El coeficiente K2 de Livingston y la fiabilidad de una decisión dicotómica en un test psicológico

René Gempp / José L. Saiz

doi:10.11144/Javeriana.UPSY13-1.eckl

Publicación preliminar

El comité editorial de *Universitas Psychologica: Panamerican Journal of Psychology* ha evaluado, revisado y aceptado el presente artículo según los criterios de la revista y siguiendo el proceso de revisión doble ciego por pares académicos externos. De la misma forma, este artículo ha iniciado el proceso de edición y la presente versión ya cuenta con la revisión de estilo final con la que aparecerá en la versión electrónica, faltando únicamente el proceso de diagramación.

Con miras a facilitar y agilizar el proceso de publicación de los contenidos de la revista, el presente artículo se publica de forma anticipada. En este punto, puede ser utilizado para su lectura, consulta, citación o distribución y deberá ser citado tal como aparece a continuación.

Finalmente, tenga en cuenta que la versión electrónica final reemplazará esta versión del archivo y se actualizarán los metadatos asociados al mismo en los diferentes sistemas de información.

Online First

The Editorial Board of *Universitas Psychologica: Panamerican Journal of Psychology* has evaluated, reviewed and accepted the article herein in accordance with the establised criteria and by means of double-blind peer review performed by external scholars. This article has likewise started the editing process. The following version has been object of copyediting; layout and proofreading are the only pending processes.

This journal makes use of Online First procedures in order to ease and fasten content release. This version can be used for reading, documentation, citing and distribution purposes, and should be cited as indicated.

Please regard this as a preliminary document that will be replaced by the final version in due time. All metadata sets related to this contribution will be updated in every applicable index and abstract service.

Para citar este artículo / To cite this article: Gempp, R. & Saiz, J. L. (2014). El coeficiente K2 de Livingston y la fiabilidad de una decisión dicotómica en un test psicológico. *Universitas Psychologica*, 13(1). doi:10.11144/Javeriana.UPSY13-1.eckl

El coeficiente *K2* de Livingston y la fiabilidad de una decisión dicotómica en un test psicológico*

Livingston *K2* Coefficient and the Reliability of a Dichotomous Decision in a Psychological Test

Recibido: diciembre 31 de 2012 | Revisado: abril 1 de 2013 | Aceptado: abril 5 de 2013

René Gempp**
Universidad Diego Portales, Santiago, Chile
José L. Saiz***
Universidad de La Frontera, Temuco, Chile

Para citar este artículo: Gempp, R. & Saiz, J. L. (2014). El coeficiente *K2* de Livingston y la fiabilidad de una decisión dicotómica en un test psicológico. *Universitas Psychologica, 13*(1). doi:10.11144/Javeriana.UPSY13-1.eckl

Resumen

La fiabilidad de los puntajes es una de las propiedades psicométricas más importantes de un test psicológico. Sin embargo, a menudo los test son utilizados para hacer clasificaciones dicotómicas de personas, como sucede en las pruebas de *screening* psicopatológico o en selección de personal. En esos casos, los coeficientes de fiabilidad convencionales no resultan apropiados para estimar la precisión de los puntajes. En este trabajo se presenta el coeficiente *K2* de Livingston (1972, 1973) y se demuestra su uso a través de dos ejemplos empíricos, para estimar la fiabilidad de una clasificación realizada a partir de un test psicológico.

Palabras clave autores

Fiabilidad, clasificación, *screening*, selección.

---

* Artículo de investigación.

** Correspondencia relativa a este trabajo puede ser enviada a: René Gempp, Facultad de Economía y Empresa, Universidad Diego Portales, Manuel Rodríguez Sur 253, antiago, Chile. Fono: (56 2) 26762274. Web: www.gempp.org. E-mails: rene.gempp@udp.cl; rgempp@gmail.com

*** Departamento de Psicología. E-mail: jsaiz@ufro.cl

Abstract

The reliability of test scores is one of the most important psychometric properties of a psychological test. However, the tests are often used for dichotomous classifications of people, as in tests used for screening or recruitment purposes. In such cases, the conventional reliability coefficients are not suitable for estimating the accuracy of the scores. This paper introduces the coefficient *K2* of Livingston (1972, 1973) and demonstrates its use through two empirical examples to estimate the reliability of a classification based on psychological tests.

Key words authors

Reliability, classification, screening, recruitment.

En el campo aplicado, especialmente en el diagnóstico clínico y selección de personal, los test psicológicos son empleados frecuentemente para clasificar a las personas en una categoría diagnóstica discreta, más que para estimar la cuantía de un atributo o rasgo psicológico. Esto sucede, por ejemplo, cuando se utilizan instrumentos de *screening* psicopatológico, los cuales, por definición, procuran decidir si un sujeto tiene o no un trastorno o rasgo específico y no la magnitud cuantitativa de este. De hecho, cuando un psicólogo clínico aplica una medida de *screening* de depresión, suele estar más interesado en determinar si su paciente pertenece o no al grupo de sujetos "depresivos", que en estimar "cuánta" depresión, tiene esa persona.

En una situación parecida, cuando se aplican pruebas de inteligencia o habilidades para seleccionar personal, a menudo el objetivo es determinar qué evaluados se encuentran por sobre o por debajo de algún puntaje de corte previamente definido, que definirá las decisiones de contratación. Un tercer ejemplo aparece cuando se utilizan instrumentos compuestos por varias subescalas (p. ej., MMPI-2, MACI, NEOPI-R), en donde la combinación de puntajes altos y bajos conforman perfiles que resultan más informativos del tipo de trastorno o personalidad de un sujeto (p. ej., "paranoia", "personalidad antisocial") que el puntaje de las subescalas en sí mismo.

En todos los casos anteriores, el uso de test psicológicos para clasificar a los evaluados en dos categorías discretas supone dos requerimientos psicométricos: fiabilidad y validez de la clasificación.

El problema de la validez de la clasificación, es decir, el grado en el cual el punto de corte utilizado permite clasificar correctamente a los evaluados, ha sido relativamente

bien resuelto mediante el uso de Curvas de Características Operantes del Receptor (Swets, 1988), denominado análisis ROC (por sus iniciales en inglés, Receiver Operating Characteristics). Esta técnica se originó en la Teoría de Detección de Señales, durante la Segunda Guerra Mundial, y desde entonces se ha diseminado velozmente en diversas áreas de la medicina clínica por su utilidad para evaluar pruebas diagnósticas. En psiquiatría, fue introducida a fines de la década de 1980 (p. ej., Mari & Williams, 1985; Mossman & Somoza, 1989; Murphy et al., 1987) hasta convertirse en una técnica ampliamente utilizada en la actualidad, tanto para seleccionar un punto de corte óptimo como para evaluar su validez concurrente.

Consiste básicamente en contrastar el resultado de la prueba contra un criterio estándar (habitualmente un diagnóstico clínico), y evaluar la sensibilidad y especificidad que alcanzaría el test con distintos puntos de corte. Formalmente, la sensibilidad se operacionaliza como la probabilidad de identificar correctamente un caso con el trastorno, mientras que la especificidad se expresa como la probabilidad de no detectar como caso clínico a quienes efectivamente no padecen el trastorno. Una revisión relativamente exhaustiva de la metodología ROC aplicada al análisis clínico puede encontrarse en Swets y Pickett (1982).

Por otro lado, el problema de la fiabilidad de una clasificación (el grado en el cual la clasificación es consistente o replicable) ha merecido bastante menos atención en psicología, al punto que muchos investigadores ignoran cómo estimar la fiabilidad de una clasificación o sencillamente no están al tanto de su importancia. Este desconocimiento probablemente se explique porque la mayoría de la tecnología psicométrica que subyace al desarrollo y uso de test psicológicos (i. e. Teoría Clásica de los Test; Teoría de Respuesta al Ítem, Análisis Factorial, Modelos de Ecuaciones Estructurales) funciona bajo la premisa de estimar puntuaciones continuas y no clasificaciones discretas. Afortunadamente, en investigación educativa se han desarrollado coeficientes de fiabilidad apropiados para este objetivo, que pueden ser fácilmente aplicados en el caso de instrumentos de *screening*, selección de personal o diagnóstico clínico. El objetivo de este breve trabajo es presentar y demostrar mediante un par de ejemplos empíricos el uso de uno de estos coeficientes, en concreto, el *K2* desarrollado por Livingston (1972, 1973). A continuación se presentan someramente los fundamentos conceptuales del coeficiente y posteriormente se ilustra su aplicación mediante dos ejemplos.

En psicometría, las metodologías para calcular la fiabilidad de las clasificaciones realizadas con instrumentos psicométricos se derivan directamente de la concepción de fiabilidad acuñada en la Teoría Clásica de los Test (TCT), también conocida como Teoría Débil de la Puntuación Verdadera. Como se recordará, en esta teoría psicométrica se propone que cualquier Puntuación Observada (X) es el resultado aditivo de la Puntuación Verdadera (V) de habilidad (o conocimientos o rasgo) del evaluado más un monto de error de medida (e), lo que se expresa en la conocida ecuación $X = V+e$. En la TCT la Puntuación Verdadera se *define* como la esperanza matemática de la Puntuación Observada, lo que equivale a afirmar que si un mismo individuo fuera evaluado infinitas veces con el mismo test, y bajo el supuesto adicional de que el resultado de cada evaluación es independiente entre sí, la media de los errores de medida sería cero y, por tanto, la Puntuación Verdadera correspondería al promedio de la distribución de las infinitas Puntuaciones Observadas.

Introduciendo algunos supuestos adicionales, los teóricos de la TCT demuestran que la misma relación es válida para las varianzas de las puntuaciones observadas, verdaderas y de error, cuando el test es aplicado a una población de evaluados, es decir:

$$\sigma_X^2 = \sigma_V^2 + \sigma_e^2 \qquad (1)$$

A partir de la Ecuación 1, la TCT define *conceptualmente* la confiabilidad de un test como la proporción de varianza verdadera respecto a la varianza observada, es decir, el coeficiente de determinación de la correlación entre las puntuaciones observada y verdadera, en la forma:

$$confiabilidad = \frac{\sigma_V^2}{\sigma_X^2} = \rho_{XV}^2 \qquad (2)$$

Como esta formulación no permite el cálculo empírico de la confiabilidad, la TCT introduce la noción de *paralelismo*. Sin profundizar en detalles, dos test se consideran paralelos si miden exactamente lo mismo, es decir, si tienen la misma puntuación verdadera (V) en la base. Al asumir que dos test son perfectamente paralelos[1], la

[1] En rigor, la TCT incluye varios modelos distintos. El más estricto (y simplista) asume paralelismo entre las pruebas, pero otros modelos asumen variaciones más relajadas del supuesto de paralelismo.

covarianza o correlación entre sus Puntuaciones Observadas necesariamente se explica por la Puntuación Verdadera común entre ambos, de lo cual se deduce que la correlación entre sus puntuaciones es un indicador directo de la fiabilidad.

Por ello, en términos prácticos, el coeficiente de fiabilidad se define *empíricamente* como la correlación entre dos test paralelos ($\rho_{xx'}$). Bajo esta lógica, se desarrolló el primer método de estimación de la fiabilidad, que consistía en construir una forma paralela del test y aplicarlo en conjunto con este para determinar directamente su correlación. Las evidentes limitaciones prácticas de este método restringen su aplicabilidad y estimularon el desarrollo de otros procedimientos. Uno de ellos es el método test-retest, en que las formas "paralelas" se definen como la aplicación anterior y posterior de un mismo test, de manera que la correlación antes-después opera como coeficiente de fiabilidad. Por último, para los casos en que se cuenta con una única aplicación del test, se desarrollaron métodos de consistencia interna, trasladando el foco del análisis a los subcomponentes del test (i.e. ítems o grupos de ítems). Asumiendo que cada componente es paralelo respecto a los demás, la covarianza entre ellos actúa como un coeficiente de confiabilidad. Existen básicamente dos variantes de esta aproximación. En la bipartición, el test se divide en dos partes supuestamente paralelas y se calcula la correlación entre ambos como estimador de la confiabilidad. En las técnicas de multipartición, cada ítem se asume como un subtest paralelo al resto y se calculan coeficientes de consistencia como estimadores de la fiabilidad. Entre estos, el alfa de Cronbach (que además equivale al promedio de todas las biparticiones posibles) se convirtió en el más popular. Debido a ello, hoy en día es habitual que la fiabilidad se defina como un indicador del grado de *consistencia* de una medición, a través del tiempo (método test-retest), a través de los subcomponentes (i.e. ítems) que la conforman (métodos de consistencia interna) o a través de sus formas paralelas (método clásico de formas paralelas).

Aplicando estos conceptos, se han planteado cuando menos dos aproximaciones generales para resolver el problema de la fiabilidad de una clasificación. La primera utiliza la noción original de paralelismo y simplemente define la fiabilidad de una clasificación como la proporción de evaluados clasificados en la misma categoría al utilizar dos formas paralelas. La segunda consiste en extender la noción tradicional de fiabilidad como consistencia interna de las puntuaciones, al caso particular en que las

puntuaciones se segmentan para construir categorías. Aunque la primera aproximación resulta conceptualmente atractiva, tiene el costo de requerir la aplicación de una forma paralela al test, lo que en muchos casos resulta imposible. Para ello, se han desarrollado soluciones en el marco de la Teoría de Respuesta al Ítem que consisten en construir una prueba paralela hipotética, predicha a partir del resultado empírico de la prueba (Lee, Hanson & Brennan). El problema de estas soluciones es que funcionan solo en el marco de la Teoría de Respuesta al Ítem y resultan, por ello, poco amigables para los investigadores aplicados. En cambio, la aproximación basada en el principio de consistencia interna resulta muy fácil de estimar, como se explica a continuación.

Este grupo de métodos ofrece la ventaja de requerir solo una aplicación de la prueba y su interpretación es equivalente a la de un índice de consistencia interna tradicional. Aunque hay varios procedimientos disponibles (Brennan & Kane, 1977; Harris, 1972; Livingstone, 1972, 1977), nos concentraremos en el coeficiente *K2* propuesto por Samuel Livingstone, pues se trata del más conocido y mejor estudiado. Originalmente, fue concebido como un coeficiente de fiabilidad para una clasificación en dos categorías, basado directamente en la noción fundacional de fiabilidad de la TCT, que la entiende como la razón entre la varianza verdadera y observada. Su formulación es:

$$K^2 = \frac{\sigma_v^2 + (\mu_V - C)^2}{\sigma_X^2 + (\mu_X - C)^2} \qquad (3)$$

donde σ_V^2 y σ_X^2 corresponden respectivamente a la varianza de las puntuaciones verdaderas y observadas, μ_V y μ_X son las medias poblacionales verdadera y observada en la prueba y C es el punto de corte.

Conceptualmente, el coeficiente *K2* definido en la Ecuación 3 es simplemente la razón entre la varianza de las puntuaciones verdaderas y observadas (concepto clásico de fiabilidad, ver Ecuación 2), 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.

Livingston (1972, 1973) argumenta 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 confiabilidad, en la

forma propuesta en la Ecuación 3. 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 *K2* la confiabilidad de la clasificación es proporcional al incremento de la confiabilidad 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 confiabilidad de la clasificación será idéntica a la confiabilidad de la prueba original.

Dado que la Ecuación 3 no permite la estimación empírica del coeficiente, Livingston demostró que este puede calcularse bajo el supuesto habitual de la TCT de que la media poblacional de las puntuaciones verdaderas y observadas es similar, y utilizando un coeficiente de consistencia interna para equiparar la varianza de la puntuación verdadera con la varianza de la puntuación observada, es decir:

$$K^2 = \frac{\alpha\sigma_X^2 + (\mu_X - C)^2}{\sigma_X^2 + (\mu_X - C)^2} \qquad (4)$$

donde α es el coeficiente de consistencia interna alfa de Cronbach.

Como se desprende de la revisión de la Ecuación 4, los datos necesarios para calcular el *K2* son la varianza y promedio de los puntajes del test, su alfa de Cronbach y, obviamente, el valor del punto de corte. Examinando cuidadosamente la Ecuación 4, se puede concluir que el coeficiente de Livingston aumenta en la medida que la consistencia interna tiene una magnitud más alta, y que su valor será exactamente igual al alfa de Cronbach si el punto de corte se fija en la media de la distribución. En otras palabras, las estimaciones de *K2* serán siempre iguales o mayores al alfa de Cronbach y se encontrarán acotadas en el rango entre 0 y 1.

El coeficiente *K2* tiene varias características atractivas: cuenta con un fundamento conceptual sólido, es fácil de calcular y requiere de solo una aplicación de la prueba. A cambio, presenta la limitación de que solo es útil para clasificaciones en dos niveles, pero ello corresponde a la típica situación en pruebas de *screening*.

A continuación, se presentan dos ejemplos empíricos para demostrar la aplicación del coeficiente.

Ejemplo 1: Fiabilidad de los puntos de corte de la CES-D

El primer ejemplo corresponde a la aplicación del coeficiente *K2* a la Escala de Depresión del Centro para Estudios Epidemiológicos (Center for Epidemiologic Studies Depression Scale [CES-D]), un autoinforme breve (Radloff, 1977) diseñado para el tamizado rápido de sintomatología depresiva en población general. El ejemplo se basa en un estudio publicado recientemente por Gempp y Thieme (2010) en el que compararon las propiedades psicométricas y puntos de corte para cuatro modalidades de puntuación de la escala, utilizando una muestra normativa de 1.143 jóvenes chilenos, no consultantes, de ambos sexos (45.7 % de hombres y 54.3 % de mujeres), con una edad promedio de 20.56 años (*DE* = 4.45). Para establecer puntos se corte se utilizó una muestra clínica de 44 pacientes. En este ejemplo, se analizarán solo los datos de la muestra normativa.

En su versión original, la CES-D consta de 20 ítems, cada uno de los cuales corresponde a un síntoma habitual y representativo del trastorno depresivo. Las instrucciones solicitan al respondiente indicar la frecuencia con la que se experimentó cada síntoma "Durante la semana pasada", utilizando una escala de cuatro alternativas acotadas por las frases: *Rara vez o ninguna vez* (1 día o menos), *Alguna vez o unas pocas veces* (1 a 2 días), *Ocasionalmente o varias veces* (3 a 4 días) *y La mayor parte del tiempo* (5 a 7 días). El método convencional y más utilizado de corrección (modalidad "ordinal"), asigna desde 0 a 3 puntos a cada alternativa, intentando valorar la presencia y gravedad de cada síntoma (Radloff & Locke, 1986), de manera que a mayor puntuación, mayor frecuencia de ocurrencia del síntoma. La puntuación total, por lo tanto, se calcula como la sumatoria simple de los ítems, pudiendo variar entre 0 a 60 puntos.

Por otro lado, Craig y Van Natta (1976, 1979) propusieron una modalidad de puntuación por "presencia" de los síntomas, en que se asigna 0 puntos a la primera alternativa (*Rara vez o ninguna vez* (1 día o menos)), indicativa de ausencia de sintomatología y 1 punto a las restantes opciones. En este caso, la puntuación total de la escala puede variar entre 0 a 20 puntos, y representa el número de síntomas que se experimentaron al menos una vez durante la semana pasada. Los mismos autores también propusieron una modalidad de puntuación por "persistencia" en la cual se asigna 1 punto a la alternativa *La mayor parte del tiempo* (5 a 7 días) y 0 puntos a las demás opciones. La puntuación total, que puede variar entre 0 a 20 puntos, se interpreta en este caso como la cantidad de síntomas que aparecieron repetidamente durante la semana.

Finalmente, McArdle, Johnson, Hishinuma, Miyamoto y Andrade (2001) propusieron una modalidad de puntuación "semanal", en que cada alternativa es transformada en el número de días en que se experimentó el síntoma depresivo. Por ejemplo, la opción *Alguna vez o unas pocas veces* (1 a 2 días) es valorada con 1.5 puntos (entre 1 a 2 días). En el caso de los ítems con puntuación inversa, los ítems reflejan "días sin depresión", así que las alternativas son transformados a "días con depresión" asignando 6.5, 5.5, 3.5 y 1 puntos. La puntuación total se calcula como el promedio de los 20 síntomas y se interpreta, sencillamente, como el número promedio de días en la semana recién pasada, en que se experimentaron síntomas depresivos (McArdle et al., 2001).

El objetivo del trabajo de Gempp y Thieme (2010) fue determinar los puntos de corte para cada modalidad de puntuación. El objetivo del presente análisis es determinar la fiabilidad de la clasificación diagnóstica que arrojaba cada modalidad, mediante el coeficiente *K2* . Adicionalmente, se pretende demostrar que el coeficiente puede calcularse fácilmente utilizando información secundaria. Para ello, se ha recuperado la información necesaria desde las tablas 2 y 4 del trabajo de Gempp y Thieme (2010, pp. 8-10), específicamente, puntajes de corte, promedios y desviaciones típicas para cada una de estas cuatro modalidades de puntuación. Aplicando la Ecuación 4 a estos datos, se pueden obtener los coeficientes *K2* presentados en la Tabla 1 y la función de fiabilidad graficada en la Figura 1, para la modalidad de puntuación ordinal.

Tabla 1

Coeficientes K2 de Livingston aplicados a los datos de Gempp y Thieme (2010)

Modalidad	alfa	Media	Varianza	P. Corte	*K2*
Ordinal	0.87	16.75	100	24	0.91
Presencia	0.84	10.52	22.7529	12	0.85
Persistencia	0.77	1.62	5.4756	3	0.83
Semanal	0.86	1.95	0.7569	2.74	0.92

Fuente: elaboración propia.

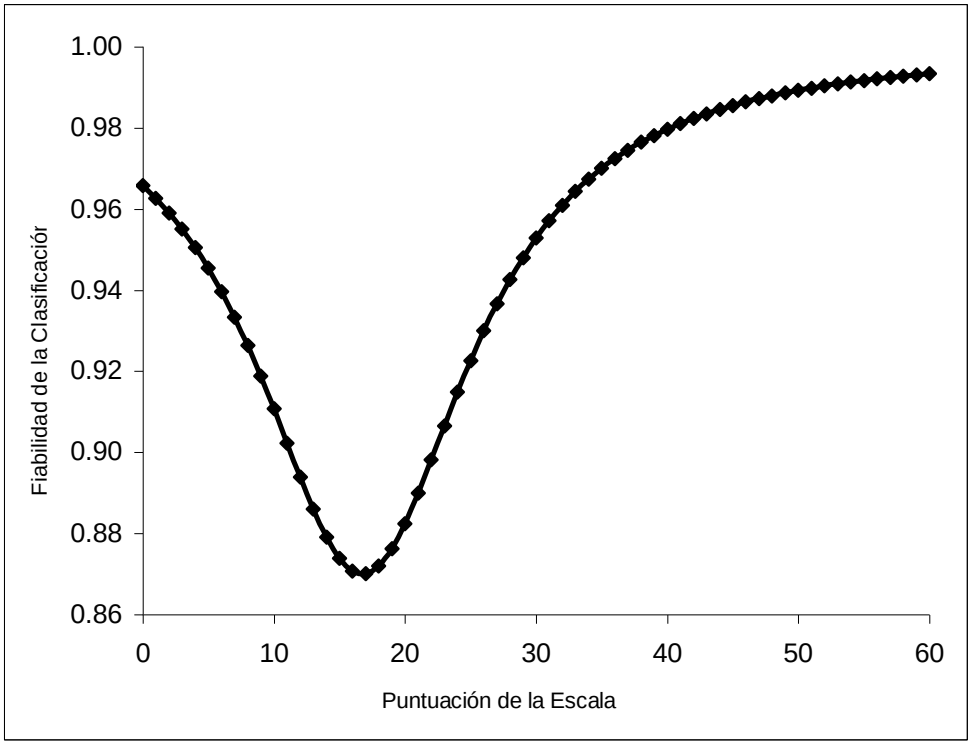

Figura 1. Coeficiente *K2* para distintos puntos de corte de la CES-D.

Fuente: elaboración propia.

Examinando la Tabla 1 se puede observar que, tal como se anticipó, los coeficientes *K2* son sistemáticamente más altos que el alfa de Cronbach, en la medida en que el punto de corte se aleja del promedio de la distribución. En otras palabras, que la fiabilidad o consistencia de la decisión dicotómica (tener o no tener depresión) es mayor que la fiabilidad de un puntaje específico.

La misma conclusión puede extraerse del análisis de la Figura 1 que muestra los coeficientes *K2* que se obtendrían para diferentes puntos de corte de la modalidad de puntuación ordinal de la CES-D.

Podría argumentarse que los elevados valores obtenidos para el *K2* son, hasta cierto punto, esperables dado que confiabilidades tradicionales (alfa de Cronbach) también son bastante altas según los cánones convencionales de análisis psicométrico. El siguiente ejemplo tiene como propósito demostrar que los valores del *K2* serán normalmente muy altos aún cuando los respectivos coeficientes alfa de Cronbach sean de reducida magnitud.

Ejemplo 2: Fiabilidad de la selección de estudiantes de psicología

El segundo ejemplo demuestra la aplicación del coeficiente *K2* a un caso de selección mediante test psicológicos. Se ha extraído un ejemplo desde un contexto no convencional de selección, básicamente porque los datos se prestan muy bien para ilustrar también la aplicación del coeficiente *K2* en casos donde la fiabilidad según alfa de Cronbach es baja[2].

Para contextualizar este ejemplo, es importante explicar que durante muchos años fue una práctica común en la formación de pregrado de psicólogos en Chile exigir exámenes de aptitud o compatibilidad psicológica a los postulantes. La mayoría de los postulantes era estudiantes egresados de educación secundaria, de ambos géneros, con un promedio de edad aproximado de 18 años. Estos exámenes normalmente incluían baterías de test y entrevistas personales con el objetivo de descartar a aquellos que estuvieran en riesgo de padecer cuadros psicopatológicos graves o trastornos de personalidad. Aunque se trató de una práctica muy común hasta fines de 1990, fue abandonada progresivamente en la década del 2000, principalmente por una serie de cambios en el sistema universitario, el aumento exponencial de escuelas de psicología en el país y un fuerte número de críticas tanto a la legitimidad como a la validez teórica y psicométrica del procedimiento.

En esa última línea de críticas, Gempp y Vinet (1997) hicieron un estudio para determinar la fiabilidad del Inventario Multifásico de Personalidad de Minesotta (MMPI) en una muestra de 1.108 estudiantes (51 % eran mujeres) que postularon a la carrera de Psicología en una universidad chilena entre la década de 1990. La muestra incluye a la totalidad de los aspirantes registrados pero por resguardos de confidencialidad no se hace un análisis por separado para cada año. Cabe destacar que se trataba de la versión original del MMPI (Hathaway & McKinley, 1943) pues era la versión del instrumento en ese momento vigente en Chile y uno de los instrumentos más utilizados para selección de postulantes a las carreras de Psicología en el país. Entre los hallazgos más importantes de ese estudio estuvo el que la fiabilidad de varias de las escalas clínicas del MMPI era inferior al límite convencionalmente aceptado de 0.7 y que, por lo tanto, algunas de las escalas no arrojaban resultados fiables. Considerando que la

[2] Los datos que se presentan en este ejemplo tienen más de una década de antigüedad y no representan un hallazgo original. Sin embargo, los hemos escogido por su valor didáctico y para demostrar cómo el uso de herramientas inapropiadas puede llevar a conclusiones equivocadas.

fiabilidad es un requisito necesario aunque no suficiente de la validez (Gulliksen, 1950), los autores razonaron que las escalas poco fiables tampoco podrían ser válidas y, en consecuencia, concluyeron que el MMPI no era un instrumento válido para seleccionar estudiantes de Psicología.

Es importante notar, sin embargo, que Gempp y Vinet (1997) basaron su análisis y conclusiones en los alfa de Cronbach para cada escala, en circunstancias que lo más recomendable habría sido determinar la fiabilidad de la clasificación. En efecto, el criterio convencional utilizado en esos años en la selección de estudiantes de psicología con el MMPI era descartar como sospechoso de riesgo psicopatológico a aquellos postulantes que presentaran una puntuación *T* mayor o igual que 70 en cualquiera de las diez escalas clínicas del instrumento, utilizando como referencia los datos normativos chilenos (Rissetti, Himmel, Maltes, González & Olmos, 1989). En otras palabras, al igual que en muchos procesos de selección, se fijaba un puntaje de corte equivalente a $T = 70$ y a partir de allí se clasificaba a los postulantes en dos categorías excluyentes (aceptado/rechazado). Como el MMPI utilizaba puntuaciones *T* con media de 50 y desviación estándar de 10 puntos, el valor $T = 70$ reflejaba el hecho de encontrarse a dos desviaciones estándar sobre el promedio normativo, que es uno de los criterios más usuales para fijar puntos de corte en test psicológicos.

En un intento por demostrar el uso del coeficiente *K2* en contextos de selección y, simultáneamente, en casos donde los alfa de Cronbach sean de baja magnitud, se ha calculado la fiabilidad de la clasificación para el estudio de Gempp y Vinet (1997) utilizando datos secundarios disponibles en el trabajo de estos autores. Específicamente, se han usado los alfa de Cronbach, promedios y desviaciones estándar reportados para cada escala y utilizado como punto de corte los valores brutos correspondientes a $T = 0$, según los datos normativos vigentes en ese momento (Rissetti et al., 1989). Los resultados se presentan en la Tabla 2.

Tabla 2

Coeficientes K2 de Livingston aplicados a los datos de Gempp y Vinet (1997)

	Hombres					Mujeres				
Escalas	alfa	Media	Var.	P.Corte	K2	alfa	Media	Var.	P.Corte	K2
Validez										
L	0.7	5.52	7.49	11	0.94	0.71	5.63	7.81	11	0.94
F	0.74	4.49	13.88	15	0.97	0.71	4.72	13.3	13	0.95
K	0.72	17.74	18.28	23	0.89	0.71	17.7	18.27	24	0.91
Clínicas										
Hs+K	0.71	13.3	15.22	21	0.94	0.62	13.28	11.4	24	0.97
D	0.55	20.67	18.36	31	0.93	0.46	20.91	15.74	33	0.95
Hy	0.53	21.63	18.29	30	0.9	0.42	21.7	15.15	33	0.94
Pd+K	0.45	23.33	14.84	31	0.89	0.47	23.37	16.14	31	0.88
Mf	0.46	32.62	19.67	35	0.58	0.46	33.29	20.16	43	0.9
Pa	0.45	9.78	9.24	17	0.92	0.33	9.87	7.62	17	0.91
Pt+K	0.3	27	18.68	39	0.92	0.2	27.04	17.83	40	0.92
Sc+K	0.35	27.75	29.99	42	0.92	0.3	27.85	28.2	40	0.89
Ma+K	0.44	21.77	14.24	30	0.9	0.43	21.7	13.88	29	0.88
Si	0.83	21.12	61.6	43	0.98	0.79	21.23	52.9	45	0.98

Fuente: elaboración propia.

El examen cuidadoso de la Tabla 2 muestra que, tal como se anticipó, los coeficientes *K2* tienden a ser bastante elevados incluso en los casos donde el alfa de Cronbach es bajo, con la única excepción de la Escala Mf para los varones. En conjunto, estos resultados evidencian que la clasificación diagnóstica sí puede ser fiable aún cuando la fiabilidad de los puntajes no lo sea, en la medida que el punto de corte se encuentre suficientemente alejado del promedio de la escala.

Como conclusión anecdótica, estos resultados sugieren que la conclusión original de Gempp y Vinet (1997) fue, tal vez, un poco apresurada, dado que psicométricamente hablando la fiabilidad de la decisión para la cual se estaba utilizando el instrumento sí era apropiada.

Discusión

En este trabajo se ha revisado, brevemente, el coeficiente *K2* de Livingston (1972, 1973) argumentado y demostrada su aplicación con dos ejemplos empíricos.

Del análisis de la Ecuación 4, que presenta la fórmula de cómputo del coeficiente, como del estudio de los ejemplos presentados, se pueden concluir varias recomendaciones prácticas. La primera, es que el coeficiente *K2* es muy fácil de calcular y que su cómputo puede hacerse incluso a partir de datos secundarios. La segunda, es que la fiabilidad de una clasificación habitualmente será más alta que el alfa de Cronbach de la escala o test, especialmente entre más distante se encuentre el punto de corte de la media de la escala. En tercer lugar, que una escala o test puede arrojar resultados de clasificación muy fiables incluso si la fiabilidad de los puntajes es baja según estándares convencionales.

Por otro lado, es importante recordar que el *K2* tiene al menos dos limitaciones importantes. Por un lado, solo es útil en el caso de clasificaciones dicotómicas. Por otro, al basarse en un modelo de consistencia interna su interpretación no es tan sencilla como lo son los coeficientes basados en consistencia entre pruebas paralelas.

Sin embargo, dada la simplicidad de su cálculo, nos atrevemos a proponer que el coeficiente *K2* de Livingston, ampliamente conocido en medición educativa, puede ser una herramienta muy útil en el análisis psicométrico de test psicológicos.

Referencias

Brennan, R. L. & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*(3), 277-289.

Craig, T. J. & Van Natta, P. A. (1976). Presence and persistence of depressive symptoms in community, clinic and mental hospital groups. *American Journal of Psychiatry, 133*(12), 1426-1429.

Craig, T. J. & Van Natta, P. A. (1979). Influence of demographic characteristics on tow measures of depressive symptoms. *Archives of General Psychiatry, 36*(2), 149-154.

Gempp, R. & Thieme, C. (2010). Efecto de diferentes métodos de puntuación sobre la fiabilidad, validez y puntos de corte de la Escala de Depresión del Centro para Estudios Epidemiológicos (CES-D). *Terapia Psicológica, 28*(1), 5-12.

Gempp, R. & Vinet, E. (1997, noviembre). *Confiabilidad del Inventario Multifásico de Personalidad de Minnesota (MMPI) en postulantes a la carrera de psicología*. Trabajo presentado en el V Congreso Nacional de Psicología, Santiago, Chile.

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

Harris, C. W. (1972, abril). *An index of efficiency for fixed-length mastery tests*. Trabajo presentado en The Annual meeting of the American Educational Research Association, Chicago, USA.

Hathaway, S. R. & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory manual*. Minneapolis: University of Minnesota Press.

Lee, W. C., Hanson, B. A. & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*(4), 412-432.

Livingston, S. A. (1972). Criterion-referenced applications of Classical Test Theory. *Journal of Educational Measurement, 9*(1), 13-26.

Livingston, S. A. (1973). A note on the interpretation of the criterion-referenced reliability coefficient. *Journal of Educational Measurement, 10*(4), 311.

Mari, J. D. J. & Williams, P. (1985). A comparison of the validity of two psychiatric screening questionnaires (GHQ-12 and SRQ-20) in Brazil, using Relative Operating Characteristics (ROC) analysis. *Psychological Medicine, 15*(3), 651-659.

McArdle, J. J., Johnson, R. C., Hishinuma, E. S., Miyamoto, R. H. & Andrade, N. N. (2001). Structural equation modeling of group differences in CES-D ratings of native Hawaiian and non Hawaiian high school students. *Journal of Adolescent Research, 16*(2), 108-149.

Mossman, D. & Somoza, E. (1989). Maximizing diagnostic information from the dexamethasone suppression test. *Archives of General Psychiatry, 46*(7), 653-660.

Murphy, J. M., Berwick, D. M., Weinstein, M. C., Borus, J. F., Budman, S. H. & Klerman, G. L. (1987). Performance of screening and diagnostic tests: Application of receiver operating characteristic analysis. *Archives of General Psychiatry, 44*(6), 550-555.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401.

Radloff, L. S. & Locke, B. Z. (1986). The Community Mental Health Assessment Survey and CES-D Scale. En M. M. Weisman, J. K. Myers & C. E. Ross (Eds.), *Community surveys of psychiatric disorders* (pp. 177-189). East Brunswick, NJ: Rutgers University Press.

Rissetti, F. J., Himmel, E., Maltes, S. G., González, J. A. & Olmos, S. (1989). Estandarización del Inventario Multifásico de Personalidad de Minnesota (MMPI), en población adulta chilena. *Revista Chilena de Psicología, 10*(1), 41-61.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.

Swets, J. A. & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from Signal Detection Theory*. New York: Academic Press.