entre los puntajes predichos y obtenidos en cada grupo, podemos determinar el funcionamiento de esa combinación específica de los predictores en muestras similares. En esa forma, respondemos a la pregunta sobre qué tan efectiva va a ser nuestra predicción en otras muestras, siempre y cuando se mantengan las mismas condiciones de administración.

Una vez obtenidos los cuatro coeficientes de correlación, dos obtenidos con la ecuación del propio grupo y dos con la del grupo contrario, se puede determinar si existen diferencias significativas entre ellos por los métodos

convencionales de la transformación a Z de Fisher y su aproximación a la curva normal. El estadístico de prueba lo provee la ecuación (5).

$$(5)\ \frac{Z_1 - Z_2}{\sqrt{\frac{1}{(N-3)} + \frac{1}{(N-3)}}}$$

En donde Z_1 y Z_2 representan los valores transformados de los coeficientes de correlación de las dos muestras y N representa el número de observaciones.

Las pruebas usuales de comparación de los coeficientes de correlación incluyendo o excluyendo algunas variables, se evalúan mediante una distribución F, haciendo uso de la ecuación (6)

$$(6)\ \frac{R_1 - R \;/ m_1 - m}{1 - R_1 \;/ N - m_1 - 1}$$

Con $(m_1\text{-}m_2)$ y $(N\text{-}m_1\text{-}1)$ grados de libertad, en donde R_1 es el coeficientes de correlación con m_1 variables independientes y R_2 es el coeficiente de correlación con m_2 variables independientes, siendo m_2 menor e incluido en m_1, y N es el número total de observaciones. Si F es significativo a un nivel determinado, se concluye que la ganancia o pérdida en precisión, es significativa (McNemar, 1969, p. 321). Los métodos modernos de cálculo de coeficientes de correlación múltiple por pasos, permiten establecer fácilmente la contribución de cada variable a la ecuación de regresión.

Validación Cruzada de la confiabilidad.

La confiabilidad se define como la consistencia o la estabilidad con que un instrumento mide lo que sea que mida. Cronbach, Rajaratman y Gleser (1963) en su muy conocido artículo, plantean la relación expresada en la ecuación (7).

$$(7)\ r_{tt} = \frac{\sigma^2_T}{\sigma^2 t} = \frac{\sigma^2_T}{\sigma^2_t + \sigma^2_e} = 1 - \frac{\sigma^2_e}{\sigma^2_t}$$

En donde r_{tt} es el coeficiente de confiabilidad, σ^2_T es la varianza de los puntajes verdaderos, σ^2_t es la varianza de los puntajes obtenidos y σ^2_e es la varianza de los puntajes de error. Esta relación define la confiabilidad como la proporción de la varianza total que se debe a varianza de error y la proporción que se debe a las diferencias individuales.

El concepto de confiabilidad también se define en función del método empleado para estimarla. Test-Retest es una correlación producto momento de Pearson entre dos conjuntos de puntajes obtenidos en dos oportunidades diferentes. División por mitades se calcula con base a la correlación entre las puntuaciones de los sujetos en cada una de las mitades definidas (pares e impares por ejemplo) que se corrige con la fórmula de profecía de Spearman Brown para el test como totalidad. Formas paralelas es igualmente una correlación entre dos conjuntos de puntajes obtenidos por los sujetos en dos tests en dos aplicaciones diferentes. En estos tres casos, la decisión sobre la diferencia entre dos coeficientes de confiabilidad, uno obtenido en una muestra normativa y otro en una muestra de replicación, se realiza mediante los métodos aplicables a la correlación producto momento de Pearson.

En el caso de la consistencia interna u homogeneidad (Kuder-Richardson, Alfa de Cronbach, Anova de Hoyt), el problema es diferente, ya que conceptualmente estos coeficientes no constituyen una correlación propiamente dicha. Feldt (1965, 1969) calculó la distribución de muestreo de KR_{20} y desarrolló fórmulas para el cómputo de intervalos de confianza y pruebas para la diferencia entre dos coeficientes.

Feldt (1969) demostró que la distribución de muestreo de KR_{20} se aproxima a una F con (N-1) y (N-1)(K-1) grados de libertad, en donde N es el número de sujetos y K es el número de ítems del test. Los intervalos de confianza pueden establecerse con la fórmula (8).

$$(8)\ p\ [1- (1-r_{tt})\ F_{(1-1/2)(dfs,dfe)} < \rho_{20} < 1 - (1- rtt)/(F_{(1-1/2)(dfe,dfs)})] = 1 - \alpha$$

En donde dfs = N-1 y dfe = (N-1)(K-1), r_{tt} es el coeficiente obtenido de confiablidad, ρ_{20} es el coeficiente de confiabilidad teórico y F son los valores teóricos superiores e inferiores que establecen los límites con un intervalo de $\alpha/2$.

Si dos coeficientes de confiabilidad se obtienen del mismo instrumento o instrumentos paralelos en muestras de la misma población, la hipótesis de igualdad se pueden evaluar mediante la fórmula (9) (Feldt, 1969).

$$(9) W = (F_A)(F_B) = \frac{1-r_A}{1-r_B}$$

En donde los subscriptos indican el grupo de donde proviene el coeficiente de confiabilidad. Feldt demostró que bajo la hipótesis nula, W se distribuye aproximadamente como una F central con (N_A-1) y (N_B-1) grados de libertad. Si W excede el valor de F preestablecido al nivel de significación, los dos coeficientes de confiabilidad son diferentes.

Una aplicación práctica.

Los datos de un estudio realizado anteriormente (Rodríguez Trujillo, 1972) sirvieron para realizar una aplicación práctica de la metodología descrita. La Figura 2 muestra el flujograma del proceso utilizado para llevar a cabo el estudio. Se trató de una muestra de 349 sujetos a los cuales se les aplicaron pruebas de Razonamiento Verbal y Razonamiento Numérico al ingresar a la Universidad de Oriente en 1970. Se contaba también con el promedio de notas de educación secundaria y las notas del primer semestre, que culminaron ese año. Se siguieron los pasos descritos a continuación:

1. La muestra total fue dividida al azar en dos grupos, A y B de 175 y 174 sujetos de acuerdo a tres variables de estratificación: carrera escogida, sexo y si habían seleccionado biología o no. Al evaluar los estadísticos de las variables predictoras y las notas universitarias, ambos grupos resultaron perfectamente equivalentes.
2. Para cada grupo se realizaron dos análisis de ítems, uno correlacionado con el criterio interno y otro con el criterio externo (promedio de notas universitarias).
3. Se seleccionaron los ítems que mostrasen una alta correlación con ambos criterios. Al aplicar la fórmula (1) se determinó que el mínimo nivel de correlación biserial significativamente diferente de 0 para el tamaño de las muestras era de 0,122. Se descartaron los ítems que tuviesen problemas de equiatracción o de discriminación en más de dos opciones. Esto redujo el número de ítems de 91 a 45 para la prueba verbal y de 50 a 20 para la prueba numérica.
4. Con los ítems seleccionados se elaboró una nueva clave de corrección que permitió puntuar nuevamente las pruebas de los dos grupos.
5. Se establecieron estimados de confiabilidad para cada grupo y se determinó la significación de la diferencia con un $\alpha = 0.01$ haciendo uso de la fórmula (9).
6. Se utilizaron las nuevas puntuaciones en los dos tests y el promedio de notas de educación secundaria como variables predictoras y como criterio externo las notas universitarias, para establecer, para cada grupo, los coeficientes correlación y las ecuaciones de regresión múltiple con el promedio de notas universitarias. Se establecieron los niveles de significación para las diferencias.
7. Se utilizaron las ecuaciones de regresión de cada grupo, para calcular las puntuaciones predichas del otro grupo.
8. Finalmente se estableció la correlación entre los puntajes predichos y los puntajes obtenidos, aplicando luego la diferencia entre los coeficientes correlación obtenidos originalmente y los obtenidos con la ecuación de regresión del otro grupo. Se evaluaron las diferencias al nivel del $\alpha = 0.01$.

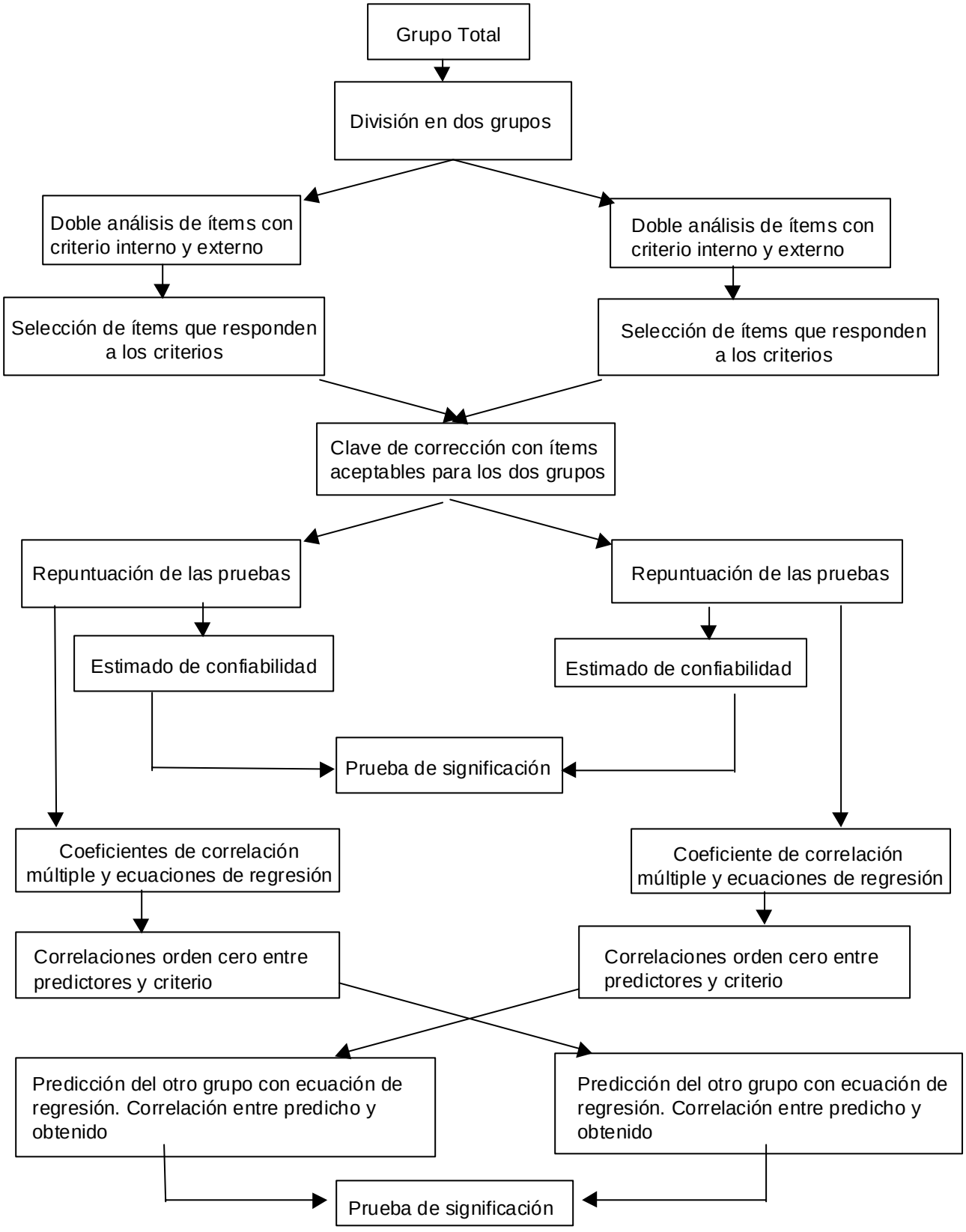

FIGURA 1. FLUJOGRAMA DEL PROCESO DE ANÁLISIS DE DATOS

Los resultados se presentan a continuación. La tabla 1 indica las correlaciones biseriales más altas, más bajas y medianas obtenidas por los ítems con el criterio externo. Puede verse que algunas de ellas son altas. Un resultado interesante son las grandes fluctuaciones de un grupo a otro de los estadísticos de los ítems. En particular en las correlaciones con el criterio externo, se producen amplias fluctuaciones en los coeficientes de un mismo ítems de un grupo a otro, por ejemplo un ítem mostró una correlación de 0,566 en el grupo B, pero de sólo 0,193 en el grupo A. Esto significa, que si no se hubiese contado con los análisis de ítems de los dos grupos, se habría aceptado, o descartado conjuntos de ítems completamente diferentes. Esto corrobora la necesidad de realizar validación cruzada para la selección de los ítems.

Tabla 1. Correlaciones biseriales de los ítems con el criterio externo

	Grupo A			Grupo B		
Subtest	alta	baja	Mediana	Alta	baja	Mediana
Verbal	0,365	-0,278	0,148	0,566	-0,143	0,142
Numérica	0,410	-0,353	0,152	0,466	-0,152	0,143

En la Tabla 2 se presentan los resultados de los coeficientes de confiabilidad de Hoyt, junto con el cálculo de W. Se puede observar que no hay diferencias significativas entre los coeficientes de confiabilidad de los dos grupos. En este caso, la razón por la cual no hay diferencias significativas, está relacionada con la mayor estabilidad de una muestra mayor de conducta que el ítem independiente. Por otro lado, la muestra de ítems de la prueba verbal son 45 ítems y la de numérico de 20. Dada la relación entre longitud del test y confiabilidad, se aplicaron las fórmulas para estimar la confiabilidad al incluir un mayor número de ítems, con 60 y 45 respectivamente para la prueba verbal y la numérica y se obtuvieron coeficientes de confiabilidad de 0,87 y de 0,83 que son aceptables para este tipo de pruebas.

Tabla 2. Coeficientes de confiabilidad de Hoyt.

Prueba y Número ítems	Grupo A		Grupo B		
	Confiab.	Error medida	Confiab.	Error medida	W
Verbal 45	0,837	2,95	0,827	2,95	1.063 (n.s.)
Numérico 20	0,686	1,91	0,637	1,92	1,154 (n.s.)

En la Tabla 3 se presentan los intervalos de confianza para los coeficientes de confiabilidad, lo que permite comparar coeficientes provenientes de otras submuestras o la comparación interpruebas.

Tabla 3. Intervalos de Confianza para los Coeficientes de Confiabilidad de Hoyt.

	Grupo A		Grupo B	
Verbal	0,796	0,873	0,783	0,865
Numérico	0,608	0,755	0,547	0,717

En las Tablas 4 y 5 se presentan las correlaciones entre los diferentes predictores y las notas de las materias que conforman el criterio. Es interesante destacar, que, aunque se considera que el promedio de notas de educación secundaria es el mejor predictor, aquí no siempre es el caso.

Pero más importante que esto en el contexto de este trabajo, son las fluctuaciones en las correlaciones de un grupo a otro, con una tendencia del grupo A a presentar correlaciones más altas que el B. Nuevamente, aquí se corrobora la necesidad de realizar estudios de validación cruzada cuando se realizan estudios de validación. Un ejemplo permite explicar esto, la Prueba Numérica tiene una correlación de 0,516 con matemáticas en el Grupo A comparada con 0,347 en el Grupo B; al elevar al cuadrado ambos coeficientes y multiplicarlos por 100, se obtiene un 26 y un 12 por ciento de varianza explicada de las notas de Matemáticas, una diferencia realmente sustancial, que podría llevar a descartar la prueba como predictor si hubiésemos trabajado solamente con el Grupo B, pero aceptarla para el Grupo A.

Tabla 4. Correlaciones entre los predictores y las materias del primer semestre. Grupo A.

Predictor	Matem.	Química	Castellano	Inglés	C. Soc.	Biología
Prom. Sec.	0,497	0,476	0,289	0,265	0,311	0,287
P. Verbal	0,301	0,399	0,294	0,230	0,519	0,336
P. Numérica	0,516	0,393	0,265	0,320	0,293	0,291

Tabla 5. Correlaciones entre los predictores y las materias del primer semestre. Grupo B

Predictor	Matem.	Quím.	Castell.	Inglés	C. Soc.	Biología
Prom. Sec.	0,368	0,424	0,292	0,250	0,281	0,325
P. Verbal	0,240	0,268	0,342	0,201	0,398	0,169 (n.s.)
P. Numérica	0,347	0,363	0,118 (n.s.)	0,201	0,211	0,140 (n.s.)

En las Tablas 6 y 7 se presentan las correlaciones simples entre los predictores y el promedio de notas para los Grupos A y B respectivamente. La tendencia del promedio de notas de secundaria se consolida aquí como un mejor predictor que las pruebas, pero solo para el grupo B. También se repite la tendencia observada en las Tablas anteriores, de correlaciones más altas para el Grupo A que para el B. Pero también hay una menor correlación entre las variables predictoras en el Grupo B, lo que puede significar también mayor independencia entre las variables predictoras, pero correlaciones significativas con el Criterio.

Tabla 6. Correlaciones entre los predictores y el Promedio de Notas del Primer Semestre Grupo A.

Predictor	X0 Promedio notas Universidad	X1 Promedio Secundaria	X2 Prueba Verbal
X1 Promedio Secund.	0,442		
X2 Prueba Verbal	0,430	0,223	
X3 Prueba Numérica	0,433	0,416	0,370

Tabla 7. Correlaciones entre los predictores y el Promedio de Notas del Primer Semestre Grupo B.

Predictor	X0 Promedio notas Universidad	X1 Promedio Secundaria	X2 Prueba Verbal
X1 Promedio Secund.	0,446		
X2 Prueba Verbal	0,379	0,239	
X3 Prueba Numérica	0,343	0,214	0,146 (n.s.)

En la Tabla 8 se presentan los coeficientes de correlación múltiple, con las tres variables predictoras, así como la ecuación de correlación múltiple. Estas correlaciones se obtuvieron calculando Y´ para cada grupo utilizando la ecuación de regresión del otro grupo, y luego estableciendo la correlación entre Y´ y Y. Ambas correlaciones son obviamente iguales. Es decir, ambos Grupos muestran un igual nivel de predictabilidad, cuando se usa la ecuación de regresión correspondiente al otro grupo.

Tabla 8. Coeficientes de Correlación Múltiple y Ecuación de Regresión para los dos Grupos.

Grupo	Coef. de Correl.	Ecuación de Regresión
A	0,584	-1.38717+0,2976X1+0,059X2+0,0978X3
B	0,584	-2,74044+0,3297X1+0,0797X2+0,1148X3

En la Tabla 9 se presenta, en la primera fila (identificada con 0.123, que significa una ecuación con los tres predictores) la correlación calculada de igual manera que la descrita anteriormente, pero utilizando para calcular Y´ la propia ecuación de regresión del Grupo. No solo las correlaciones para ambos grupos son iguales, sino que son iguales a las calculadas cruzando las ecuaciones de regresión de un grupo al otro. Es decir, a pesar de las diferencias en las correlaciones simples con las materias de un grupo al otro, en realidad ambos tienen igual magnitud en las correlaciones múltiples.

Tabla 9. Correlaciones múltiples entre diferentes combinaciones de predictores con el Criterio.

Predictores en La Ecuación [(a)]	GRUPO A		GRUPO B	
	Correlación	Test F	Correlación	Test F
0.123	0,585		0,585	
0.13	0,520	23,92 ($p<0,01$)	0,528	16,47 ($p<0,01$)
0.23	0,522	18,14 ($p<0,01$)	0,477	32,39 ($p<0,01$)
0.12	0,557	8,08 ($p<0,01$)	0,541	12,96 ($p<0,01$)

(a) X0 Promedio universitario, X1 Promedio secundaria, X2 Prueba verbal, X3 Promedio numérico.

La Tabla 9 incluye información adicional que conviene discutir ahora. Se trata de las diferentes combinaciones de los predictores en las ecuaciones de regresión. Se evaluaron con el Test F descrito por la ecuación (6). Todas ellas son diferentes de 0 y hacen una contribución significativa a la predicción del promedio de notas, siendo la combinación más eficiente la que contiene las tres variables predictoras.

Conclusión.

Cuando se seleccionan ítems para un instrumento, se establecen estimados de confiabilidad o se determinan coeficientes de validez es necesario realizar estudios de replicación, para asegurar que se tiene estimados estables y que se están tomando las decisiones correctas con fundamento en esos estadísticos. Pero, no siempre se cuenta con recursos para realizar varios estudios sucesivos con muestras diferentes.

Este estudio muestra la posibilidad de utilizar las proposiciones de Katzell con relación a subdividir las muestras y realizar estudios paralelos que permiten corroborar y verificar los resultados obtenidos, lográndose así estudios de validación cruzada de manera relativamente económica.

Los estadísticos de muestreo que se presentan, permiten además verificar las hipótesis de igualdad o de diferencia entre los resultados, así como consolidarlos en estadísticos que tienen mayor nivel de confianza que si se trabajase con el grupo como totalidad.

Bibliografia

Anastasi, A. Urbina, S. (1998) Tests Psicológicos. Prentice Hall, México.

Cronbach, L., Rajaratman, J., Gleser G. (1963) Theory of generalizability: a liberation of reliability theory. The British Journal of Statistical Psychology, Vol. XVI, Part 2, 1-163.

Feldt, L.S. (1969). A test of the hypothesis that Cronbach´s Alpha or Kuder Richardson Coefficient Twenty is the same for the two tests. Psychometrika, 34, number 3, 363-373.

Feldt, L.S.: (1965). The approximate sampling distribution of Kuder Richardson Reliability Coefficient Twenty, Psychometrika, 30, Number 3, 375-380

Henryssen, S. (1971) Gathering, analyzing and using data on test items. R.L. Thorndike Educational Measurement, Washington, D.C.: American Council on Education.

Katzell, R. (1951) Cross validation of item analyses, Educational and Psychological Measurement, 11, 16-22

Magnusson, D. (1966) Test Theory. Reading, Mass.; Addison-Wesley Publishing Co.

McNemar, Q. (1969) Psychological Statistics, Tokyo Japan, Toppan Co, Ltd.

Mosier, C. (1951) Problems and designs of cross-validation, Educational and Psychological Measurement, 11, 5-11

Mosteller, F. y Tukey, J.W. (1968) Data analysis including statistics, The Handbook of Social Psychology, Vol. II, Reading, Mass: Addison-Wesley.

Nunnaly, J. (1987) Teoría Psicométrica. Editorial Trillas, México.

Rodríguez Trujillo, N. (1972). Construction and Analysis of a Scholastic Aptitude Test for Use in Venezuelan High School and University Populations. Tesis de maestría, Universidad de Wisconsin.

Autor: Nelson Rodríguez Trujillo PhD.
Profesor de Psicometría. Escuela de Psicología.
Universidad Central de Venezuela.
Director Gerente de Psico Consult C.A.

Titulo en Español: VALIDACIÓN CRUZADA DE PRUEBAS PSICOMÉTRICAS
Palabras Clave: Tests psicométricos, Validación cruzada, Estadísticos de los Tests.

Título en Inglés: CROSS VALIDATION IN TEST CONSTRUCTION
Key Words: psychometric tests, cross validation, tests statistics.

Resumen en español:

En el proceso de construcción de instrumentos psicométricos, es recomendable evaluar su funcionamiento en muestras sucesivas provenientes de la misma población, a fin de tomar en consideración el efecto de los errores de muestreo sobre los valores de los estadísticos de los ítems y de los tests. Este proceso se denomina Validación Cruzada. En este artículo se analizan los fundamentos de la validación cruzada, se propone una metodología para realizarla en muestras relativamente pequeñas y se presenta una aplicación tomada del contexto de la admisión estudiantil a nivel superior, que ilustra tanto el proceso de validación, como la toma de decisiones sobre la estabilidad de los estadísticos obtenidos en muestras sucesivas.

Resumen en inglés

In the development of psychometric instruments it is recommended to evaluate in different samples from the same population the results of both item and test statistics. This is called Cross Validation. In this article, the author analyzes the rationale for Cross validation and presents a methodology for its realization in small samples, as well as the sampling statistics to decide on the obtained differences. An example illustrates both the procedure and the decision making process.