Revista de Psicología
Universidad de Chile

2014, 23(1), 12-20
ISSN impreso : 0716-8039
ISSN en línea : 0719-0581
www.revistapsicologia.uchile.cl

Coeficiente Phi(Lambda) y la fiabilidad de las decisiones sobre selección de personal

Phi(Lambda) coefficient and the reliability of decisions on personnel selection

René Gempp
Universidad Diego Portales, Santiago, Chile

Resumen

En el contexto de selección de personal es habitual que los tests psicológicos sean utilizados para hacer clasificaciones dicotómicas de personas. En esos casos, los coeficientes de fiabilidad convencionales, diseñados para estimar la fiabilidad de un puntaje, no resultan apropiados para estimar la precisión de la clasificación resultante. En este trabajo se presenta el coeficiente Phi(Lambda) desarrollado por Brennan y Kane (1977) en el marco de la Teoría de la Generalizabilidad (Brennan, 2001), para estimar la fiabilidad de un punto de corte. Además, se presenta un software gratuito desarrollado para facilitar la estimación del coeficiente y se demuestra su uso a través de un ejemplo empírico, con un test de selección de personal.

Palabras clave: selección de personal, fiabilidad, clasificación, Phi(Lambda), Teoría de la Generalizabilidad.

Abstract

In the context of personnel selection it is usual that psychological tests are employed for dichotomous classifications of applicants. In such cases, the conventional reliability coefficients, designed to estimate score reliability, are not suitable for estimating the accuracy of classifications. This paper introduces the Phi(Lambda) coefficient developed by Brennan and Kane (1977) in the framework of Generalizability Theory (Brennan, 2001) for estimating the reliability of a cutting point. In addition, a free software was developed to facilitate the estimation of the coefficient and its use is demonstrated through an empirical example, using a recruitment test.

Keywords: personnel selection, reliability, classification, Phi(Lambda), Generalizability Theory.

Contacto: R. Gempp. Facultad de Economía y Empresa, Universidad Diego Portales, Av. Santa Clara 797, Huechuraba, Santiago, Chile.
rene.gempp@udp.cl

Cómo citar este artículo:
Gempp, R. (2014). El coeficiente Phi(Lambda) y la fiabilidad de las decisiones sobre selección de personal. *Revista de Psicología, 23*(1), 12-20.
doi: 10.5354/0719-0581.2014.32869

Introducción

En los procesos de selección de personal, los psicólogos habitualmente utilizan instrumentos psicométricos para evaluar habilidades, conocimientos, intereses y/o personalidad de los postulantes, con la finalidad de identificar a los candidatos con mayor probabilidad de ajustarse a las necesidades del cargo. Teóricamente, los procesos de selección descansan sobre varios supuestos. El más básico es que las personas difieren en función de sus atributos psicológicos, mientras que, por otro lado, los puestos de trabajo difieren en términos de las habilidades y cualidades que requieren para desempeñarse correctamente en ellos. Bajo este supuesto, en los procesos de selección se intenta localizar a la persona más idónea en el cargo más compatible con su perfil de competencias y personalidad. Un segundo supuesto, es que las habilidades, conocimientos, intereses y/o personalidad de los postulantes en el momento de ser evaluados pueden predecir su comportamiento o desempeño futuro en el trabajo. Por último, y no menos importante, un tercer supuesto es que los instrumentos psicométricos que se utilizan en el proceso de selección son capaces de medir con un grado razonable de precisión las variables psicológicas evaluadas.

Desde un punto de vista psicométrico, este último supuesto implica que los instrumentos utilizados deben demostrar evidencia de validez predictiva y de fiabilidad. Aunque la validez predictiva de los tests de selección de personal ha generado un amplio debate y un cúmulo de investigaciones a través de los años, su fiabilidad ha merecido menor interés por parte de los investigadores. De hecho, la mayoría de las guías técnicas y trabajos de revisión (e.g., Sackett y Lievens, 2008; Society for Industrial and Organizational Psychology, 2003) se concentran en cuestiones relativas a la validez en desmedro de la fiabilidad de los instrumentos de selección. A modo de ejemplo, el número más reciente del *Annual Review of Psychology* dedicado a selección de personal (Ryan y Ployhart, 2014) aborda minuciosamente tópicos de validez, pero no de fiabilidad en instrumentos de selección. Asimismo, el recientemente publicado *Handbook of Employee Selection* (Farr y Tippins, 2010) dedica apenas medio capítulo (Putka y Sackett, 2010) al problema de la fiabilidad de las puntuaciones en selección de personal, pero no aborda el problema de la fiabilidad de las decisiones de selección. En la misma línea, tampoco es posible localizar trabajos sobre fiabilidad de las decisiones de selección de personal al hacer una revisión exhaustiva de los artículos publicados en los últimos 15 años en revistas especializadas en el área, como *Personnel Psychology, Educational and Psychological Measurement, Applied Psychological Measurement, Organizational Research Methods* o el *International Journal of Selection and Assessment*. En suma, el problema de la fiabilidad de las decisiones de selección personal parece haber generado poco interés por parte de los investigadores y usuarios de tests psicológicos.

Es importante tener presente que los resultados de los instrumentos psicométricos usados en selección de personal pueden emplearse de dos maneras. La primera, menos frecuente, es hacer un ranking de candidatos y seleccionar a la proporción de postulantes con mejor desempeño relativo hasta completar un número finito de vacantes (un ejemplo paradigmático, son las pruebas de selección universitaria). La segunda, mucho más habitual, consiste en definir a priori un punto de corte en la escala de resultados y usarlo para clasificar a los postulantes en categorías discretas (e.g., aceptado o rechazado; aprobado o reprobado; recomendable o no recomendable).

En el primer caso, la fiabilidad puede estimarse con cualquiera de los procedimientos convencionales (e.g., test-retest, bipartición, α de Cronbach), pues la mayoría de los modelos psicométricos que subyacen al desarrollo y uso de tests (i.e., Teoría Clásica de los Tests; Teoría de Respuesta al Ítem, Análisis Factorial, Modelos de Ecuaciones Estructurales) funcionan bajo la premisa de estimar rasgos continuos. La fiabilidad, en el caso convencional, es la consistencia entre las replicaciones de un procedimiento de medición (Gulliksen, 1950). Por ello, la fiabilidad suele definirse en términos operativos como la correlación entre el resultado obtenido por los evaluados y el resultado que obtendrían si volvieran a evaluarse con el mismo instrumento o con un instrumento perfectamente equivalente (i.e., paralelo). Incluso los métodos de consistencia interna, como el coeficiente α de Cronbach, no son sino el límite inferior (*lower bound*) de la correlación entre la prueba y una versión paralela de sí misma (Guttman, 1945).

Una sutileza, a menudo inadvertida, de esta definición correlacional de la fiabilidad es que esta se refiere al ordenamiento relativo o ranking de los evaluados. De este modo, mayor fiabilidad implica mayor probabilidad de que los evaluados vuelvan a ocupar la misma posición en el ranking de resultados, lo que no es equivalente a decir que obtendrán el mismo resultado. A modo de ejemplo, la tabla 1 ilustra los resultados hipotéticos de 10 evaluados en un estudio test-retest. Aunque todos los examinados incrementaron su puntaje en la segunda aplicación, el ordenamiento relativo entre ellos se mantiene estable, lo cual es indicativo de una alta fiabilidad. De hecho, la fiabilidad test-retest en este ejemplo corresponde a $r = .99$.

Tabla 1

Ejemplo hipotético de estudio test-retest en 10 evaluados

Test	15	20	20	32	35	40	40	45	50	60
Re-test	25	29	29	42	45	49	50	55	57	65

Esta definición correlacional de los coeficientes de fiabilidad tradicionales resulta problemática cuando los tests son utilizados para clasificar a los examinados en categorías discretas, a partir de puntos de corte previamente establecidos. En el ejemplo anterior, supongamos que el punto de corte que define la decisión de selección es un resultado igual o superior a 33 puntos. En ese caso, en la primera aplicación seleccionaríamos a seis postulantes, mientras que en la segunda aplicación serían nueve los seleccionados. En otras palabras, pese a la alta consistencia test-retest, en la primera aplicación seleccionaríamos a poco más de la mitad de los evaluados y, en la segunda, a casi todos ellos. Examinando cuidadosamente la tabla 1 podemos ver que solo el primer postulante resultaría no seleccionado en ambas aplicaciones (test = 15 puntos y re-test = 25 puntos) y que seis postulantes obtendrían resultados iguales o mayores al punto de corte en ambas ocasiones. La suma de los casos no seleccionados y seleccionados en ambas ocasiones (n = 1 + 6) respecto del total de casos (n = 10) corresponde a una proporción de .7 casos clasificados consistentemente. Como puede observarse en este simple ejemplo, aun cuando exista una fiabilidad de r = .99 la consistencia de la clasificación resultante puede ser más baja (.70 en el presente caso). Por ello, los procedimientos tradicionales para la estimación de la fiabilidad no resultan apropiados en el caso de instrumentos utilizados para clasificar a las personas en categorías dicotómicas, como lo son las escalas de *screening* o, muy a menudo, los tests de selección de personal.

Entonces, ¿cómo estimar la fiabilidad de la clasificación resultante a partir de un punto de corte en un instrumento psicométrico? Existen varias aproximaciones disponibles (cf., Muñiz, 2001). La más simple de todas, recomendada por Hambleton y Novick (1973), es hacer un estudio test-retest y calcular la consistencia de la clasificación mediante la proporción simple de casos clasificados consistentemente, tal como se hizo en el ejemplo anterior. Una variante más sofisticada de la misma aproximación, propuesta por Swaminathan, Hambleton y Algina (1974), consiste en calcular la proporción de casos clasificados consistentemente corrigiendo el efecto de consistencia al azar, mediante el índice de Kappa de Cohen (1960). El problema de ambos métodos es que requieren dos aplicaciones del mismo instrumento, lo que a menudo resulta difícil o incluso implausible en contextos aplicados.

Debido a ello, en la literatura especializada en Psicometría y medición educativa se han desarrollado varios coeficientes que pueden aplicarse cuando existe una única aplicación del test. La mayoría de ellos (cf., Lee, Hanson y Brennan, 2002) requieren del uso de modelos psicométricos bastante sofisticados, como Teoría de Respuesta al Ítem (TRI) y la llamada Teoría Fuerte de la Puntuación Verdadera (Strong True Score Theory; Allen y Yen, 1979), lo que impide su uso en la mayoría de los tests de selección de personal, que han sido construidos en el marco de la Teoría Clásica de los Test (TCT).

Afortunadamente, en Psicometría también se han desarrollado al menos un par de coeficientes para estimar la fiabilidad de una clasificación a partir de una única aplicación del test. El objetivo de este artículo es presentar un software de distribución gratuita, denominado *Phi(Lambda)*, programado para facilitar el cálculo del coeficiente homónimo desarrollado por Brennan y Kane (1977) en el marco de la Teoría de la Generalizabilidad (Brennan, 2001), con la finalidad de estimar la fiabilidad de una clasificación. El uso del software y la interpretación del coeficiente son demostrados con las respuestas dadas a un test de selección por una muestra real de postulantes que participaron en un proceso masivo de selección de personal. El programa *Phi(Lambda)* es un asistente de cómputo que debe utilizarse en conjunto con el conocido software SPSS (IBM Corp., 2013).

Fundamento teórico del coeficiente Phi(Lambda)

En Psicometría, el concepto de fiabilidad está estrechamente ligado al paradigma de la puntuación verdadera, cuya expresión más conocida, la TCT, constituye el fundamento técnico de la gran mayoría de los tests, escalas, cuestionarios y autoinformes actualmente en uso en la comunidad psicológica. Aunque la Psicometría contemporánea privilegia el paradigma de variables latentes, especialmente bajo la versión de la TRI, la mayoría de los usuarios están más familiarizados con los conceptos de la TCT. Una presentación sintética, en castellano, de los supuestos, fundamentos, limitaciones y modelos de la TCT puede encontrarse en la obra de Muñiz (2001).

En términos simples, la TCT (también conocida como Teoría Débil de la Puntuación Verdadera; *Weak True Score Theory;* Allen y Yen, 1979) postula que la puntuación obtenida por

un sujeto (X) está compuesta de dos elementos: una puntuación verdadera (V) y un error (*e*). Este simple principio se expresa en la conocida ecuación:

$$X = V + e \quad (1)$$

que indica que cualquier puntuación observada contiene error. Por ejemplo, si un niño obtiene una puntuación de 115 en una prueba de inteligencia (X = 115), la ecuación (1) indica que estos 115 puntos contienen su puntuación verdadera de inteligencia más un error. Tal vez la puntuación verdadera del sujeto corresponda a 128 o 94 puntos, pero no tenemos forma de saberlo hasta encontrar una manera de aislar el error y determinar cuánto de la puntuación observada es puntuación verdadera. La TCT se funda en esta distinción y tiene por propósito cuantificar la magnitud del error. El concepto de fiabilidad nace con el objetivo instrumental de permitir la estimación del error de medida.

Haciendo una serie de supuestos matemáticos (cf., Gulliksen, 1950; Muñiz, 2001), la TCT plantea que si los errores fueran aleatorios y no correlacionaran con la puntuación verdadera ni entre sí, la varianza de las puntuaciones observadas en una población de personas evaluadas correspondería a la suma de las varianzas verdaderas y de error, es decir:

$$\sigma^2_X = \sigma^2_V + \sigma^2_e \quad (2)$$

que es simplemente la ecuación (1) expresada en términos de varianzas. A partir de esta ecuación fundacional, la TCT define el concepto de fiabilidad, como la proporción de varianza observada que corresponde a varianza verdadera:

$$\text{fiabilidad} = \frac{\sigma^2_V}{\sigma^2_X} = \frac{\sigma^2_V}{\sigma^2_V + \sigma^2_e} \quad (3)$$

Como se observa en la ecuación (3), la fiabilidad también se puede definir como la proporción de varianza verdadera, respecto a la suma de la varianza verdadera y la varianza de error. Este planteamiento se deriva directamente de la manera en que se define la varianza observada en la ecuación (2) y es muy importante para comprender el origen del coeficiente Phi(Lambda).

Es evidente que en la TCT el concepto de error o de varianza de error juega un papel fundamental en la definición de la fiabilidad. Sin embargo, un problema de la TCT es que no permite distinguir entre las diferentes fuentes de error que pueden afectar el resultado de una medición. Por ejemplo, podría suceder que distintos grupos de postulantes respondan el test en diferentes días, horarios o *settings*, o que algunos protocolos sean corregidos por evaluadores distintos, utilizando criterios idiosincráticos (como suele ocurrir con los tests proyectivos o con las preguntas abiertas de algunos tests de inteligencia, como el *Wechsler Adult Intelligence Scale*, WAIS). Incluso, podría ocurrir que algunos postulantes tengan experiencia previa con el test, que interfiera (favorable o desfavorablemente) con sus respuestas. En todos estos casos, hay múltiples fuentes de error pero la fiabilidad estimada en el marco de la TCT no permite diferenciar entre ellas.

Como una manera de superar esta limitación, varios autores han propuesto el uso de la Teoría de la Generalizabilidad (TG), una extensión de la TCT originada en el contexto de la medición educativa (Brennan, 2001). En lo que sigue intentaremos explicar los conceptos más elementales de la TG, necesarios para comprender el coeficiente Phi(Lambda), reduciendo al mínimo el desarrollo estadístico (para una revisión estadística, cf., Brennan, 2001).

La TG desagrega el error en sus diferentes componentes, mediante Análisis de Varianza (ANOVA). Aunque se trata de un enfoque psicométrico bastante complejo, la idea básica es que la variación observada en las puntuaciones de los ítems o del test puede tener distintos orígenes o "fuentes": la dificultad de los ítems, la habilidad de los evaluados, las condiciones de aplicación, los sesgos de los correctores, el número de mediciones, etc. El objetivo de la TG es cuantificar la proporción de varianza explicada por estas fuentes, con la finalidad de separar entre la varianza originada por las diferencias verdaderas entre las personas evaluadas (i.e., "objeto de medida") de la varianza originada por el resto de las variables, que son denominadas "facetas" de medición en este contexto, y que equivalen a los "factores" de la terminología ANOVA convencional.

El diseño más sencillo de la TG, que equivale al modelo clásico de medida en la TCT, es denominado diseño de una faceta (Brennan, 2001) porque identifica una única faceta de medición (i.e., los ítems). En este diseño, hay tres grandes fuentes de varianza o variabilidad en los resultados. El primer lugar, las diferencias verdaderas entre las personas evaluadas (σ^2_p). En segundo lugar, las diferencias sistemáticas en la dificultad de los ítems (σ^2_i), que representan la varianza explicada por la faceta de medición. En tercer lugar, hay varianza que se explica por la interacción entre las personas y los ítems (para algunas personas algunos ítems pueden resultar más fáciles y para otras más difíciles) y hay varianza no explicada, que se suma a la anterior y constituye la varianza residual del modelo ($\sigma^2_{pi,e}$). Todo lo anterior se puede representar como:

$$\sigma^2(X_{pi}) = \sigma^2_p + \sigma^2_i + \sigma^2_{pi,e} \quad (4)$$

Si se comparan las ecuaciones (2) y (4), se podrá ver que en la TG la varianza verdadera de la TCT corresponde a la varianza originada en las diferencias sistemáticas entre las personas, una vez controladas las otras fuentes de variación (σ^2_p). Sin embargo, es el tratamiento de la varianza error lo que marca la diferencia entre la TG y la TCT y que permitirá estimar la fiabilidad de una clasificación.

En este punto, la TG se diferencia entre dos tipos de varianza de error, según el tipo de decisión que pretenda tomar a partir de los resultados del test. Por una parte, los resultados pueden utilizarse para tomar *decisiones relativas*, es decir, decisiones en que importa el ordenamiento o posición relativa de los examinados en comparación con la población. Este es el caso convencional de los test interpretados de acuerdo a normas o baremos y es también el caso señalado anteriormente en que el test se usa para hacer un ranking de evaluados y seleccionar a aquellos con mejor desempeño. Para este tipo de decisiones, la varianza de error se calcula a partir de la varianza residual ($\sigma^2_{pi,e}$), sin tomar en cuenta las diferencias sistemáticas en dificultad de los ítems (σ^2_i), pues estas últimas no alteran la posición relativa de los examinados en el ranking de resultados.

Sin embargo, los resultados de los test también pueden usarse para tomar *decisiones absolutas*, en las que el foco está puesto en el desempeño de cada evaluado con independencia del resultado de los demás. Un ejemplo de evaluación con decisiones absolutas corresponde a las pruebas de aula convencionales, en donde la calificación final depende del porcentaje de respuestas correctas respecto a un máximo teórico, con independencia del resultado del resto del curso. Otro ejemplo, es el de las listas de cotejo de síntomas, en los que el riesgo de patología aumenta conforme más síntomas presente el paciente, sin referencia a resultados normativos. Para las decisiones de tipo absoluto, las diferencias sistemáticas en dificultad de los ítems (σ^2_i) sí tienen un efecto en el resultado. Por ello, en este caso la varianza de error se calcula a partir de la varianza residual ($\sigma^2_{pi,e}$) y de la varianza originada en los ítems (σ^2_i).

Considerando ambos tipos de decisión y sus respectivas varianzas de error, en la TG se definen dos coeficientes de fiabilidad que son denominados coeficientes de *Generalizabilidad Relativa* y *Absoluta*, convencionalmente representados con las letras *G* y *Phi* (Φ), respectivamente. De este modo, el coeficiente de *Generalizabilidad Relativa* es:

$$G = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{eREL}} \quad (5)$$

Mientras el coeficiente de *Generalizabilidad Absoluta* es:

$$\phi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{eABS}} \quad (6)$$

Como puede apreciarse, ambas ecuaciones son equivalentes a la ecuación (3), que define la fiabilidad convencional y solo difieren entre sí en el tipo de varianza error. Es interesante notar, además, que la fiabilidad convencional también utiliza varianza de error relativo, de modo que el coeficiente *G* es el coeficiente de fiabilidad de la TCT. En Brennan (2001) puede revisarse cómo la ecuación del coeficiente G es simplemente la fórmula del coeficiente α de Cronbach, cuando este es estimado a partir de componentes de varianza.

En el marco conceptual de la TG, en particular en torno a la noción de decisiones absolutas y al coeficiente de *Generalizabilidad Absoluta*, Brennan y Kane (1977) desarrollaron un coeficiente de generalizabilidad especial para el problema de la fiabilidad de una clasificación, al que denominaron Φ(λ) (léase: Phi[Lambda]), es decir, un coeficiente de generalizabilidad para decisiones absolutas (coeficiente Φ), en relación a un punto de corte (λ).

La definición más general de Φ(λ) es:

$$\Phi(\lambda) = \frac{\sigma^2_p + (\mu_{pi} - C)^2}{\sigma^2_p + (\mu_{pi} - C)^2 + \sigma^2_{eABS}} \quad (7)$$

donde μ_{pi} corresponde a la media poblacional de ítems respondidos correctamente, *C* representa el valor del punto de corte y los términos σ^2_p y σ^2_{eABS}, son las varianzas de los sujetos (varianza verdadera) y la varianza de error absoluto, en la nomenclatura de la TG.

Conceptualmente, el coeficiente Φ(λ) definido en la ecuación (7) es simplemente la razón entre la varianza de las puntuaciones verdaderas y observadas (concepto clásico de fiabilidad, ver ecuación [3]), en que cada varianza es "corregida" agregando la distancia cuadrática entre la media y el punto de corte. Se utiliza la distancia cuadrática en lugar de la diferencia simple porque el objetivo es introducir en la ecuación la discrepancia entre el punto de corte y la media grupal, sin importar su signo.

Brennan y Kane (1977) desarrollaron su coeficiente basándose en los trabajos previos de Livingston (1972, 1973), que había argumentado que cuando el punto de corte es distinto del promedio de la muestra, la diferencia entre ambos ($\mu_V - C$) es una fuente de varianza verdadera y, por lo tanto, es necesario incorporarlo al coeficiente de fiabilidad, en la forma propuesta en la ecuación (7). El razonamiento detrás de este procedimiento es que si el punto de corte se aleja del centro de la distribución, la clasificación de los evaluados en dos grupos debería ser más confiable, en términos globales, que la medición con los puntajes originales. Por ello en el coeficiente Φ(λ) la fiabilidad de la clasificación es proporcional al incremento de la fiabilidad de las puntuaciones originales de la prueba y a la distancia entre el punto de corte y la media de la distribución. Obviamente, si el punto de corte se fija en la media de la distribución, la fiabilidad de la clasificación será idéntica a la fiabilidad de la prueba original.

Aunque se han desarrollado otros coeficientes para estimar la fiabilidad de un punto de corte (para una discusión del coeficiente K^2, cf. Gempp y Saiz, 2014), el coeficiente Φ(λ) es superior a otras alternativas. Primero, cuenta con un sustento teórico muy riguroso, está construido explícitamente para evaluar la fiabilidad de puntos de corte y permite contrastar directamente la fiabilidad de la prueba con la fiabilidad de las clasificaciones mediante la comparación entre Φ y Φ(λ). Por otro lado, el uso de la TG permite desagregar y corregir el efecto de las distintas fuentes de error, lo que redunda en resultados más precisos. Además, permite trabajar con diferentes formatos de ítems y, eventualmente, modelar otros efectos, como el de los correctores, en el caso de preguntas abiertas.

La única limitación práctica del coeficiente Φ(λ) es que su cálculo requiere de una operatoria bastante tediosa, pues primero es necesario estimar los componentes de varianza con algún programa estadístico y luego usar una calculadora de bolsillo o planilla de cálculo para estimar los coeficientes, lo que además incrementa las posibilidades de cometer error de cómputo. Por esa razón, se estimó pertinente elaborar un software que pudiera facilitar el cálculo y que operara en conjunto con el ampliamente difundido programa estadístico SPSS (IBM Corp., 2013).

Descripción del programa *Phi(Lambda)* y ejemplo empírico

El software *Phi(Lambda)* es un asistente de cómputo para la estimación del coeficiente homónimo, en el marco del diseño de TG de una faceta. Se trata de una calculadora que opera en ambiente Windows y que trabaja en conjunto con el SPSS (IBM Corp., 2013). De este modo, el usuario debe primero hacer un análisis de varianza para proveer al programa con los insumos necesarios para estimar el coeficiente Phi(Lambda).

En el marco del diseño de TG de una faceta, los componentes de varianza pueden estimarse mediante varias rutinas del SPSS (IBM Corp., 2013). No obstante, por simplicidad, se escogió la metodología de análisis de varianza disponible en la rutina "Reliability" del programa.

Para la demostración se usarán las respuestas dadas por una muestra de 663 postulantes (52.6% mujeres) a una prueba de rendimiento máximo, compuesta por 30 ítems de selección múltiple de cinco alternativas con solo una correcta. La prueba en cuestión forma parte de la batería de instrumentos de selección utilizada por una empresa del área industrial. Como política de la empresa, los candidatos son preseleccionados sobre la base de antecedentes académicos y posteriormente sometidos a una capacitación intensiva de dos días, al cabo de los cuales se aplica la prueba en cuestión. Solo los postulantes que aprueben este instrumento pueden continuar en el proceso. El criterio de selección (punto de corte) corresponde a 60% de rendimiento, es decir, al menos 18 preguntas correctas.

Para comenzar, se estimó la fiabilidad convencional mediante el coeficiente α de Cronbach, utilizando el módulo de estimación de fiabilidad del programa SPSS (IBM Corp., 2013).

Como resultado, se obtuvo un α de Cronbach = .85. Este resultado es equivalente al de *Generalizabilidad Relativa* e indica que si se volviera a evaluar a las mismas personas, se encontraría una correlación de .85 entre su posición relativa en el test y su posición relativa en el re-test. Bajo parámetros convencionales de interpretación de la fiabilidad, el resultado obtenido puede considerarse bueno, aunque no excelente, para la toma de decisiones individuales, pues se encuentra bajo el umbral de .90.

A continuación, se procedió a estimar el coeficiente Phi(Lambda). Para ello, en el mismo módulo de fiabilidad del SPSS se presionó el botón *Estadísticos* para ingresar al cuadro de diálogo. Una vez allí, se seleccionó la opción Prueba F en el recuadro tabla ANOVA.

ANOVA

		Sum of Squares	df	Mean Square	F	Sig
Between People		546,967	662	,826		
Within People	Between Items	229,832	29	7,925	65,581	,000
	Residual	2320,001	19198	,121		
	Total	2549,833	19227	,133		
Total		3096,800	19889	,156		

Grand Mean = .1929

Figura 1. Salida de análisis de varianza de SPSS (IBM Corp., 2013).

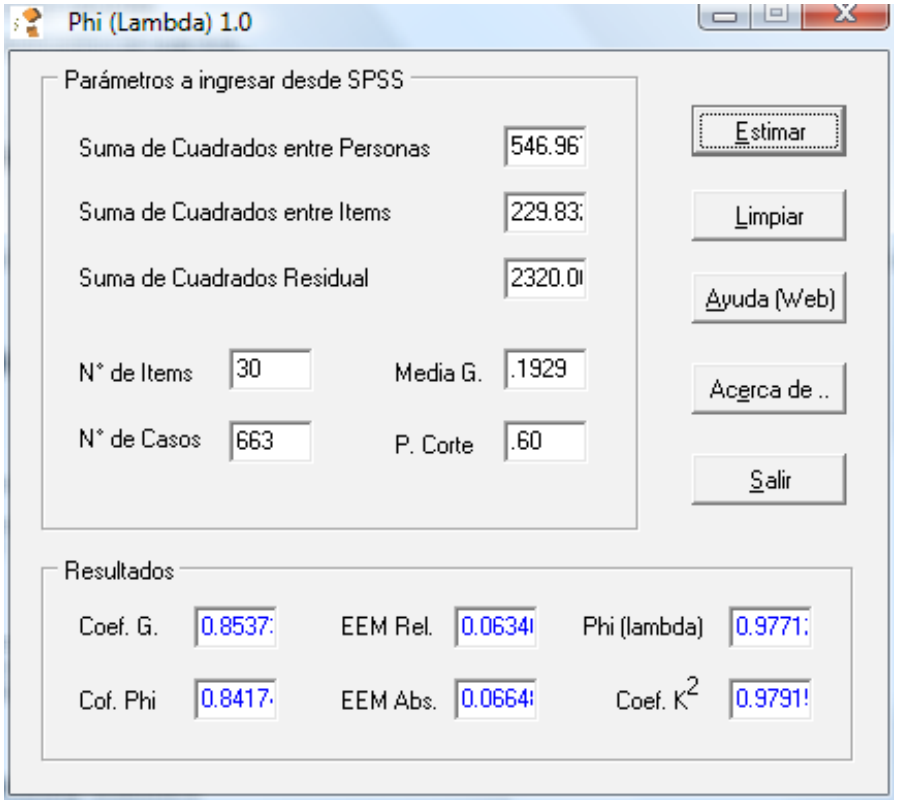

Figura 2. Ingreso de datos al software Phi(Lambda).

Con este procedimiento se obtuvo la tabla de ANOVA presentada en la figura 1, en el formato original arrojado por SPSS. Los resultados de esta tabla fueron usados como insumo (*input*) para el programa *Phi(Lambda)*, cuyo único cuadro de diálogo es presentado en la figura 2. Puede observarse que los datos más importantes son la Suma de Cuadrados entre Personas, Ítems y Residual, que corresponden a la primera columna y las tres primeras líneas de la tabla de resultados de ANOVA presentada en la figura 1. Además, es necesario indicar al programa el número de ítems (i = 30), el número de casos (n = 663), la media global, a veces llamada "gran media" (M = .193) y el punto de corte (C = .60). En este caso, la gran media y el punto de corte están expresados como la proporción de ítems respondidos correctamente.

Una vez ejecutado el programa con el botón *Estimar*, los resultados arrojan los coeficientes de *Generalizabilidad Relativo* (.85) y *Absoluto* (.84), además de sus respectivos errores de medida. Nótese cómo el coeficiente de *Generalizabilidad Relativo* es exactamente el mismo que el α de Cronbach, tal como se había advertido previamente. Respecto a la fiabilidad de la decisión, el coeficiente Phi(Lambda) es igual a .977. Adicionalmente, el programa entrega el coeficiente K^2 de Livingston (1972, 1973) (para mayores antecedentes sobre K^2, cf. Gempp y Saiz, 2014).

Conclusiones

En este trabajo hemos revisado el coeficiente Phi(Lambda) desarrollado por Brennan y Kane (1977) para estimar la fiabilidad de un punto de corte en el marco de la Teoría de la Generalizabilidad (Brennan, 2001). Además, hemos presentado una pequeña herramienta informática, disponible gratuitamente, para facilitar el cómputo del coeficiente en conjunto con SPSS (IBM Corp., 2013).

Tal como esperamos haber demostrado, la TG ofrece un marco conceptual mucho más sólido y flexible que la TCT, además de herramientas psicométricas más útiles para problemas como el planteado en el artículo (estimación de la fiabilidad de un punto de corte en selección de personal). En tal sentido, el software presentado también puede ser útil como primera aproximación a la TG por parte de usuarios o estudiantes que no estén familiarizados con este enfoque.

Para seguir profundizando, el texto de Brennan (2001) es una referencia obligada.

Respecto al coeficiente Phi(Lambda), es interesante observar que la fiabilidad de una clasificación habitualmente será más alta que la fiabilidad de una puntuación, especialmente entre más distante se encuentre el punto de corte de la media de la escala. Esto supone que una escala o test puede arrojar resultados de clasificación muy fiables incluso si la fiabilidad de los puntajes es baja según estándares convencionales. Conceptualmente, el coeficiente Phi(Lambda) puede interpretarse como el porcentaje de casos que serían clasificados en la misma categoría si respondieran nuevamente el mismo test. Al igual que los coeficientes de fiabilidad tradicionales, valores iguales o superiores a .75 pueden considerarse aceptables.

Por otro lado, es preciso advertir algunas limitaciones del coeficiente. La primera es que solo es útil en el caso de clasificaciones dicotómicas. Otra limitación es que al estar basado en un modelo de consistencia interna, su interpretación no resulta tan sencilla como en el caso de los coeficientes basados en consistencia entre pruebas paralelas, que se han desarrollado más recientemente (Lee et al., 2002), cuyo cálculo es más complejo, aunque útil en el contexto de la Teoría de Respuesta al Ítem.

Finalmente, respecto a la disponibilidad del software presentado, este incluye un breve manual y puede obtenerse contactando a la dirección electrónica del autor.

Referencias

Allen, M. J. y Yen, W. M. (1979). *Introduction to measurement theory*. Illinois: Waveland Press.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Brennan, R. L. y Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*, 277-289. doi: 10.1111/j.1745-3984.1977.tb00045.x

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46. doi: 10.1177/001316446002000104

Farr, J. L. y Tippins, N. T. (2010). *Handbook of employee selection*. NY: Routledge.

Gempp, R. y Saiz, J. L. (2014). El coeficiente K^2 de Livingston y la fiabilidad de una clasificación dicotómica en un test psicológico. *Universitas Psychologica*, *13*(1), 217-226. doi:10.11144/Javeriana.UPSY13-1.eckl

Gulliksen, H. (1950). *Theory of mental tests*. NY: Wiley.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255-282. doi: 10.1007/BF02288892

Hambleton, R. K. y Novick, M. R. (1973). Toward an integration of theory and method for criterion referenced tests. *Journal of Educational Measurement*, *10*, 159-170. doi:10.1111/j.1745-3984.1973.tb00793.x

IBM Corp. (2013). IBM SPSS Statistics for Windows (Version 22.0) [*Software* de computación] Nueva York: IBM Corp.

Lee, W. C., Hanson, B. A. y Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*, 412-432. doi: 10.1177/014662102237797

Livingston, S. A. (1972). Criterion-referenced applications of Classical Test Theory. *Journal of Educational Measurement*, *9*, 13-21. doi: 10.1111/j.1745-3984.1972.tb00756.x

Livingston, S. A. (1973). A note on the interpretation of the criterion-referenced reliability coefficient. *Journal of Educational Measurement*, *10*, 311. doi: 10.1111/j.1745-3984.1973.tb00809.x

Muñiz, J. (2001). *Teoría clásica de los test*. Madrid: Pirámide.

Putka, D. J. y Sackett, P. R. (2010). Reliability and validity. En J. L. Farr y N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 9-49). Nueva York: Routledge.

Ryan, A. M. y Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology, 65*, 693-717. doi: 10.1146/annurev-psych-010213-115134

Sackett, P. R. y Lievens, F. (2008). Personnel selection. *Annual Review of Psychology, 59*, 419-450. doi: 10.1146/annurev.psych.59.103006.093716

Swaminathan, H., Hambleton, R. K. y Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, *11*, 263-268. doi: 10.1111/j.1745-3984.1974.tb00998.x

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4ª Ed.). Bowling Green, OH: Autor.

Fecha de recepción: 20 de marzo de 2014
Fecha de aceptación: 18 de junio de 2014