Estudios Pedagógicos XLII, N° 3: 467-481, 2016
DOI: 10.4067/S0718-07052016000400025

REVISIONES

Coeficientes edumétricos para la validez y dificultad de un test: Propuesta*

Edumetrics coefficients for validity and difficulty of a test: Proposal

José González Campos,[a] *Cristian Carvajal Muquillaza,*[ab]
Francisca Viveros Reyes[c]

[a] Universidad de Playa Ancha. Valparaíso
Telf.: (32) 2500550. Correo electrónico: jgonzalez@upla.cl

[b] Telf.: (32) 2500550. Correo electrónico: cristian.carvajal@upla.cl

[c] Universidad Andrés Bello. Viña del Mar
Telf.: (32) 2845122. Correo electrónico: f.viverosr@gmail.com

RESUMEN

El siguiente artículo presenta dos propuestas de coeficientes edumétricos para la calibración de test, específicamente para la cuantificación de la validez, basado en técnicas de análisis factorial, y para la cuantificación de la dificultad de un test o ítem. Estas propuestas vienen a superar algunas de las limitaciones de los coeficientes propuestos en la teoría clásica de los test y teoría de respuesta al ítem. Se hace evidente la inconsistencia de proponer ítems difíciles sin medir aquello que se quiere, esto es, la dificultad no debe ser un elemento edumétrico ajeno y, por tanto, debe relacionarse con la fiabilidad y validez del test.

Palabras clave: validez, alfa de Cronbach, fiabilidad, edumetría, dificultad.

ABSTRACT

This article presents two proposals for edumetric calibration coefficients for a test, specifically for quantification of validity, based on factorial analysis techniques and for quantifying the difficulty of a test or item. These proposals overcome some of the limitations of the coefficients proposed in classical test theory and item response theory. It's not consistent, difficult items without measuring what we want, that is, the difficulty should not be a stranger edumetric element and therefore must relate to reliability and validity of the test.

Key words: validity, Cronbach alpha, reliability, edumetric, difficulty.

* Tercer fondo de investigación CD/PMI UPA 1203. Laboratorio de investigación Lab[e]saM. "Plan de nivelación de competencias básicas para estudiantes vulnerables académicamente de los primeros años de la Universidad de Playa Ancha" (PMI FIAC2 UPA 1104).

1. INTRODUCCIÓN

La cuantificación en educación es un proceso vivo y dinámico de suma importancia, pues por ella se evidenciará, en gran medida, el logro del proceso de enseñanza y aprendizaje, del que se desprenden variadas líneas de investigación. Existen algunos autores que intentan mejorar este proceso, enriqueciendo las propiedades métricas de los instrumentos de cuantificación, como la confiabilidad y validez, por ejemplo Brennan (2011) presenta una sustancial mejora con la teoría de generalizaciones y la manera en que esta se relaciona con la teoría de respuesta al ítem (TRI). El ajuste de un buen modelo es un factor clave en esta perspectiva, por esta razón, autores como Patarapichayatham, Tamata y Kanjanawasee (2012) discuten en profundidad esta problemática, proponiendo una serie de alternativas estadísticas que van desde lo clásico hasta herramientas Bayesianas. En este mismo sentido, Taylor (1994) plantea que si el modelo seleccionado no es el correcto se puede caer en falsas interpretaciones, por ejemplo no tener clara la diferencia entre un modelo de medida, un modelo de norma y sus proyecciones analíticas.

Por otro lado, existe una línea de investigación preocupada de estudiar y mejorar las características métricas de los test, tales como la validez y la confiabilidad (Thomas y Zumbo, 2012). Por su parte Moses y Kim (2012) orientan respecto a cómo llevar un buen proceso de comparación y clasificación de grupos, y Jacob, Goddard y Kim (2014) indican la importancia de no desperdiciar ningún tipo de información. Además, no se debe desconocer el rol del contexto y la evaluación integrada como lo indican Wilson y Sloane (2000), los alumnos son parte de un contexto y están en constante interacción.

El proceso de evaluación es un proceso vivo, dinámico y, por tanto, en constante cambio. Las metodologías de cuantificación deben ir evolucionando y buscando la representatividad, además de evolucionar y enriquecerse. Al respecto, Stiggins (1991) plantea que la tecnología se ha puesto al servicio de la evaluación, permitiendo obtener informaciones de manera mucho más rápida y procesos de resúmenes estadísticos en tiempos que día a día sorprenden. Zenisky y Sireci (2002) reconocen el impacto de las tecnologías y presentan el enriquecimiento que esto trae para los procesos de cuantificación, posteriormente Mostert y Snowball (2013) consolidan el aporte de la tecnología en la evaluación, y específicamente en cuantificación en una situación práctica. Ahora, independiente de todas las líneas de investigación antes indicadas, se debe tener una clara distinción entre evaluar tareas destinadas a facilitar el aprendizaje (método) y evaluar el aprendizaje (objetivo del proceso de enseñanza y aprendizaje), pues la trascendencia de los análisis que se derivan de cada tipo difiere enormemente (Crisp, 2012).

Los objetivos del artículo son describir brevemente el análisis factorial, describir algunos fundamentos de la teoría clásica de los test y la teoría de respuesta al ítem, proponer un coeficiente de validez, proponer un coeficiente de dificultad, relacionar los dos coeficientes, presentar aplicaciones simples y promover una línea de investigación en torno a estos conceptos que enriquezcan los procesos de calibración de instrumentos de medición de aprendizajes.

2. CONCEPTUALIZACIONES EDUMÉTRICAS

Los siguientes análisis permitirán disponer de los elementos conceptuales necesarios para la comprensión de la propuesta, con la finalidad de hacer el artículo autocontenido y así facilitar la lectura.

2.1. PRELIMINARES EDUMÉTRICOS

La mayoría de los atributos físicos resultan directamente medibles, los constructos psicosociales resultan ser conceptualizaciones teóricas que no son accesibles a la medición directa, para los que no existen "Metros" o "Balanzas" diseñadas para medirlos de manera precisa. La actitud hacia el aborto, el nivel de cohesión grupal, el coeficiente intelectual, la postura hacia el consumo de drogas, el grado de liderazgo y muchos otros ejemplos son constructos que deben medirse mediante instrumentos específicamente diseñados: Los test, cuestionarios o inventarios. Nadie dudaría que un metro bien diseñado mide longitud y que lo hace de manera precisa, pero la bondad y precisión de un cuestionario no se puede presuponer; más bien es una cuestión de grado y siempre susceptible de mejorar.

En muchos sistemas educativos surge la necesidad de disponer de un mecanismo para evaluar la adquisición de conocimiento de los alumnos, algo fundamental para identificar el éxito o fracaso en el proceso de aprendizaje. Los test son probablemente hoy en día el modo más habitual de efectuar dicha tarea, pudiendo ser utilizados en diferentes contextos educativos.

En la actualidad, los estudios se centran en poder determinar y optimizar el grado de estabilidad, precisión o consistencia que manifiesta el test como instrumento de medición del constructo. Es por esto que se han generado nuevas tendencias en el ámbito de los test, de tal manera que el constructo que determina a cada sujeto sea definido con mayor precisión (De Kohan, 2003).

Históricamente, la primera teoría fue la que hoy se conoce como teoría clásica de los test (TC), que surgió a comienzos del siglo XX. Si bien no recibió su forma axiomática hasta mediados de la década de los 60, la TC se fundamenta en el denominado modelo clásico que establece una relación lineal entre el nivel del constructo del examinado y la puntuación obtenida en el test realizado (López, Pérez y Armendáriz, 2004).

$$X = V + \varepsilon$$

Este modelo se expresa en términos de la puntuación empírica del test, el elemento central sobre el que gira toda la teoría. Concretamente, se considera que la puntuación empírica del sujeto, esto es, el valor observado en el test, es igual a la suma de dos componentes hipotéticos y desconocidos a priori: la puntuación verdadera o nivel del constructo real del evaluando y un determinado error de medida (Muñiz, 1997).

La principal limitación de la TC es que en su contexto las características del test y las del examinado son dependientes, esto es, las mediciones obtenidas dependen por lo general de la naturaleza del test utilizado y, a la inversa, las propiedades de los test dependen de los sujetos a los cuales se aplica. Así, para la TC, el nivel del constructo del alumno se mide mediante el número de respuestas acertadas en el test realizado. Por lo tanto, los resultados estarán siempre relacionados con el test administrado: Si las preguntas son difíciles, el nivel del constructo de los examinados resultará ser bajo porque habrá pocos que la respondan correctamente; y viceversa. Así, la dificultad estimada de un test según la TC dependerá de quién lo realice: Si los examinados son muy listos, responderán correctamente a las preguntas y, por tanto, el test será considerado fácil. Estas dependencias entre test y examinados supone que las medidas obtenidas en la TC no se pueden utilizar en otros contextos de manera generalizada, y por tanto comparar individuos que hayan hecho exámenes distintos resultará muy difícil (Hambleton, Swaminathan y Rogers, 1991).

La teoría de respuesta al ítem supera esta limitación, pues propone modelos matemáticos orientados a las preguntas (que en este contexto reciben el nombre de ítem), en contraposición a la TC, para el cual la evaluación de conocimientos gira en torno al test como unidad. Se trata de una teoría relativamente joven, data de comienzos de los años 60, implementada por primera vez durante los 80 (López et al., 2004).

En síntesis, la psicometría y edumetría se ocupan de los problemas de medición en Psicología y Educación, utilizando la estadística como pilar básico para la elaboración de teorías y para el desarrollo de métodos y técnicas específicas de medición (Olea et al., 2004).

Visión general de ambas teorías

En este apartado se analizarán dos de las principales teorías que se utilizan en el ámbito de la medición educacional: La TC y la teoría de respuesta al ítem (TRI).

En primer lugar, se hace necesario definir algunos términos. Distintas pruebas miden diferentes características de los examinados, por ejemplo: conocimiento matemático, razonamiento científico, capacidad de memorización, vocabulario, etc. En medición educacional, por "constructo" se entenderá a la característica del examinado que va a ser medida a través de la prueba y, por consiguiente, esta se utiliza para representar de manera genérica cualquiera de esas características. Por otro lado, en lo que respecta a los ítems de una prueba, hay parámetros (o descriptores) que permiten describir sus atributos particulares. Los parámetros de un ítem que suelen ser más frecuentemente utilizados y, por consiguiente estimados, son

a. Nivel de dificultad.
b. Nivel de discriminación de un ítem.
c. Efectos del azar.

Una forma de medición que se utiliza con mucha frecuencia en pruebas de gran escala, como las de admisión a la universidad (por ejemplo, en el Scholastic Aptitude Test (SAT) de Estados Unidos y en la Prueba de Selección Universitaria (PSU) en Chile), es la llamada teoría clásica. En teoría clásica, el indicador del constructo de un estudiante corresponde al puntaje que se obtuvo en la prueba, construido a partir del número de respuestas correctas (o número de respuestas correctas netas) que obtuvo. Como indicador de la dificultad de una pregunta, en este sistema se utiliza la razón entre el número de personas que acertaron correctamente el ítem y el total de alumnos que dieron una respuesta a este. Por otro lado, el índice de discriminación de un ítem se calcula como la correlación entre el puntaje de la respuesta a este y el puntaje en la prueba total. Además, en los test formados por ítem de opción múltiple, en las que solamente una alternativa es la correcta, se puede sobrestimar la puntuación directa de una persona dado que alguno de sus aciertos ha podido producirse por azar. El problema, entonces, consiste en establecer un procedimiento para descontar del número total de aciertos que se han producido por azar. Si se asume que no se conoce la respuesta correcta a un ítem, todas las alternativas de respuesta son equiprobables y la probabilidad de acertar al azar ese ítem se puede establecer como: $^{1}/_{n}$, donde n es el número de alternativas del ítem. Estos descriptores de los ítems pueden calcularse ya sea en el contexto de una prueba piloto o experimental, o en la prueba definitiva u operacional.

Como se puede apreciar, en la TC el grado del constructo de una persona depende del grupo de ítems (vale decir, de sus diferentes parámetros) que contiene la prueba. Por ejemplo, si la prueba es fácil, un mismo alumno tendrá un puntaje mayor que si la prueba es difícil. Con esto resulta complejo hacer comparaciones entre estudiantes que han rendido pruebas diferentes. A su vez, la estimación de los parámetros depende del grupo de personas a las cuales se les aplicó el test. Así, un mismo ítem puede ser catalogado como fácil si el grupo que rindió la prueba es excepcionalmente hábil, pero como difícil si el grupo que rindió la prueba es desaventajado. Con respecto a la discriminación, un ítem puede aparecer muy discriminativo en el contexto de un grupo con nivel heterogéneo del constructo, pero poco discriminativo si el grupo que rindió la prueba es homogéneo (es decir, si todos los estudiantes tienen un nivel del constructo similar). Esta debilidad que se genera en la estimación del constructo de los examinados y los parámetros de los ítems ha llevado a buscar un método que permitiese una medida del constructo que sea independiente de los ítems a los que estos se han enfrentado, una caracterización de los ítems independiente de la población a la que se aplica y, al mismo tiempo, una medida más fiel de la precisión con que se está midiendo el constructo.

La satisfacción de estos requerimientos se ha intentado encontrar en la teoría de respuesta al ítem. Los modelos de la TRI se centran en los ítems e intentan establecer, para cada uno de ellos, la probabilidad de ser respondidos correctamente. Esta probabilidad depende del constructo del examinado y de ciertas características de los ítems, entre las que se encuentra el grado de dificultad, discriminación y la probabilidad de ser respondido correctamente debido al azar por una persona de bajo nivel del constructo. Hay varios modelos en TRI de distinta complejidad. El modelo más simple es aquel que solo diferencia los ítems según su grado de dificultad. Sin embargo, otros modelos permiten, además, incluir nuevos parámetros a los ítems, como su grado de discriminación y la probabilidad de responderlo correctamente al azar.

Además, la TRI permite conocer el nivel de precisión que un ítem aporta a la estimación para cada nivel del constructo. En términos técnicos esto es lo que se conoce como información de Fischer, que contextualizado se denomina información del ítem. Mientras mayor es la información que aporta el ítem a un determinado nivel del constructo, mayor es la precisión en la estimación de ese nivel.

La principal ventaja teórica de la TRI es que mediante su utilización se lograría que un estudiante obtuviese siempre la misma estimación de su nivel del constructo, independientemente de las preguntas que deba responder. También, con la TRI un ítem tendría siempre los mismos parámetros que lo describen (dificultad, discriminación, azar), independientemente del grupo que rindió la prueba. Esta notable propiedad se llama invarianza, y constituye la principal ventaja que distingue a la TRI de la teoría clásica. Sin embargo, es preciso destacar que la invarianza se cumple siempre y cuando se satisfagan ciertos supuestos y requisitos que se enunciarán a continuación.

a. Unidimensionalidad: este supuesto consiste en que en una prueba todos los ítems están midiendo uno y solamente un constructo en los examinados.
b. Independencia local: postula que, dado un nivel del constructo, las respuestas a los ítems no pueden estar correlacionadas entre sí.
c. Igualdad de educación: que todos los alumnos que rindan la prueba hayan tenido experiencias educacionales similares.

d. Que la prueba no haya sido apurada, es decir que los tiempos estimados en el desarrollo de la prueba se respeten.
e. Efectos de contextos no controlados: se refiere a que algunas preguntas se comportan de modo diferente según la posición que tenga en la prueba.

La teoría de respuesta al ítem presenta una serie de potenciales ventajas sobre la teoría clásica. La principal de ellas es la invarianza de los puntajes de la prueba y de las características de las preguntas. También surge, gracias a las curvas de información, herramienta exclusiva de la TRI, la posibilidad de optimizar el proceso de selección de preguntas según el objetivo que se busca.

Todas las ventajas anteriores se pierden cuando los supuestos y requisitos de la TRI no se cumplen. Para que haya invarianza, principal ventaja de la TRI sobre la TC, es fundamental que exista, como se señalaba anteriormente, unidimensionalidad e independencia local. Sin embargo, en muchas ocasiones la naturaleza de las disciplinas mismas les impide someterse a tales restricciones. Es por ello que en la práctica los supuestos y requisitos de la TRI a menudo se transgreden (Dussaillant, 2003). Este incumplimiento de los supuestos lleva no solo a perder la invarianza, sino que afecta directamente a la estimación del constructo e introduce errores en aplicaciones secundarias de la teoría.

2.2. PRELIMINARES ESTADÍSTICOS

El análisis factorial es una técnica multivariada de reducción de dimensionalidad de un conjunto de variables, las que generalmente pueden ser utilizadas para estudiar propiedades métricas de los test, identificando cada ítem que constituye el test como una variable aleatoria (Flores et al., 2012). La reducción de dimensionalidad se debe a la capacidad del análisis factorial de agrupar conjuntos de variables en función de su cohesión, denominándolas factores. El análisis factorial consiste en definir un nuevo conjunto de variables no correlacionadas, estas nuevas variables son ponderadas por las cargas de los factores, que dependen de la fuerza de la correlación de las variables originales (Castrillón y Borrero, 2005). Por ejemplo, si un conjunto de variables tiende a correlacionar más fuertemente que otro, entonces la ponderación o carga de ese factor ha de ser mayor.

El análisis por factores tiene los siguientes propósitos (Batista-Foguet, Coenders y Alonso, 2004):

– Crear un nuevo conjunto de variables no correlacionadas, llamadas factores subyacentes, con la esperanza de que estas proporcionen una mejor comprensión de los datos que se están analizando.
– Determinar si existe un conjunto más pequeño de variables no correlacionadas que expliquen las relaciones que existen entre las variables originales.
– Determinar el número de variables o factores subyacentes.
– Interpretar estas nuevas variables.

Por tanto, la metodología factorial es usada como una medida de validez.

3. HERRAMIENTAS EDUMÉTRICAS PROPUESTAS

La metodología de trabajo es de tipo propositiva, en el sentido de presentar dos nuevos coeficientes edumétricos en beneficio de mejorar las características métricas de los test.

3.1. PROPUESTA 1

La primera propuesta que presentamos es un coeficiente de validez, el cual busca ser coherente a la definición: un test será válido en la medida que mida la dimensión para la cual fue diseñado (Corral, 2009). Para ello se basa en técnicas de análisis factorial y propiedades matriciales.

Diremos que R es la matriz de correlación, de un test, con p ítems, en donde la puntuación a cada ítem puede ser expresada dependiendo en forma lineal de nuevas variables denominados factores y cargas respectivas. Es decir, sea la matriz de respuestas emitidas al test, el cual está constituido por p preguntas o ítems y que fue respondido por m personas. Formalmente escribiremos $X_{pxm} = \begin{pmatrix} \kappa_{11} & . & \kappa_{1k} & . & \kappa_{1m} \\ . & & . & & . \\ . & & . & & . \\ \kappa_{j1} & . & \kappa_{jk} & . & \kappa_{jm} \\ . & & . & & . \\ \kappa_{p1} & . & \kappa_{pk} & . & \kappa_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ : \\ f_m \end{pmatrix}$, donde $f_k; k = 1,...,m$ representa el k-ésimo factor y $\kappa_{jk}; j = 1,...,p : k = 1,...,m$ cargas de los factores. El método de los factores principales sobre R, inicialmente requiere de las estimaciones apropiadas de las comunidades o, lo que es equivalente, estimación de las varianzas específicas de cada ítem, lo que simbolizaremos por $\psi_1, \psi_2, ..., \psi_p$, las que son utilizadas para modelar R de la siguiente manera, $R = \Lambda\Lambda^t + \Psi$, donde la $\Lambda^t\Lambda$ es utilizada para representar los niveles medios de correlación para cada ítem. A partir de esta ecuación podemos despejar la siguiente relación $\Lambda\Lambda^t = R - \Psi$. Sin embargo, para llegar a obtener una solución única para Λ, es necesario forzar a $\Lambda^t\Lambda$ ser una matriz diagonal (Barbolla, 1998), lo que anotaremos $\Lambda^t\Lambda = D$. Luego si $\Lambda\Lambda^t = R - \Psi$ podemos post-multiplicar por la matriz , obteniendo $\Lambda\Lambda^t\Lambda = (R - \Psi)\Lambda$, por propiedades de las matrices diagonales (Sanz y Barbolla, 1998), tenemos que $\Lambda D = (R - \Psi)\Lambda$, particularmente si observa una componente de esta relación tendremos: $[d_1\kappa_1, ..., d_m\kappa_m] = [(R - \Psi)\kappa_1, ..., (R - \Psi)\kappa_m]$. Finalmente, la única forma de que las ecuaciones antes planteadas puedan ser verdaderas es si los elementos de la diagonal de D son autovalores de la matriz caracterizada por $R - \Psi$ y si las columnas de Λ son los autovectores correspondientes. Cualquier subconjunto de *m* autovalores y autovectores resolverían las ecuaciones de los factores principales, pero si se eligen los vectores correspondientes a los *m* autovalores más grandes, puesto que los elementos de D son comunidades, entonces pueden ser expresados como $d_k = \sum_{j=1}^{p} \kappa_{jk}^2$, para $k = 1,2,...,m$. Al elegir los vectores correspondientes a los *m* autovalores más grandes, se puede maximizar las comunidades, cuyo efecto recae en la maximización de los κ_{jk} (las

cuales identifican las cargas de los factores), de donde estos vectores deben corresponder a los factores más importantes.

Observemos que si todas las comunidades toman el valor 1, entonces la matriz de varianzas específica a cada ítem será 0, esto es $\Psi = 0$ y el método de los factores principales se reduce a un análisis de componentes principales sobre la matriz de correlación.

Retomando la relación $\Lambda D = (R - \Psi)\Lambda$, luego como las columnas de la matriz de cargas de los factores, Λ, son los autovectores de $R - \Psi$, la que cumple con ser simétrica y tener la propiedad de $\Lambda^{-1} = \Lambda^t$. Esta última relación permite generar una nueva relación, esto es $\Lambda D\Lambda^{-1} = (R - \Psi)$ puede ser escrita como $\Lambda D\Lambda^t = (R - \Psi)$. Finalmente, estas ecuaciones permiten establecer que la correlación entre los ítems y la varianza de ellos es expresada en función de los autovalores y autovectores, por lo tanto, mientras mayor es un autovalor, mayor es la correlación y varianza de los ítems que está siendo explicada por el autovalor.

De esta forma, diremos que un test es válido en la medida que las variables originales se agrupen en torno a un solo factor, de manera equivalente, la dimensión de los ítems tiende a la unidad.

Bajo esta concepción de validez y los supuestos de la teoría clásica de los test (Muñiz, 1997) se define el siguiente *coeficiente de validez.* $\Gamma = \dfrac{1 + (n-1)\rho_{xx'}}{n}$, en donde $\rho_{xx`}$ es el coeficiente de fiabilidad del test, *n* es el número de ítems del test.

Una relación importante basada en los autovalores de una matriz es la siguiente: si A es una matriz cuadrada de orden *n,* es decir el número de filas y columnas es el mismo, tal que cada componente de la matriz A, que lo anotaremos por α_{ii}, será considerado como un número real, específicamente, $a_{ii} = k$, donde los subíndices pueden asumir los valores $i : 1,..., n$, entonces la suma de los autovalores de la matriz A, que lo simbolizaremos por λ_i, es igual a *k* veces el orden de la matriz, es decir: $\sum_{i=1}^{n} \lambda_i = nk$.

Otro resultado importante para poder entender algunas de las propiedades del coeficiente Γ es el siguiente: si A es la matriz de e*quicorrelación*, esto es que todos los ítems tiene el mismo grado de asociación lineal, entonces $\sum_{i=1}^{n} \lambda_i = n$.

Finalmente, otro importante resultado de este artículo es la siguiente relación: Sea λ_1 el mayor de los valores propios de una matriz de factores. Si $\lambda_1 \geq 2$ entonces el *coeficiente de fiabilidad alfa de Cronbach* (Cronbach, 1951; Cronbach y Meehl, 1955) es una cota superior para *el coeficiente de validez*, es decir: La demostración de esta importante relación necesita del conocimiento de ciertos elementos algebraicos, sin embargo se han explicitado cada uno de los cálculos que fundamentan la anterior afirmación.

Demostración: Para efectos de escritura, diremos que: $\lambda = \lambda_1$, luego, dado $\lambda \geq 2$, se tiene que, donde n simboliza a un número natural mayor que 2. El origen de esta afirmación se debe a que alcanza como máximo el valor 2. Por otro lado bajo los supuestos de la Teoría Clásica de los Test se tiene que $\lambda = 1 + (n-1)\rho_{xx'}$, así $1 + (n-1)\rho_{xx'} \geq \dfrac{n}{n-1}$, equivalentemente $1 + (n-1)\rho_{xx'} \geq 1 + \dfrac{1}{n-1}$, luego $\rho_{xx'} \geq \dfrac{1}{(n-1)^2}$, esta última relación nos permite implicar que: $1 + (n-1)\rho_{xx'} \leq n\sqrt{\rho_{xx'}}$, ahora la siguiente secuencia de igualdades son equivalentes:

$$\frac{\lambda^2}{n^2} \leq \rho_{xx'}.$$

$$\frac{\lambda^2}{n} \leq \left(\frac{n}{n-1}\right)(n-1)\rho_{xx'}.$$

$$\frac{1+(n-1)\rho_{xx'}}{n} \leq \left(\frac{n}{n-1}\right)\left(\frac{n(n-1)\rho_{xx'}}{n+n(n-1)\rho_{xx'}}\right)$$

$$\frac{1+(n-1)\rho_{xx'}}{n} \leq \left(\frac{n}{n-1}\right)\left(\frac{\sigma_x^2 n(n-1)\rho_{xx'}}{\sigma_x^2\left(n+n(n-1)\rho_{xx'}\right)}\right).$$

$$\frac{1+(n-1)\rho_{xx'}}{n} \leq \left(\frac{n}{n-1}\right)\left(\frac{\sum_{j\neq k}^{n}\sum Cov(x_j,x_k)}{\sigma_x^2 n+\sigma_x^2 n(n-1)\rho_{xx'}}\right).$$

$$\frac{1+(n-1)\rho_{xx'}}{n} \leq \left(\frac{n}{n-1}\right)\left(\frac{\sum_{j\neq k}^{n}\sum Cov(x_j,x_k)}{\sum_{j=1}^{n}\sigma_j^2+\sum_{j\neq k}^{n}\sum Cov(x_j,x_k)}\right).$$

$$\frac{1+(n-1)\rho_{xx'}}{n} \leq \left(\frac{n}{n-1}\right)\left(\frac{\sum_{j\neq k}^{n}\sum Cov(x_j,x_k)}{\sigma_X^2}\right).$$

$\Gamma \leq \alpha$, lo que permite concluir la demostración y fundamentar formalmente la afirmación.

El coeficiente de Validez Γ, aquí definido, representa la razón entre la fuerza de agrupación de las variables en torno al primer factor, es decir el mayor de los autovalores, que se representará por $\lambda_{(1)}$, respecto de la Correlación y Varianza total del test.

Algunos de los aspectos que se deben indicar como ventajas de la propuesta son los siguientes. El coeficiente de Validez Γ supera la limitación dada por la subjetividad y dependencia de un criterio existente. Otro atributo de consistencia del coeficiente de Validez, es su relación directa con el coeficiente de fiabilidad (alfa de Cronbach), pues al aumentar la fiabilidad del test, la validez también lo hace.

Finalmente, se observa la consistencia asintótica de este estimador de la validez, esto es: en la medida que la fiabilidad tiende a ser perfecta, también lo hace el coeficiente de validez, simbólicamente esto es representado por: $\lim_{\rho_{xx'}\to 1}\Gamma = \lim_{\rho_{xx'}\to 1}\frac{1+(n-1)\rho_{xx'}}{n} = 1$

3.2. PROPUESTA 2

La segunda propuesta que se presenta es la definición de un coeficiente de dificultad. La definición del concepto de dificultad que se presenta en numerosos artículos está sujeta a la muestra, esto es, un mismo test puede arrojar diferentes coeficientes de dificultad, a tal punto de pasar de dificultad extrema a un test extremadamente fácil y en otras ocasiones, no menores, se recurre al juicio de experto.

La motivación para esta propuesta surge a partir de las reflexiones del filósofo alemán Heidegger, que en síntesis plantea lo siguiente:

> Se dice: "Ser" es el más universal y vacío de los conceptos. En cuanto tal, resiste a todo intento de definición. Es el más universal de los conceptos y, por ende, indefinible; no ha menester de definición. Todos lo usamos constantemente y comprendemos también lo que en cada caso queremos decir con él. De esta suerte, lo que como algo oculto sumió y mantuvo en la inquietud el filosofar de la antigüedad, se convirtió en una cosa comprensible de suyo y tan clara como el sol, hasta el punto de que a quien sigue haciendo aún la pregunta se le tacha de error metódico. (Canals, 1974, p. 53).

Existen muchos conceptos o expresiones que son utilizadas frecuentemente, sin embargo son expresiones automáticas y que todo el mundo entiende al ser utilizadas, particularmente el concepto de dificultad, el cual es un concepto altamente conocido, utilizado y entendido por cualquier persona, sin embargo, a la hora de buscar una definición de este concepto encontramos que no existe definición concreta u objetiva. En el ámbito de evaluación educacional, y específicamente en la calibración de instrumentos de evaluación o en Teoría de respuesta al ítem, es usado el concepto de dificultad proponiendo algunos coeficientes cuya dependencia del grupo al que se aplica caracteriza su limitación, además de que la obtención de buenas estimaciones solo son alcanzadas a tamaños muestrales excesivamente grandes y atípicos, como es el caso de la teoría de respuesta al ítem.

En el camino a superar estas limitaciones y en coherencia a la reflexión previa, se propone un método para aproximar al concepto de dificultad y un coeficiente que permita su operatividad.

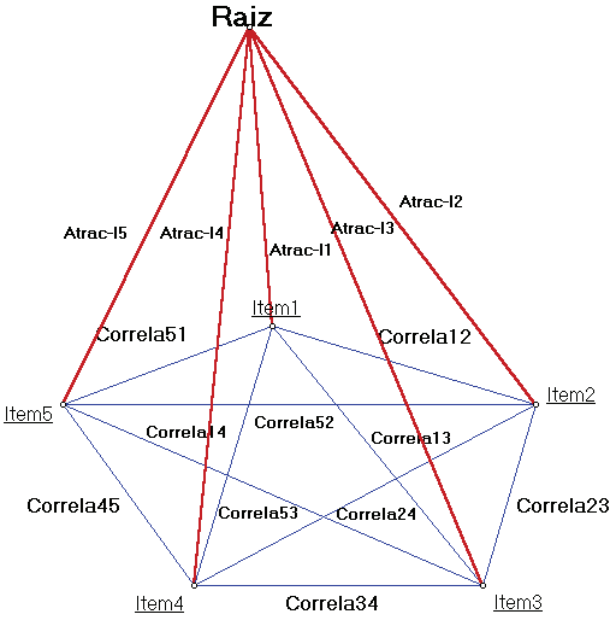
Considérese un test con n ítems, donde I_i es el *i-ésimo* ítem del test, en donde i puede tomar los valores $i = 1,2,...,n$. Diremos que $\delta(I_i)$ es el peso del ítem *i-ésimo* del test, que es definido como: $\delta(I_i) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} \rho_{ij}}{n-1}$, el cual viene a caracterizar una media entre las correlaciones del ítem **i** con cada uno de los demás ítems que constituyen el test, por lo tanto cada ítem será caracterizado por su peso.

Una relación interesante que se puede desprender a partir de esta definición y bajo los supuestos de la teoría clásica de los test, es que el coeficiente *Alfa de Crombach* puede ser representado por: $\alpha = \frac{1}{\lambda_1} \sum_{i=1}^{n} \delta(I_i)$, además, es llamado atracción del ítem respecto de la característica que mide el test al coeficiente $R_i = \frac{f_{1i}^2}{\lambda_1}$, donde:

f_{1i}^2 : Representa la cantidad de varianza del ítem ***i*** explicada por el primer factor.
λ_1 : Representa la proporción de varianza de los ítems explicada por el primer factor.

Este coeficiente representa la razón entre la varianza del ítem **i** respecto de la varianza total del test, *ambas explicadas por el primer factor*. Para facilitar la comprensión de dicha propuesta, se recurre a la teoría de grafos para su visualización.

Figura 1. Diagrama de correlaciones entre ítems



Además de las limitaciones de las propuestas de coeficientes de dificultad existentes, se hacen evidentes las siguientes inquietudes: ¿Tiene sentido un ítem "*difícil*" si no está midiendo lo que se quiere medir con el test? ¿Por qué se dice que un ítem es "*difícil*"? ¿Es acaso porque la información que entrega es un claro reflejo de la posesión de la característica de interés o porque refleja procesos mentales complejos?

Estas inquietudes vienen a orientar esta segunda propuesta a describir.

Definición: un ítem es *difícil* si su peso y su atracción son altos. En otras palabras, si la correlación con los demás ítems es alta y además mide aquello que debe medir. Bajo estos supuestos se operacionaliza el coeficiente de dificultad a través de la siguiente propuesta: $D_i = R_i + \delta(I_i),\ i = 1,2,...,n$.

Por medio de la definición de este coeficiente se acota el concepto de dificultad a un contexto de un test condicionado a la *fiabilidad* y *validez,* siendo coherente a las inquietudes antes planteadas.

Lo importante de contar con este coeficiente es su objetividad, el desligamiento de la dificultad de ítems solo por dificultad, sin tener un objetivo particular y además por las siguientes propiedades:

Propiedades de D_i

1. $0 \le D_i \le 2$
2. $\sum_{i=1}^{n} D_i = 1 + \lambda_1 \cdot \alpha$
3. $\overline{D} = \frac{1}{n} + \Gamma \cdot \alpha$

La tercera propiedad es sumamente interesante, pues se relaciona a la fiabilidad y validez del test, en donde representa la dificultad media del test.

Todos estos resultados son ideas basales para nuevas líneas de investigación en torno a las propiedades métricas de las dos propuestas aquí presentadas.

4. APLICACIÓN

Para la aplicación fue considerado un test de 4 ítems o preguntas, que fue aplicado a 5 alumnos. Se indica que esta situación es solo para ejemplificar. La siguiente tabla resume los resultados de esta aplicación, en donde las columnas caracterizan los ítems del test y las filas a los alumnos sometidos a este test. Cada ítem podría obtener una puntuación entre 1 y 7, ambos incluidos.

Tabla 1. Tabla de datos

Sujeto	Ítem1	Ítem2	Ítem3	Ítem4
1	1	5	2	7
2	2	3	4	6
3	4	4	3	3
4	5	5	6	7
5	6	7	6	7

Basado en esta aplicación fueron calculados algunos elementos, como por ejemplo:

- $\lambda_1 = 2.477$
- $f_{11}^2 = 0.852$
- $f_{14}^2 = 0.517$

Un coeficiente que está presente en todo proceso de cuantificación de la consistencia interna de un test es el coeficiente alfa de Cronbach, cuya cuantificación fue $\alpha = 0.77$, de manera similar fue determinado nuestro índice de fiabilidad propuesto en este artículo, $\Gamma = 0.62$, en el cual se evidencia la superioridad del coeficiente alfa de Cronbach, tal como fue demostrado en los desarrollos teóricos de esta propuesta.

En relación al coeficiente de dificultad, solo se hará la estimación en los ítems 1 y 4, obteniéndose lo siguiente:

Dificultad del ítem 1: $D_1 = 0.827$

Dificultad del ítem 4: $D_4 = 0.49$

Considerando los pesos de cada ítem y particularmente la suma de ellos se tiene $\sum_{i=1}^{4}\delta(I_i)=1.907$, lo que permite evidenciar la relación con el coeficiente alfa de Cronbach, esto es $\alpha=\frac{1}{2.477}\cdot 1.907=0.7$, permitiendo concluir la consistencia con los estadígrafos clásicos de la teoría de los test, además permite visualizar una serie de relaciones funcionales entre ellos y situar en contextos formales la dificultad, esto es fiabilidad y validez, por otro lado su interpretabilidad adquiere significado y se vuelve objetiva.

5. CONCLUSIONES

Se indica que las propuestas aquí orientadas están sujetas a restricciones métricas, esto es considerar el estatus métrico de la puntuación de cada ítem. Por tanto, si los estatus métricos son cualitativos, la metodología no es aplicable. Existe un error frecuente en el abuso del uso de la estadística, pudiendo llegar a encontrar estadígrafos como media o varianza en variables cualitativas. Por otro lado, con estas propuestas no se resuelve el problema, sin embargo supera algunas limitaciones de la teoría clásica y teoría de respuesta al ítem.

Es establecida una conexión entre los conceptos de fiabilidad, validez y dificultad, dando sentido al concepto de dificultad, esto es, desde el punto de vista educacional no tendría sentido hablar de una pregunta o ítem difícil si este no mide aquello que se quiere. Por lo tanto estos coeficientes edumétricos nos permiten enriquecer las características métricas de las pruebas o test, caracterizando procesos de calibración independientes de las personas que rinden la prueba o test, permitiendo la comparación objetiva entre grupos sometidos a la evaluación. También se superan algunas limitaciones de la teoría de respuesta al ítem, como tamaños muestrales excesivamente grandes, que desde el punto de vista práctico difícilmente serán observados, además se supera la limitación de depender del grupo de alumnos al cual se aplica, es decir es una propuesta de estimación de la dificultad objetiva.

Se propone un coeficiente de validez que acota inferiormente la fiabilidad o confiabilidad, además de presentar consistencia asintótica, basado en técnicas de análisis factorial y reducción de dimensionalidad del test. Con esta propuesta no se soluciona el problema de calibración de un test o de sus características métricas, pero sí se mejoran algunas cualidades de los ya existentes. Por otro lado, la implementación es sencilla para cualquier software. En trabajos futuros, se pretende formalizar aspectos inferenciales de las propuestas.

La definición de este conjunto de herramientas edumétricas permitirá dar mayor consistencia a las conclusiones vertidas de los análisis de datos obtenidos por medio de test o instrumentos de medición.

REFERENCIAS BIBLIOGRÁFICAS

Barbolla, R., & Sanz P. (1998). *Álgebra y Teoría de Matrices*. Madrid: Prentice Hall.
Batista-Foguet, J., Coenders, G., & Alonso, J. (2004). Análisis factorial confirmatorio. Su utilidad

en la validación de cuestionarios relacionados con la salud. *Medicina Clínica*, *122*, 21-27. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=edo&AN=13090084&lang=es&site=eds-live

Brennan, R. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1-21. doi:10.1080/08957347.2011.532417

Canals, F. (1974). *Textos de los grandes filósofos: Edad contemporánea. Selección de textos de Marx, Kierkegaard, Nietzsche, Comte, Bergson, Blondel, Husserl, Marcel y Heidegger.* Barcelona: Herder.

Castrillón, D., & Borrero, P. (2005). Validación del inventario de ansiedad estado-rasgo (STAIC) en niños escolarizados entre los 8 y 15 años. *Acta Colombiana de Psicología*, *8*(1), 79-90. Recuperado de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-91552005000100005&lng=en&tlng=en

Corral, Y. (2009). Validez y confiabilidad de los instrumentos de la investigación para la recolección de datos. *Revista Ciencias de la Educación*, *19*(33), 228-247. Recuperado de http://www.riuc.bc.uc.edu.ve/bitstream/123456789/1949/1/ycorral.pdf

Crisp, G. (2012). Integrative assessment: reframing assessment practice for current and Future learning. *Assessment & Evaluation in Higher Education*, *37*(1), 33-43. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=eric&AN=EJ950338&lang=es&site=eds-live

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=edb&AN=71554458&lang=es&site=eds-live

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281-302. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=mdc&AN=13245896&lang=es&site=eds-live

De Kohan, N. (2003). Posibilidad de integración de las teorías cognitivas y la psicometría moderna. *Interdisciplinaria: Revista de Psicología y Ciencias Afines*, *22*(1), 29-58. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=fua&AN=21093393&lang=es&site=eds-live

Dussaillant, F. (2003). *Técnicas de medición en pruebas de admisión a las universidades.* Santiago: Centro de Estudios Públicos. Recuperado de http://www.cepchile.cl/dms/archivo_3192_1472/rev90_dussaillant.pdf

Flores, F., Peinado, J., Ornelas, M., & López, L. (2012). Composición Factorial de una Escala Autoeficacia en Conductas de Cuidado de la Salud en Estudiantes de Ingeniería. *Formación universitaria*, *5*(3), 43-54. doi:10.4067/S0718-50062012000300006

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory* (2da ed.). London: Sage publications.

Jacob, R., Goddard, R., & Kim, E. (2014). Assessing the use of aggregate data in the evaluation of school-based interventions implications for evaluation research and state policy regarding public-use data. *Educational Evaluation and Policy Analysis*, *36*(1), 44-66. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=eric&AN=EJ1019191&lang=es&site=eds-live

López, J., Pérez, T., & Armendáriz, A. (2004). Evaluación Mediante Test, Departamento de Lenguas y Sistemas Informáticos. *Revista Iberoamericana de Educación*, *3*(1), 664-668. Recuperado de http://rieoei.org/deloslectores/1040Lopez.PDF

Moses, T., & Kim, S. (2012). Evaluating ranking strategies in assessing change when the Measures differ across time. *Educational and Psychological Measurement*, *72*(1), 78-98. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=eric&AN=EJ956062&lang=es&site=eds-live

Mostert, M., & Snowball, J. (2013). Where angels fear to tread: Online peer-assessment in a large

first-year class. *Assessment & Evaluation in Higher Education*, *38*(6), 674-686. doi:10.1080/02602938.2012.683770

Muñiz, J. (1997). *Teoría de Respuesta al Ítem*. Oviedo: Pirámide.

Olea, J., Ponsoda, V., Abad, F., Revuelta, J., Gil, B., & García, C. (2004). *Introducción a la Psicometría, Teoría Clásica de los Test y Teoría de la Respuesta al Ítem*. Madrid: Universidad Autónoma de Madrid.

Patarapichayatham, C., Kamata, A., & Kanjanawasee, S. (2012). Evaluation of model selection strategies for cross-level two-way differential item functioning analysis. *Educational and Psychological Measurement*, *72*(1), 44-51. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=eric&AN=EJ956067&lang=es&site=eds-live

Stiggins, R. (1991). Facing Challenges of a New Era of Educational Assessment. *Applied measurement in education*, *4*(4), 263-273. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=a9h&AN=7366090&lang=es&site=eds-live

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal*, *31*(2), 231-262. Recuperado de http://recursosbiblioteca.unab.cl:2173/stable/1163308

Thomas, D., & Zumbo, B. (2012). Difference Scores From the Point of View of Reliability and Repeated-Measures ANOVA In Defense of Difference Scores for Data analysis. *Educational and Psychological Measurement*, *72*(1), 37-43. doi:10.1177/0013164411409929

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*(2), 181-208. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=a9h&AN=3341990&lang=es&site=eds-live

Zenisky, A., & Sireci, S. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, *15*(4), 337-362. Recuperado de http://recursosbiblioteca.unab.cl:2059/login.aspx?direct=true&db=eric&AN=EJ667253&lang=es&site=eds-live