

ERROR DE MEDICIÓN ALREDEDOR DE LOS PUNTOS DE CORTE EN EL MBI-GS

MEASURING ERROR AROUND THE CUTTING POINTS IN THE MBI-GS

Manuel Fernández Arata* y César Merino Soto**
Universidad de San Martín de Porres, Perú

Recibido: 04 de agosto de 2014

Aceptado: 26 de septiembre de 2014

RESUMEN

En la calificación de los resultados del MBI-GS (Maslach Burnout Inventory - General Survey) como de sus antecesores, se usan interpretaciones dimensionales, pero también puntos de corte para clasificar a los sujetos y diferenciarlos con respecto a la presencia del síndrome de *burnout*. Sin embargo, este uso está orientado al tipo de tests referidos a criterios, por lo que la estimación de la confiabilidad debe converger con este tipo de uso. El objetivo del presente estudio fue estimar la confiabilidad alrededor de varios puntos de corte en el MBI-GS (agotamiento emocional, eficacia profesional y cinismo), mediante el coeficiente K2 de Livingston. Los participantes fueron 741 sujetos de varias ocupaciones, provenientes de instituciones públicas y privadas de Lima y de otras ciudades. Previo a la obtención de la confiabilidad criterial, se verificó la unidimensionalidad, el modelo equivalente tau, la distribución y la bimodalidad de los puntajes. Se halló que los puntos de corte más extremos muestran mayor confiabilidad; que la confiabilidad se incrementa en relación al grado de extremidad de los puntajes; que el puntaje de agotamiento emocional obtiene mayor confiabilidad criterial y que cinismo obtiene la menor, pero todos los puntajes generalmente muestran buenas confiabilidades. Se discute el uso de los puntos de corte en relación a su validez y confiabilidad.

Palabras clave: Puntos de corte, burnout, MBI-GS, confiabilidad, diagnóstico, prueba referida a criterios.

ABSTRACT

In the rating of the MBI-GS results (Maslach Burnout Inventory - General Survey) as its predecessors, we used dimensional interpretations, but also cut-off points for classifying and differentiating the subjects with respect to the presence of the burnout syndrome. However, this use is oriented to the type of tests relating to criteria so that the estimate of the reliability must converge with this type of use. The objective of this study was to estimate the reliability around various cutting points in the MBI-GS (emotional exhaustion, professional effectiveness and cynicism), through Livingston's K2 coefficient. The participants were 741 subjects of various occupations, from public and private institutions in Lima and other cities. Prior to obtaining the reliability criterion, the unidimensionality was verified, the equivalent model tau, distribution and bimodality of the scores. It was found that the most extreme cut-off points show greater reliability; that the reliability increases in relation to the degree of extremity of the scores; the scores of emotional exhaustion get greater reliability criterion and that cynicism gets the lowest, but all the scores generally show good reliability. This paper discusses the use of the cut-off points in relation to their validity and reliability.

Key words: Cut-off points, burnout, MBI-GS, reliability, diagnosis, criterion-referenced tests

* mfernandezal@usmp.pe

** sikayax@yahoo.com.ar

Los puntos de corte utilizados para interpretar los puntajes de una escala de medición psicológica, aunque permiten categorizar a los sujetos, pero pueden resultar con frecuencia poco confiables e imprecisos. Si bien los manuales de algunas escalas están acompañados de los baremos correspondientes, estos pueden presentar dificultades cuando se los utiliza para realizar interpretaciones en poblaciones con características culturales distintas. Debido a que los puntos de corte se construyen sobre los estadísticos descriptivos de muestras de referencia, estos presentan también el mismo problema de posible invalidez intercultural. Estos problemas son aún más críticos en instrumentos que se usan en el trabajo clínico y en la investigación para definir grupos de sujetos caracterizados por aspectos desadaptativos. Uno de estos instrumentos es el *Maslach Burnout Inventory* - MBI (Maslach, Jackson & Leiter, 1996), que además de su gran expansión intercultural (Schaufeli, Leiter & Maslach, 2009), se utiliza para identificar sujetos con elevados síntomas de agotamiento emocional e indiferencia, así como disminuida eficacia profesional y compromiso con el trabajo realizado.

El MBI adolece también de este problema práctico cuando se requiere realizar un diagnóstico con fines de intervención. La misma autora de la escala manifiesta su disconformidad con el uso de los puntos de corte en la calificación e interpretación del MBI, dado que sus puntajes se interpretan esencialmente de manera dimensional (Schaufeli et al., 2009). Sin embargo, al igual que cualquier otro instrumento que pretende describir alguna característica psicológica de un grupo poblacional, el MBI requiere de un mecanismo válido y confiable para identificar los niveles de burnout que experimentan los trabajadores de una organización (Schaufeli, et al. 2009); este requerimiento es más urgente al considerar los altos niveles de burnout en trabajadores, así como su globalización progresiva alrededor del mundo.

Maslach et al. (1996) sugieren que la identificación de los grupos con diferentes niveles de burnout se puede efectuar en tres agrupaciones de similar tamaño, lo que básicamente divide a una muestra de sujetos en tercios correspondientes a los percentiles 33 y 66. Esto produciría un número mayor de sujetos identificados cuando se los compara con otros métodos que usan puntajes más alejados de la media, por ejemplo, el uso de cuartiles. Algunos

estudios con el MBI han aplicado cuartiles para la identificación de los niveles del burnout (Brenninkmeijer & Van Yperen, 2003; Lindblom, Linton, Fedeli & Bryngelsson, 2006; Schaufeli, Bakker, Schaap, Kladler & Hoogduin, 2001), mientras que otros han usado la mediana como criterio de diferenciación (Maslach & Leiter, 2008; Golembiewski, Munzenrider & Carter, 1983). Estos criterios pueden ser convenientes debido a que la mediana es menos afectada por datos extremos. En nuestro medio, Fernández (2002, 2008, 2010), reporta niveles de burnout en maestros de educación primaria y secundaria, clasificando a los investigados en tres tercios pero sin utilizar un criterio adicional externo que justifique el método utilizado.

Un método poco reportado aún es el uso de la desviación estándar, en el que se separa a los grupos de acuerdo con el grado de desviación respecto a la media de la muestra de estudio; típicamente se aplica $\pm 2 DE$ como criterio del grado de alejamiento del puntaje (Llaja, Sarria & García, 2007; Peters, Jelacic, Heij & Merckelbach, 2006). Este criterio podría suponer una mejor detección de verdaderos positivos y, consecuentemente, el decremento de falsos positivos (Schaufeli & Enzmann, 1998), ya que los grupos con puntajes mayores (o menores) a dos desviaciones de la media representan al 2.5% de una población normalmente distribuida. Sin embargo, dada la naturaleza del burnout, es probable que la forma que asume la distribución no sea normal sino asimétrica, lo que puede crear un límite en la aplicación de este criterio.

Otro método que podría mejorar la sensibilidad clínica del MBI-GS es el propuesto por Gil-Monte, Carretero, Desamparados y Núñez-Román (2005) para el Cuestionario de Evaluación del Síndrome de Quemarse por el Trabajo (Gil-Monte, 2004), que usa los percentiles 10, 25, 75, 90. En cualquiera de los métodos aplicables, la obtención de valores de referencia propios de la muestra de estudio y la validez clínica de los mismos, serán la piedra angular para el uso efectivo de un método de clasificación e identificación de los niveles de burnout (Gil-Monte et al., 2005; Schaufeli & Van Dierendonk, 1995).

Si bien es cierto que reportar la caracterización categórica del burnout puede facilitar la comprensión de la intensidad del mismo en un grupo de sujetos, ésta requiere no solo simplicidad de su comunicación sino

* mfernandezal@usmp.pe

** sikayax@yahoo.com.ar

también un buen respaldo de validez. En este sentido, algunos investigadores han resaltado las dificultades para establecer estrategias diagnósticas bien fundamentadas para identificar el burnout (Dyrbye, West & Shanafelt, 2009; Olivares & Gil-Monte, 2009; Schaufeli, et al, 2009; Schaufeli & Van Dierendonk, 1995), principalmente porque (a) los puntos de corte elegidos en el MBI no tienen un suficiente respaldo en la evidencia acumulativa que combine criterios clínicos y estadísticos, (b) el diagnóstico es menos aceptado cuando este se basa en un solo criterio (los puntajes de un solo test) y (c) los puntos de corte y las normas tienden a obtenerse de muestras no representativas y arbitrarias (Olivares & Gil-Monte, 2009; Schaufeli et al., 2009). Esto último no permite identificar con seguridad el error muestral de los estadísticos de referencia (por ejemplo, la media y la desviación estándar) ni asegurar la generalización. También es posible reconocer que los puntos de corte no se acompañan muchas veces de una suficiente racionalidad del autor para elegirlos, por ejemplo entre elegir cuartiles, la mediana, terciles o $\pm 2 DE$.

Tomando en cuenta las limitaciones de los procedimientos anteriores en la construcción de normas y puntos de corte con el MBI, hay una clara recomendación por usar normas locales y puntos de corte validados clínicamente como la mejor opción para llevar al MBI a la evaluación con propósitos individuales (Schaufeli et al., 2009). Uno de los pocos puntos de corte con un buen respaldo de validez clínica lo ha reportado Schaufeli et al. (2001) para la estandarización de la escala en Holanda, en la que usaron criterios de la Clasificación Internacional de Enfermedades (ICD-10, 1994). Sin embargo, la eficiencia diagnóstica mediante puntos de corte del MBI no parece replicarse de manera efectiva en recientes estudios (Kleijweg, Verbraak & Van Dijk, 2013). Esta inconsistencia parece deberse al criterio del DSM con el cual se podría comparar la clasificación del MBI.

Efectivamente, las limitaciones referidas en los párrafos anteriores no son las únicas; ya que otros aspectos críticos asociados a los puntos de corte, no han sido tratados a profundidad. Uno de ellos es el fingimiento de los síntomas de burnout en relación a los puntajes extremos (Peters, et al. 2006) y otro, es el error de medición.

Sobre el primero, Peters et al. (2006) hallaron que las diferencias entre examinados instruidos para exagerar

síntomas de burnout y los que sí padecían el burnout no eran claras, especialmente en los puntajes extremos. Por otro lado, respecto al error de medición, este parece haber sido considerado de manera habitual (usando coeficientes de consistencia interna) pero sin tomar en cuenta su impacto sobre los puntos de corte. El presente estudio abordó la obtención de los puntos de corte incluyendo la información referida al error de medición y su comparabilidad con las obtenidas de manera tradicional. Este error de medición se estima directamente desde un coeficiente de confiabilidad, como los que se obtienen al calcular la consistencia interna (Nunnally & Bernstein, 1995).

El reporte de la consistencia interna del MBI es una práctica estándar para presentar el grado de confiabilidad de sus puntajes, y el coeficiente α (Cronbach, 1951), es el método comúnmente elegido (Feldt & Brennan, 1989). Este coeficiente provee una representación estática de la confiabilidad, y se le asume como constante en cada nivel de puntuación (Feldt & Brennan, 1989). También es la elección más frecuente entre los investigadores cuando usan puntajes de *pruebas referidas a normas (PRN)* en el que se ubica el nivel de desempeño de un sujeto comparado con un grupo normativo.

La interpretación normativa de los puntajes del MBI corresponde a la naturaleza dimensional de sus constructos, pero con la construcción de puntos de corte, el uso del MBI tiene un diferente propósito y aparentemente, muchos de sus usuarios lo han preferido por décadas. La clasificación y distinción de sujetos con alto o bajo burnout sugiere que los puntajes del MBI se usan en el marco de las *pruebas referidas a criterios (PRC)*, en los que el objetivo básico es separar dos grupos de sujetos: los que aprueban (pasan) o desaproveban (no pasan) un test (Woolfolk, 2010). Esta distinción se practica frecuentemente con el MBI, pero no se ha reconocido su rol como instrumento referido a un criterio. Dentro de este tipo de uso del MBI, la estimación de la confiabilidad debe enfocarse en la correcta identificación de los sujetos en el punto de corte elegido, y por lo tanto el coeficiente α ya no es el mejor estimador del mismo. Hay varios modelos creados para estimar la confiabilidad en una prueba de criterio, basados en la consistencia de las desviaciones, de las decisiones o de las estimaciones de dominio (Mateo, 1992), pero hasta la fecha no se dispone de ninguna evidencia de que hayan sido aplicados al MBI.

* mfernandezal@usmp.pe

** sikayax@yahoo.com.ar

LIBERABIT: Lima (Perú) 20(2): 209-218, 2014

El uso de coeficientes de confiabilidad para pruebas referidas a criterios derivadas de la teoría clásica de los tests para pruebas referidas a normas (PRN), ha sido demostrado y estimulado desde hace varios años atrás (Lopez, Mostafavi, Nibut & Smith, 2005; Lopez, Perez, Smith & Castillo, 2007; Lopez, Quan & Carvajal, 2010). También, recientemente en la literatura de habla hispana ha sido valorado como un método complementario tan importante como los coeficientes de confiabilidad para PRN (Gempp & Saiz, 2014), cuando un instrumento permite también una interpretación criterial de sus puntajes.

El objetivo del presente estudio es aplicar la metodología de la confiabilidad de PRC en el MBI-GS (Maslach et al., 1996), en el que se evaluará la precisión de la clasificación relativa a los puntos de corte (Haertel, 2006), un aspecto relevante al MBI-GS cuando la interpretación de sus puntajes sea criterial. Se utilizó el MBI-GS porque es una escala que permite evaluar el burnout en diversos grupos ocupacionales, además de ser breve comparada con sus otras versiones, y parece ser de uso creciente en la literatura científica y práctica profesional.

Método

Participantes

El diseño para la elección de la muestra fue no probabilístico pero se trató de obtener la mayor heterogeneidad respecto a las ocupaciones. Está conformada por 741 trabajadores, 57.2% del sexo masculino y 42.5% femenino (dos sujetos no respondieron). El 76% laboraba en instituciones públicas y el resto en instituciones privadas; en cuanto al estado civil, el 39% son solteros, el 43% casados, el 12.4% practica la unión libre, el 3.6% son divorciados y el 1.2% están viudos (5 sujetos no reportaron esta información). Con respecto al tipo de ocupación el 46.4% son docentes, el 34.3% empleados manuales, 8.4% empleados administrativos, el 10.1% supervisores o jefes y el .8% directivos. La edad varió entre 18 y 69 años ($M = 39.31$, $DE = 11.13$).

Instrumento

La evaluación del síndrome de burnout se realizó mediante el Maslach Burnout Inventory General Survey (MBI-GS; Maslach, et al., 1996), en su traducción española (Juárez et al., 2011). El MBI-GS mide respuestas

relacionadas con el trabajo y está conformada por 16 ítems y tres subescalas: Agotamiento emocional (cinco ítems, p. ej., «Me siento desgastado al final de la jornada laboral»); cinismo (cinco ítems, p. ej., «Dudo de la relevancia que tiene mi trabajo»); y eficacia profesional (seis ítems, p. ej., «En mi opinión, soy bueno en mi trabajo»). Todos los ítems obtienen un puntaje en una escala de frecuencia de rango entre 0 (*nunca*) y 6 (*diariamente*).

El ítem 13 no fue incluido en el puntaje de cinismo por dos motivos: primero, porque los resultados previos en población nacional (Fernández, Merino & Guimet, en prensa) e internacional (Bakker, Demerouti & Schaufeli, 2002; Gil-Monte, 2002; Oramas, González & Vergara, 2007; Salanova & Schaufeli, 2000; Schutte, Toppinnen, Kalimo & Schaufeli, 2000), consistentemente recomiendan no usarlo porque no se ajusta empíricamente a su dimensión; y, segundo, porque la correlación ítem-test en el presente estudio fue cerca de cero.

Procedimiento

La información recogida incluyó tres grupos de trabajadores: un grupo de 344 docentes de educación básica y secundaria de 18 escuelas públicas de Lima, quienes fueron encuestados en sus propias instituciones educativas; un segundo grupo de 254 trabajadores de nivel operativo de una Municipalidad limeña, quienes respondieron la encuesta en su lugar de trabajo; y, un tercer grupo de 143 trabajadores de un empresa agroindustrial del norte del país conformado por empleados administrativos y profesionales, quienes también respondieron el instrumento en su centro laboral. En los tres casos se solicitó permiso a los responsables de los centros laborales para poder realizar el recojo de información en un lugar y horario apropiado. También, se les informó del propósito de la investigación, así como de su participación voluntaria en el estudio.

Se probaron cuatro presupuestos estadísticos univariados y de medición relacionados con las escalas del MBI-GS, antes de realizar el análisis principal. En primer lugar, como los procedimientos matemáticos para lograr el objetivo principal del presente estudio intentan extraer conclusiones usando la media y la varianza, se tuvo especial cuidado en la detección de datos extremos (*outliers*). Por lo tanto, se aplicó un procedimiento basado

* mfernandezal@usmp.pe

** sikayax@yahoo.com.ar

en la desviación absoluta de la mediana (Leys, Ley, Klein, Bernard & Licata, 2013), dado que el método usual del puntaje z es generalmente menos eficiente (Leys et al., 2013), especialmente en distribuciones altamente asimétricas (Leys et al., 2013; Thompson, 2006).

En segundo lugar, se modeló individualmente cada escala del MBI-GS, mediante el análisis de ecuaciones estructurales, aplicando el método de Satorra y Bentler (1994), $SB-\chi^2$, y complementándolo con índices de ajuste ($CFI \geq 0.95$ y $SRM \leq 0.08$). No se usó como indicador del ajuste a RMSEA debido a que el pequeño grado de libertad de cada modelo produciría estimaciones artificialmente RMSEA altas (Kenny, Kaniskan & McCoach, en prensa; Quintana & Maxwell, 1999). En tercer lugar, se probó el modelo de medición (paralelo y equivalente tau) en cada subescala para asegurar la estimación de confiabilidad más apropiada, es decir, si era apropiado usar α (Cronbach, 1951) o el coeficiente Gilmer-Feldt, r_{gr} (Gilmer & Feldt, 1983); este último es recomendado para modelos congénéricos (Lopez et al., 2005; Lopez et al., 2007; Lopez et al., 2010). En último lugar, considerando que las distribuciones bimodales son problemáticas para obtener un punto medio válido y su efecto es directo sobre la exactitud del $K2$ (Subkoviak, 1976), se examinó el grado de bimodalidad mediante la observación de gráficos de histogramas aplicándose un método formal mediante el coeficiente de bimodalidad - CB (SAS Institute, 1989).

Para la estimación de la confiabilidad en el punto de corte, se usó el coeficiente $K2$ (Livingston, 1972), recomendado para aplicaciones como la del presente estudio (Gempp & Saiz, 2014; Lopez et al., 2005; Lopez et al., 2007; Lopez et al., 2010). Además de las ventajas de su cálculo sencillo (Gempp & Saiz, 2014), parece estar débilmente relacionado con la forma distribucional de los puntajes (Rim & Bresler, 1974) y se deriva directamente del modelo clásico de los tests (Gempp & Saiz, 2014; Livingston, 1972, 1973). La ecuación es la siguiente:

$$K2 = \frac{r_{xx}DE^2 + (\bar{X} - C)^2}{DE^2 + (\bar{X} - C)^2}$$

En la ecuación, r_{xx} es el coeficiente de consistencia interna, DE la desviación estándar, \bar{X} la media y C el

puntaje que sirve como punto de corte. Se estimaron los puntos de corte en niveles de puntuación basados en cuartiles (25%, 50%, 75%), terciles (33% y 66%) y unidades de desviación estándar ($\pm 1 DE$, $\pm 2 DE$).

Resultados

Análisis preliminar

En esta sección se verificó la ausencia de puntajes extremos, la unidimensionalidad, el modelo de medición subyacente y la ausencia de bimodalidad. En la identificación de puntajes extremos, se detectaron 24 sujetos con Z_{MAD} extremos en algunos de los puntajes del MBI-GS; considerando que el efecto de removerlos es estadísticamente más conveniente que retenerlos (Osborne & Overbay, 2004), estos sujetos fueron separados de los datos. Respecto al modelamiento de las subescalas, el resultado del ajuste de la dimensionalidad se presenta en la Tabla 1. Exceptuando CIN, AE y EP presentaron ajuste satisfactorio sin hacer re-especificaciones. De acuerdo a los índices de modificación de Lagrange y la inspección de residuales, la subescala CIN requirió modelar el término de error entre los ítems 8 y 9, lo que produciría sustancial mejoría al ajuste: $SB-\chi^2 (gl: 1) = 7.202, p < 0.01, CFI = 0.99, SRMR = 0.014$. Sin embargo, se omitió esta re-especificación pues la estimación de la confiabilidad congénérica no incluía este término del error correlacionado en su definición matemática.

Por otro lado, en la misma Tabla 1 se presenta la comparación entre el modelo congénérico y equivalente tau, estimado mediante la diferencia escalada de $SB-\chi^2$ (Satorra & Bentler, 2001). El modelo congénérico produce una mejora estadísticamente significativa en todos los factores del MBI-GS, por lo tanto, parece representar mejor el modelo de medición subyacente al MBI-GS. En la Tabla 2 se muestran los coeficientes de confiabilidad y se puede ver que no hay una diferencia práctica entre los coeficientes, aún cuando las diferencias entre los modelos evaluados fueron estadísticamente significativas. Adicionalmente, se observa que la remoción del ítem 13 produce una importante mejora en los coeficientes de confiabilidad de CIN. Para tener consistencia con el modelamiento previo se usó el coeficiente Gilmer-Feldt (r_{gr}).

* mfernandezal@usmp.pe

** sikayax@yahoo.com.ar

LIBERABIT: Lima (Perú) 20(2): 209-218, 2014

Tabla 1*Resultados del ajuste de los modelos de medición de cada subescala del MBI-GS*

	SB- χ^2 (gl)	CFI	SRMR	M	DE
Agotamiento Emocional (AE)				8.25	5.70
Congenerico	60.124 (5)	0.97	0.043		
Equivalente tau	73.465 (8)	0.97	0.054		
Diferencia ^a	17.247 (3)***	-	-		
Eficacia Profesional (EP)				29.88	5.348
Congenerico	28.49 (9)**	0.98	0.031		
Equivalente tau	49.33 (13)**	0.96	0.060		
Diferencia ^a	17.861 (4)**	-	-		
Cinismo (CIN)				3.66	4.156
Congenerico	81.205 (2)**	0.893	0.091		
Equivalente tau	139.309 (4)**	0.817	0.107		
Diferencia ^a	62.600 (2)***	-	-		

^a Diferencia escalada (Satorra & Bentler, 2001). M: media. DE: desviación estándar.

Tabla 2*Estimaciones de consistencia interna para los puntajes del MBI-GS*

	Nº ítems	M (DE)	r_{ii} (mín, máx)	α	ω	r_{gf}
Agotamiento emocional (AE)	5	8.25 (5.70)	0.513(0.383, 0.647)	0.841	0.841	0.843
Eficacia profesional (EP)	6	29.88 (5.34)	0.355(0.241, 0.555)	0.768	0.771	0.754
Cinimos (CIN) con 13	5	-	0.296(0.090, 0.709)	0.678	0.692	0.633
Cinismo (CIN) sin ítem 13	4	3.66 (1.15)	0.422(0.280, 0.709)	0.745	0.746	0.731

r_{ii} : correlación inter-ítem promedio. α : coeficiente alfa. ω : coeficiente omega. r_{gf} : coeficiente Gilmer-Feldt.

Confiabilidad en los puntos de corte

Para identificar si se construirían puntos de corte de acuerdo a alguna diferencia muestral importante, se examinó y comparó la función de distribución empírica (prueba Kolmogorov-Smirnov, Z_{KS}) de las escalas MBI-GS respecto al sexo y tipo de profesión (profesores vs. no profesores); también, se examinó la relación entre la edad y los puntajes del MBI-GS. Los resultados indican que respecto al género se detectaron diferencias sistemáticas en las distribuciones de AE ($Z_{KS} = 2.067$, $p < 0.001$, $r = 0.15$), IN ($Z_{KS} = 1.371$, $p < 0.05$, $r = 0.10$) y EP ($Z_{KS} = 1.464$, $p < 0.01$, $r = 0.10$). Sobre el tipo de profesión, únicamente se detectaron diferencias en CIN ($Z_{KS} = 1.896$, $p < 0.01$, $r = 0.13$) y EP ($Z_{KS} = 2.411$, $p < 0.01$, $r = 0.17$). Asimismo, la

edad covarió bajo con AE ($r = -0.127$, $p < 0.01$) y EP ($r = -0.115$, $p < 0.01$). Aunque estos resultados dan evidencias para no aceptar la hipótesis nula, son de pequeña magnitud, y por lo tanto, el análisis principal se hará considerando toda la muestra de estudio.

Los coeficientes de confiabilidad criterial alrededor de los puntos de corte se presentan en la Tabla 3. Se observa que, como era de esperar, los puntos de corte más extremos obtienen menor error de medición alrededor de ellos. En los mismos niveles de corte, el puntaje de CIN generalmente obtiene los coeficientes más bajos comparados con AE y EP.

* mfernandezal@usmp.pe

** sikayax@yahoo.com.ar

Tabla 3
Confiabilidad criterial en el nivel de los puntos de corte

	Agotamiento emocional (AE)		Eficacia Profesional (EP)		Cinismo (CIN)	
	Puntaje	K2	Puntaje	K2	Puntaje	K2
PC Percentiles						
5	1	0.939	19	0.955	0	0.857
25	4	0.898	27	0.822	0	0.857
33	5	0.880	29	0.777	1	0.820
50	7	0.848	31	0.781	2	0.781
66	10	0.855	33	0.829	5	0.770
75	12	0.889	34	0.856	6	0.807
95	15	0.934	36	0.901	12	0.949
PC DE						
-1DE	2 (12%)	0.928	24 (13%)	0.896	0 (16%)	0.857
+1DE	14 (84%)	0.921	36 (93%)	0.901	8 (85%)	0.879

Desviación estándar. PC: puntos de corte. K2: coeficiente de Livingston.

AE es la escala que muestra mejor confiabilidad. Las escalas del MBI-GS, aunque generalmente producen elevadas confiabilidades criteriosales alrededor de los puntos de corte, exhiben alguna variabilidad y en consecuencia el error de medición también tiende a ser variable.

Discusión

Las advertencias declaradas en el manual sobre los puntos de corte (Maslach et al., 1996) parece que no han evitado que un número de investigadores (p. ej., en Perú, Arias & Jiménez, 2013; Dueñas, Merma & Ucharico, 2003; Fernández, 2002; Gamonal, García & Silva, 2008; Ponce, Bulnes, Aliaga, Atalaya & Huertas, 2005) los usen para identificar los casos que padecen burnout entre aquellos que no lo experimentan, y para establecer la posición del sujeto relativo a la distribución de la muestra de estandarización para propósitos diagnósticos. El mismo manual también no aconseja el uso del puntaje transformado en categorías para identificar los niveles de burnout como usualmente se realiza en la práctica, y en general, el uso de los puntajes del MBI para propósitos diagnósticos. Sin embargo, esta recomendación contra el uso de puntos de corte es razonable en ausencia de evidencias de validez y de error de medición. El objetivo del presente trabajo fue obtener una apropiada estimación de la confiabilidad en el MBI-GS, dentro del marco de las

pruebas referidas a criterios, pues el uso de puntos de corte que identifica a este tipo de pruebas puede servir para interpretar el MBI.

Los resultados indican que los puntos de corte más extremos obtienen confiabilidades más altas, satisfaciendo el criterio de > 0.90 indicado por Nunnally y Bernstein (1995) cuando se refieren a la intolerancia de los errores de medición en aplicaciones clínicas, en las decisiones sobre los sujetos evaluados. Efectivamente, las clasificaciones más extremas pueden proveer no solo de estimaciones más confiables en términos del método K2, sino también covariaciones más altas con criterios desadaptativos; por ejemplo, los cuartiles más extremos de despersonalización en el MBI-HSS (Maslach et al., 1996) se relacionan fuertemente con la frecuencia de prácticas subóptimas de cuidado a pacientes (Shanafelt, Bradley, Wipf & Back, 2002). Por lo tanto, una adecuada combinación de evidencias sobre la confiabilidad en el punto de corte elegido y su efectividad para clasificar sujetos darían el respaldo que se requiere para usar un específico punto de corte en el MBI-GS.

La obtención de la confiabilidad alrededor de los puntos de corte demuestra que un apropiado enfoque para modelar el error de medición debe ser incluido con el MBI-GS cuando se requiera usar, clasificar o diferenciar grupos. Sin embargo,

* mfernandezal@usmp.pe

** sikayax@yahoo.com.ar

esto no resuelve el problema de la validez externa de los puntos de corte elegidos, pues en ausencia de resultados diagnósticos externos, habría poco respaldo para usarlos. Esto significa que ambas informaciones psicométricas, la confiabilidad y la validez, no son intercambiables para valorar el uso de puntos de corte en el MBI-GS.

Del mismo modo en que las propiedades psicométricas no son necesariamente transferibles entre estudios, las propiedades entre las escalas del MBI-GS tampoco lo son. Esto se resalta por la relativa baja confiabilidad criterial y normativa de CIN frente a AE y EP. Esto no parece una limitación de considerable magnitud si se toma el valor absoluto de sus coeficientes que son mayores o iguales a 0.85 en los puntos de corte más extremos; sin embargo, señala el efecto acumulativo de la confiabilidad normativa y el grado de ajuste estructural. Por lo tanto, CIN tenderá a producir un mayor rango de variación en las clasificaciones basadas en los puntos de corte comparado con AE, que puede parecer un pequeño aumento en el error de medición pero que sería de especial cuidado cuando se clasifique a los sujetos para decisiones críticas, como recomendar intervención recuperativa.

En el proceso de establecer la confiabilidad criterial se deben considerar dos aspectos metodológicos: uno sobre el modelamiento de medición de cada subescala y otro sobre K2. En primer lugar, en la evaluación del ajuste de los modelos de medición, se toleró un tipo de especificación que puede ser común en los resultados SEM. Aunque la potencial re-especificación era óptima de acuerdo a los índices de ajuste, la decisión de retenerlo estuvo de acuerdo con Chen, Curran, Bollen, Kirby y Paxton (2008), en que algún grado de especificación impropia puede ser tolerado, y se evitó capitalizar el pequeño número de variables en cada modelamiento. Sin embargo, esto puede aumentar espuriamente los coeficientes de confiabilidad como α (Komaroff, 1997; Zimmerman, Zumbo & Lalonde, 1993) y razonablemente, también al coeficiente r_{gf} o disminuirlos, como ocurriría con los coeficientes basados en el análisis factorial (Lucke, 2005). En el futuro se requerirá una modificación al r_{gf} para incluir en su formulación matemática el término correspondiente a la covariación de los errores, tal como ocurre con la modificación del coeficiente α en Komaroff (1997) o Rae (2006).

En segundo lugar, debido a la sensibilidad de K2 a los cambios en la relación media-PC (Marshall & Haertel, 1975), las confiabilidades criterioles obtenidas no podrían ser generalizables a otros grupos a menos que se compruebe que las diferencias de media entre grupos sean de poca importancia práctica. Esto también se aplica a las diferencias entre los coeficientes de consistencia interna obtenidas en las mismas muestras de estudio, que pueden ser evaluados mediante pruebas inferenciales (p. ej., Feldt & Brennan, 1989; Haertel, 2006). Por otro lado, hay otros métodos potentes de estimación como los modelos basados en conocidas propiedades distribucionales, provenientes de la teoría fuerte del puntaje verdadero, como beta-binomial o beta-binomial de 4 parámetros, pero requieren aproximaciones de alguno de sus parámetros que no son apropiados para ítems politómicos como el uso del coeficiente KR-21 para el modelo beta-binomial de 4 parámetros. Por lo tanto, los coeficientes derivados de condiciones similares (p. ej., Subkokiav, 1976; Huynh, 1976), parecen inapropiados para los puntajes de este estudio.

Un estudio ha reportado que K2 parece integrar las dos tradiciones de medición (criterial y normativa), y puede ser más útil que otros tipos de estimación (Downing & Mehrens, 1978), por lo tanto, este método parece la opción más apropiada para estimar la confiabilidad y construir bandas de error de medición alrededor de los puntos de corte del MBI-GS.

Referencias

- Arias, W. & Jiménez, B. (2013). Síndrome de burnout en docentes de educación básica regular de Arequipa. *Educación*, 22(42), 53-76.
- Bakker, A. B., Demerouti, E. & Schaufeli, W. B. (2002). Validation of the Maslach Burnout Inventory - General Survey: An Internet study. *Anxiety, Stress, and Coping*, 15, 245-260.
- Brenninkmeijer, V. & Van Yperen, N. (2003). How to conduct research on burnout: advantages and disadvantages of a unidimensional approach in burnout research. *Occupational and Environmental Medicine*, 60, 16-20.
- Chen, F., Curran, P. J., Bollen, K. A., Kirby, J. & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA Test Statistic in Structural Equation Models. *Sociological Methods and Reserarch*, 36(4), 462-494.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

* mfernandezal@usmp.pe

** sikayax@yahoo.com.ar

- Downing, S. M. & Mehrens, W. A. (marzo,1978). *Six single-administration reliability coefficients for criterion-referenced tests: A comparative study*. Paper presented at the Annual Meeting of the American Educational Research Association, 62nd, Toronto, Ontario, Canada.
- Dueñas, M., Merma, L. & Ucharico, R. (2003). Prevalencia de Burnout en médicos de la ciudad de Tacna. *Ciencia e Investigación Médica Estudiantil Latinoamericana*, 8(1), 33-37.
- Dyrbye, L. N., West, C. P. & Shanafelt, T. D. (2009). Defining burnout as a dichotomous variable. *Journal of General Internal Medicine*, 24(3), 440.
- Feldt, L. S. & Brennan, R. (1989). Reliability. En R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.
- Fernández, M. (2002). Desgaste psíquico en profesores de primaria de Lima Metropolitana. *Persona*, 5, 27-66.
- Fernández, M. (2002). *Realidad psicosocial del maestro de primaria*. Lima: Universidad de Lima, Fondo de Desarrollo Editorial.
- Fernández, M. (2008). Burnout, autoeficacia y estrés en Maestros Peruanos: Tres Estudios Fáclicos. *Ciencia & Trabajo*, 10(30), 120-125.
- Fernández, M. (2010). El estrés laboral en los peruanos: hallazgos recientes. *Teoría e investigación en Psicología*, 19, 37-59.
- Fernández, M., Merino, C. & Guimet, M. (en prensa). Propiedades psicométricas del Maslach Burnout Inventory-General Survey en una muestra de docentes de Lima, Perú. En R. León (Ed.), *Libro en Homenaje al Dr. Reynaldo Alarcón*. Lima: Universidad Ricardo Palma.
- Gamonal, Y., García, C. & Silva, Z. (2008). Síndrome de Burnout en el profesional de enfermería que labora en áreas críticas. *Revista de Enfermería Herediana*, 10(1), 33-39.
- 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), 217-226.
- Gilmer, J. S. & Feldt, L. S. (1983). Reliability estimations for a test with parts of unknown lengths. *Psychometrika*, 48, 99-111.
- Gil-Monte, P. R. (2002). Validez factorial de la adaptación al español del Maslach Burnout Inventory-General Survey. *Salud pública de México*, 44(1), 36.
- Gil-Monte, P. R. (2004). *La evaluación del síndrome de quemarse por el trabajo («burnout») en centros para personas con discapacidad: Construcción y validación test-restest del «Cuestionario para la Evaluación del Síndrome de Quemarse por el Trabajo» (CESQT-versión PD)*. Memoria anual no publicada del Proyecto de Investigación «Sd. Téc./226.04 FA/cm».
- Gil-Monte, P., Carretero, N., Desamparados, M. & Núñez-Román, E. (2005). Prevalencia del síndrome de quemarse por el trabajo (burnout) en monitores de taller para personas con discapacidad. *Revista de Psicología del Trabajo y de las Organizaciones*, 21(1-2), 107-123.
- Golembiewski, R. T., Munzenrider, R. F. & Carter, D. (1983). Phases of progressive burnout and their work site covariants: Critical issues in OD research and praxis. *Journal of Applied Behavioral Science*, 19(4), 461-481.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger.
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 13, 253-264
- ICD-10 (1994). *International classification of diseases*. Ginebra: World Health Organization.
- Juárez-García, A., García, J., Camacho, A., Gómez, V., Vera, A., Fernández, M. & García, I. (2011). *Traducción y adaptación del MBI GS en población latina*. Documento no publicado.
- Kenny, D. A., Kaniskan, B. & McCoach, D. B. (en prensa). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*.
- Kleijweg, J. H., Verbraak, M. J. & Van Dijk, M. K. (2013). The clinical utility of the Maslach Burnout Inventory in a clinical population. *Psychological Assessment*, 25(2), 435-441.
- Komaroff, E. (1997). Effect of simultaneous violations of essential tau equivalence and uncorrelated errors on coefficient alpha. *Applied Psychological Measurement*, 21, 337-348.
- Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4), 764-766.
- Lindblom, K. M., Linton, S. J., Fedeli, C. & Bryngelsson, I. L. (2006). Burnout in the working population: Relations to psychosocial work factors. *International Journal of Behavioral Medicine*, 13(1), 51-59.
- 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.
- Llaja, V., Sarria, C. & García, P. (2007). *Manual Inventario Burnout-Muestra Peruana*. Lima: Pinedo.
- Lopez, M. C., Mostafavi, B., Nibut, L. P. & Smith, W. E. (2005). Psychometric properties of the folstein mini-mental state examination. *Assessment*, 12(2), 137-144.
- Lopez, M. N., Perez, J. J., Smith, W. E. & Castillo, W. (2007). Psychometric properties of the Bender Gestalt Test using Lacks' version of the Hutt-Briskin Scoring System. *Applied Neuropsychology*, 14(4), 284-290.
- Lopez, M. N., Quan, N. M. & Carvajal, P. M. (2010). A psychometric study of the Geriatric Depression Scale. *European Journal of Psychological Assessment*, 26(1), 55-60.

* mfernandez1@usmp.pe

** sikayax@yahoo.com.ar

- Lucke, J. F. (2005). «Rassling the Hog»: the influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. *Applied Psychological Measurement*, 29(2), 106-125.
- Marshall, J. L. & Haertel, E. H. (abril, 1975). *A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement*. Reporte presentado at the Annual Meeting of the American Educational Research Association, Washington, DC.
- Maslach, C., Jackson, S. E. & Leiter, M. P. (1996). *Maslach Burnout Inventory Manual*, 3rd ed. Palo Alto, CA: Consulting Psychologists Press.
- Maslach, C. & Leiter, M. P. (2008). Early predictors of job burnout and engagement. *Journal of Applied Psychology*, 93, 498-512.
- Mateo, J. (1992). La fiabilidad y la validez desde la perspectiva criterial. En E. Abalde & J. Muñoz (Coord.), *Metodología educativa I. Jornadas de Metodología de Investigación Educativa* (pp. 77-86). A Coruña: Universidade da Coruña.
- Nunnally, J. C. & Bernstein, I. J. (1995). *Teoría psicométrica* (3^a ed). México, DF: McGraw-Hill Latinamericana.
- Olivares, V. & Gil-Monte, P. (2009). Análisis de las principales fortalezas y debilidades del «Maslach Burnout Inventory» (MBI). *Ciencia & Trabajo*, 11(33), 160-167.
- Oramas, A., González, A. & Vergara, A. (2007). El desgaste profesional. Evaluación y factorialización del MBI-GS. *Revista Cubana de Salud y Trabajo*, 8(1), 37-45.
- Osborne, J. W. & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, 9(6). Recuperado de <http://PAREonline.net/getvn.asp?v=9&n=6>
- Peters, M. J. V., Jelacic, M., Heij, C. D. H. & Merckelbach, H. (2006). Detection of feigned burnout symptoms using standard clinical questionnaires. *The German Journal of Psychiatry*, 9(1), 10-16. Recuperado de <http://www.gjpsy.uni-goettingen.de/gjp-article-peters.pdf>
- Ponce, C., Bulnes, M., Aliaga, J., Atalaya, M. & Huertas, R. (2005). El síndrome del quemado por estrés laboral asistencial en grupos de docentes universitarios. *Revista del Instituto de Investigaciones Psicológicas*, 8(2), 87-112.
- Quintana, S. M. & Maxwell, S. E. (1999). Implications of recent developments in structural equation modeling for counseling psychology. *The Counseling Psychologist*, 27, 485-527.
- Rae, G. (2006). Correcting coefficient alpha for correlated errors: Is á_k a Lower Bound to Reliability? *Applied Psychological Measurement*, 30, 56-59.
- Rim, E. & Bresler, S. (abril, 1974). *Index of efficiency: An empirical study of the two reliability coefficients for criterion-referenced tests*. Reporte presentado at a Joint Session of the American Educational Research Association and the National Council on Measurement in Education, Chicago, Illinois.
- Salanova, M. & Schaufeli, W. B. (2000). Exposure to information technologies and its relation to burnout. *Behavior & Information Technology*, 19, 385-392.
- SAS Institute Inc. (1989). *SAS/STAT user's guide*. Cary, NC: Author.
- Satorra, A. & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. En A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks.
- Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507-514.
- Schaufeli, W. B. & Van Dierendonk, D. (1995). A cautionary note about the cross-national and clinical validity of cut-off points for the Maslach Burnout Inventory. *Psychological Reports*, 76, 1083-1090.
- Schaufeli, W. B. & Enzmann, D. (1998). *The burnout companion to study and practice: A critical analysis*. Londres: Taylos & Francis.
- Schaufeli, W. B., Bakker, A., Schaap, C., Kladler, A. & Hoogduin, C. A. L. (2001). On the clinical validity of the Maslach Burnout Inventory and the Burnout Measure. *Psychology & Health*, 16, 565-82.
- Schaufeli, W. B., Leiter, M. P. & Maslach, C. (2009). Burnout: 35 years of research and practice. *Career Development International*, 14(3), 204-220.
- Schutte, N., Toppinnen, S., Kalimo, R. & Schaufeli, W. B. (2000). The factorial validity of the Maslach Burnout Inventory–General Survey across occupational groups and nations. *Journal of Occupational and Organizational Psychology*, 73, 53-66.
- Shanafelt, T. D., Bradley, K. A., Wipf, J. E. & Back, A. L. (2002). Burnout and self-reported patient care in an Internal Medicine Residency Program. *Annals of Internal Medicine*, 136(5), 358-367.
- Subkoviak, M. J. (1976). Estimating reliability from single administration of a criterion-referenced test. *Journal of Educational Measurement*, 13, 265-276.
- Thompson, G. L. (2006). An SPSS implementation of the nonrecursive outlier deletion procedure with shifting z score criterion (Van Selst & Jolicoeur, 1994). *Journal Behavior Research Methods*, 38(2), 344-352.
- Woolfolk, A. (2010). *Psicología Educativa* (11^a ed.). México: Pearson Educación.
- Zimmerman, D. W., Zumbo, B. D. & Lalonde, C. (1993). Coefficient alpha as an estimate of reliability under violation of two assumptions. *Educational and Psychological Measurement*, 53, 33-49.

* Instituto de Investigación de la Escuela Profesional de Psicología, Universidad de San Martín de Porres, Perú.

* mfernandez1@usmp.pe

** sikayax@yahoo.com.ar

LIBERABIT: Lima (Perú) 20(2): 209-218, 2014

ISSN: 1729-4827 (Impresa)

ISSN: 2233-7666 (Digital)