TERAPIA PSICOLÓGICA
2010, Vol. 28, Nº 1, 5-12

ISSN 0716-6184 (impresa) · ISSN 0718-4808 (en línea)

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)

Effect of different scoring methods on the reliability, validity, and cut points of the Center for Epidemiologic Studies Depression Scale (CES-D)

René Gempp
SIMCE [1] [2], Chile.

Claudio Thieme [2]
Facultad de Economía y Empresa
Universidad Diego Portales, Chile.

(Rec: 4 noviembre 2009 / Acep: 23 abril 2010)

Resumen

Este estudio compara el efecto de cuatro métodos de puntuación para la Escala de Depresión del Centro para Estudios Epidemiológicos (CES-D), sobre la fiabilidad, validez concurrente, puntos de corte, sensibilidad, especificidad y fiabilidad de clasificación de la escala. La CES-D fue puntuada utilizando el método "ordinal" convencional, dos métodos binarios ("presencia" y "persistencia" de los síntomas) y un nuevo sistema de puntuación "semanal". A partir de análisis psicométricos y de curvas ROC, realizados sobre una muestra normativa (n=1143) y clínica (n=44), se encontró que los métodos "ordinal" y "semanal" se desempeñan mejor en la detección de depresión. Por otro lado, el método por "persistencia" de los síntomas mostró el peor desempeño. En general, los resultados muestran que el método de puntuación tiene efecto en la fiabilidad y validez de la CES-D.

Palabras Clave: CES-D, método de puntuación, validez concurrente, curvas ROC.

Abstract

This study compare the effect of four scoring methods for the Center for Epidemiologic Studies Depression Scale (CES-D) on score reliability, concurrent validity, cut points, sensibility, specificity and classification reliability of the scale. The CES-D was scored using the conventional "ordinal" method, two binary methods ("presence" and "persistence" of symptoms) and a new "weekly" scoring system. On the basis of both psychometric analysis and receiver operating characteristic (ROC) analysis, performed on normative (n=1143) and clinical (n=44) samples, it was found that the "ordinal" and "weekly" methods performed best in detecting depression. On the other hand, the "persistence" of symptoms method resulted in worse performance. Overall, the results indicate that the scoring method has an effect on the reliability and validity of the CES-D.

Keywords: CES-D, scoring method, reliability, concurrent validity, ROC curves.

Correspondencia a: René Gempp, Equipo de Medición y Diseño de Instrumentos, SIMCE, Alameda 1146-B, Piso 7, Santiago de Chile. Email rene.gempp@mineduc.cl, rgempp@gmail.com. Web: www.sigmas.cl/rgempp

[1] La adscripción institucional del autor es sólo informativa. El contenido de este trabajo, incluyendo las opiniones que en él se expresan y los datos en que se basa, no representa ninguna línea de investigación pasada o actual del SIMCE ni ha recibido financiamiento o subvención alguna de dicha institución.

[2] Este artículo se enmarca en una línea de trabajo sobre Modelos de Ecuaciones Estructurales, desarrollada por ambos autores.
Más información sobre la interpretación del coeficiente y un software gratuito para estimarlo, en combinación con el módulo *RELIABILITY* de SPSS, puede obtenerse escribiendo al autor.

Introducción

La *Escala de Depresión del Centro para Estudios Epidemiológicos* (*Center for Epidemiologic Studies Depression Scale* [CES-D]) es un autorreporte breve (Radloff, 1977) diseñado para el tamizado rápido de sintomatología depresiva en población general. De acuerdo a la revisión de Santor, Gregus y Welch (2006), es una de las tres escalas de depresión más utilizadas en el mundo y cuenta con versiones validadas en distintos idiomas, países, grupos culturales y etáreos. En el caso de Chile, Gempp, Avendaño y Muñoz (2004) presentaron, en esta misma revista, una versión adaptada y estandarizada para población juvenil chilena, que mostró alta fiabilidad (*alfa*=0.87) y validez concurrente, entendida esta última como eficacia diagnóstica en una muestra clínica, evaluada con metodología ROC (*AUC*=0.96). En el mismo estudio, se observó que el punto de corte tradicional de la escala (16 puntos) era sensible (97.7%) pero poco específico (57%), por lo que se sugirió un punto de corte de 24 puntos para optimizar simultáneamente la sensibilidad (97.7%) y especificidad (79%) del diagnóstico en esa población.

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. Cuatro de ellos (el 4, 8, 12 y 16) son presentados de manera positiva, por lo cual es necesario invertirlos en el proceso de corrección. Las instrucciones de respuesta solicitan 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, presentado en la Tabla 1 (modalidad "ordinal"), asigna desde 0 a 3 puntos a cada alternativa, intentando valorar la presencia y gravedad de cada síntoma (Radloff y Locke, 1986), de manera que a mayor puntuación, mayor frecuencia de ocurrencia del síntoma. La puntuación total se calcula como la sumatoria simple de los ítems, pudiendo variar entre 0 a 60 puntos.

Desde su creación varios autores han propuesto modalidades alternativas para puntuar la escala, con el objetivo de mejorar su eficacia diagnóstica o propiedades psicométricas. Una de las más conocidas se origina en el trabajo de Craig y Van Natta (1976), quienes enfatizaron la importancia de diferenciar entre la "*presencia*" de los síntomas, entendida como su mera aparición durante una semana normal, y la "*persistencia*" de éstos, es decir, su manifestación reiterada durante varios días. Mientras la "*presencia*" de síntomas sería útil para la investigación epidemiológica, la "*persistencia*" de ellos resultaría más eficiente para identificar a sujetos con alto riesgo de sufrir depresión.

Aplicando estos conceptos a la CES-D, Craig y Van Natta (1979) propusieron una modalidad de puntuación por "*presencia*" de los síntomas, en que se asigna 0 puntos a la primera alternativa, indicativa de ausencia de sintomatología, y 1 punto a las restantes opciones (ver Tabla 1). 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, como indica la Tabla 1, 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.

Con los años, la práctica de diferenciar entre modalidades de puntuación "*ordinal*", por "*presencia*" y "*persistencia*" se ha vuelto habitual en los estudios de estandarización o aplicación de la CES-D (e.g. Aguilera-Guzmán, Carreño y Juárez, 2004; Salgado de Snyders & Maldonado, 1994), pese a que su efecto sobre la fiabilidad y validez de la escala no está suficientemente documentado y la poca evidencia disponible es contradictoria. Por ejemplo, López Pina (2005) comparó varios métodos de dicotomización de las alternativas de la CES-D y concluyó que cualquiera de ellos entregaba tanta información y tan fiable como el método de puntuación tradicional ("*ordinal*"). Sin embargo, López-Pina no evaluó las consecuencias sobre la validez de la CES-D, como sí lo hicieron Gelin y Zumbo (2003), quieres compararon específicamente la modalidad "*ordinal*" con los

Tabla 1. Modalidades de puntuación de la CES-D

	Modalidad de puntuación			
Alternativa	Ordinal [a]	Presencia [a]	Persistencia [a]	Semanal [b]
Rara vez o ninguna vez (1 día o menos)	0	0	0	0.5
Alguna vez o unas pocas veces (1 a 2 días)	1	1	0	1.5
Ocasionalmente o varias veces (3 a 4 días)	2	1	0	3.5
La mayor parte del tiempo (5 a 7 días)	3	1	1	6

Notas:

[a] En los ítems 4, 8, 12 y 16, la puntuación se invierte.

[b] En el caso de los cuatros ítems positivos, las alternativas son puntuadas 6.5, 5.5, 3.5 y 1.0

métodos de "*presencia"* y "*persistencia*", y demostraron que la manera en que se puntuaran las alternativas afectaba la sensibilidad de la escala para detectar diferencias válidas entre hombres y mujeres. En particular, la puntuación según "*persistencia"* acentuaba levemente la probabilidad de sesgo de los ítems. Por otro lado, en un estudio más antiguo, realizado en Japón, Furukawa et al., (1997) compararon el efecto de estos tres métodos de puntuación sobre la sensibilidad y especificidad diagnóstica de la CES-D, y concluyeron, a partir de metodología ROC, que el método "*ordinal*" mostraba mejor desempeño que los demás y que la puntuación según "*persistencia*" del síntoma resultaba poco recomendable.

En una línea de razonamiento distinta, y apelando a criterios psicométricos, McArdle, Johnson, Hishinuma, Miyamoto y Andrade (2001) han argumentado a favor de una modalidad de puntuación "*semanal*", mediante la cual se evitarían algunos de los problemas de las medidas ordinales y dicotómicas. Como se expone en la Tabla 1, en esta modalidad 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 cuatro ítems positivos, éstos reflejan "días sin depresión", así que son convertidos a "días con depresión" asignando 6.5, 5.5, 3.5 y 1.0 puntos a cada alternativa. 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. De acuerdo a McArdle et al., (2001) el método "*semanal*" es sólo una transformación monotónica de la puntuación *"ordinal*", así que preserva el ranking de las respuestas y el ordenamiento relativo de los sujetos evaluados, pero con la ventaja adicional de linearizar las puntuaciones, mejorando así los análisis psicométricos. Sin embargo, los autores de la propuesta de puntuación "*semanal"* no ofrecen una comparación sistemática con la modalidad de puntuación tradicional ("*ordinal")* ni entregan evidencia sobre su efecto en la fiabilidad y validez de la escala. Lamentablemente, tampoco hay otros estudios que avalen empíricamente la pretendida superioridad del método, o que lo comparen con las modalidades de puntuación por *"presencia*" o "*persistencia*".

¿Qué tan extendido es el uso de éstos u otros métodos de puntuación de la CES-D, en Chile? ¿Cuál es el impacto sobre la fiabilidad y validez de la escala? Aunque una búsqueda exhaustiva en publicaciones especializadas no entrega antecedentes para responder estas preguntas, la evidencia informal sugiere que el empleo de modalidades de puntuación por *"presencia*" y "*persistencia*" es, al menos, conocido. Esta presunción se funda en el hecho que desde la publicación del estudio de estandarización de Gempp et al., (2004), en el que sólo se reportaron resultados para el método de puntuación *"ordinal*", los autores han recibido periódicamente solicitudes de investigadores consultando por los puntos de corte, sensibilidad, especificidad y propiedades psicométricas para variaciones en la modalidad de puntuación de la escala.

¿Es recomendable puntuar la CES-D con una modalidad distinta a la tradicional? ¿Hay un efecto sobre la fiabilidad o validez concurrente de las decisiones diagnósticas que entregue la escala? ¿Cuál método de puntuación es más recomendable? ¿Cuál es el punto de corte óptimo para cada modalidad de puntuación? ¿Cuál es la sensibilidad y especificidad de esos puntos de corte? En nuestra opinión, responder estas preguntas con evidencia empírica y objetiva es relevante tanto desde un punto de vista teórico-psicométrico como clínico-aplicado. Desde la perspectiva teórico-psicométrica, no existen estudios que comparen sistemáticamente las modalidades de puntuación *"ordinal*", por "*presencia", "persistencia" y "semanal*", y sus consecuencias sobre las propiedades psicométricas de la CES-D, por lo que resulta importante evaluar su desempeño diferencial. Por otro lado, desde un punto de vista clínico-aplicado, la evidencia informal indica que algunos investigadores están utilizando modalidades de puntuación por "*presencia" y "persistencia*" con la versión adaptada de la CES-D para población juvenil chilena, y requieren información psicométrica robusta para orientar su trabajo.

Con el propósito de responder sistemática y objetivamente estas preguntas, este breve trabajo se propuso comparar las propiedades psicométricas de las modalidades de puntuación "*ordinal", por "presencia", "persistencia" y "semanal*" de la CES-D. Específicamente se analizan: fiabilidad de las puntuaciones, validez concurrente, puntos de corte, sensibilidad, especificidad y fiabilidad de clasificación para cada método de puntuación. Todos los análisis se basan en los datos originales utilizados en Gempp et al. (2004) y constituyen, por lo tanto, un análisis secundario. Aunque esto es obviamente una limitación del presente estudio, la revisión de publicaciones indexadas no arrojó ningún estudio de estandarización más reciente de la CES-D en Chile, así que los datos en referencia pueden considerarse todavía vigentes. Además, el uso de esos datos garantiza la comparabilidad de los resultados con el estudio de validación original, lo cual, desde un punto de vista práctico, implica que los resultados que a continuación se presentan constituyen un complemento y profundización de aquella investigación.

Método

Se utilizaron dos muestras. La *muestra normativa* estuvo integrada por 1143 jóvenes no consultantes, de ambos sexos (45.7% de hombres y 54.3% de mujeres), con una edad promedio de 20.56 años (*SD*=4.45). Para establecer la validez concurrente y fijar puntos de corte, se utilizó como criterio una *muestra clínica*, no aleatoria, definida como aquellos

pacientes diagnosticados con depresión (a lo menos desde un mes a la fecha) y que no hubieran comenzado tratamiento en el momento de la evaluación. Se muestrearon pacientes desde la población de consultantes espontáneos por problemas de salud mental en el sistema de salud pública, lográndose una muestra final de 44 pacientes (40.9% hombres y 59.1% mujeres), con edades entre los 20 y 35 años. El diagnóstico, en todos los casos, fue realizado en una entrevista psiquiátrica o psicológica especializada. De este modo, la razón entre los casos clínicos y normales es de 3.8%, mientras que en el estudio de estandarización original de la CES-D (Radloff, 1977), fue de 2.7%. Los participantes de ambas muestras respondieron el instrumento de manera anónima y no recibieron retribución por su participación.

Una descripción acabada de la metodología original del estudio de estandarización, incluyendo características de las muestras, procedimiento e instrumento puede consultarse en Gempp et al., (2004). Una copia de la escala CES-D, en formato PDF, puede solicitarse por correo electrónico al autor.

Resultados

Se estimó la fiabilidad (coeficiente *alfa* de Cronbach) y se calcularon las medias y desviaciones estándar de la CES-D, en ambas muestras, para las cuatro modalidades de puntuación. El resultado, presentado en la Tabla 2, indica que al puntuar según "*persistencia"* del síntoma se obtiene un resultado menos fiable que con las demás modalidades en la muestra normativa, mientras que la tendencia se invierte en el caso de la muestra clínica. Además, la puntuación por "*persistencia"* resulta poco apropiada para la muestra normativa, considerando su promedio y desviación estándar y, como es lógico, resulta más útil en la muestra clínica que en los sujetos no consultantes. Para todas las modalidades de puntuación las diferencias entre los promedios de la muestra normativa y clínica resultan significativos ($p<0.001$, utilizando pruebas t de student) en tanto los tamaños de efecto (*d* de Cohen) para las modalidades "*ordinal",* por "*presencia", "persistencia" y "semanal"* fueron de 2.45, 1.63, 2.59, 2.51, respectivamente.

Los resultados anteriores son confirmados cuando se comparan las distribuciones de ambas muestras según modalidad de puntuación, presentadas en la Figura 1. Para compensar el efecto de las diferencias de tamaño entre las muestras normativa y clínica, se presentan las distribuciones suavizadas con un método kernel, ajustando la densidad a cada tamaño muestral (Bowman y Azzalini, 2004). Las gráficas resultantes arrojan información muy interesante sobre cada sistema de puntuación. Por ejemplo, se observa que los métodos "*ordinal" y "semanal*" obtienen distribuciones muy similares entre sí y bastante disímiles a las modalidades según *"presencia" y "persistencia*", indicando que éstas últimas introducen algún grado de distorsión en las puntuaciones totales. Además, examinando la zona en que se intersectan las distribuciones, se puede anticipar que las modalidades "*ordinal" y "semanal*" localizarán sus puntos de corte en zonas equivalentes de la escala de puntuación, mientras en los métodos por *"presencia" y "persistencia*" los puntos de corte se ubicarán en las zonas alta y baja de la escala, respectivamente.

Otra manera de verificar el grado de equivalencia entre las cuatro modalidades de puntuación, es analizar las correlaciones entre los totales de la escala para cada una de ellas, presentadas en la Tabla 3. Corroborando los resultados previos, se observa que en las muestras

Tabla 2. Fiabilidad, media y desviación estándar para cada modalidad de puntuación en ambas muestras

	Muestra Normativa			*Muestra Clínica*		
Modalidad	*alfa*	*M*	*SD*	*alfa*	*M*	*SD*
Ordinal	0.87	16.75	10.00	0.78	41.11	8.03
Presencia	0.84	10.52	4.77	0.71	18.20	2.18
Persistencia	0.77	1.62	2.34	0.79	7.93	4.23
Semanal	0.86	1.95	0.87	0.78	4.13	0.78

Tabla 3. Correlaciones entre los totales de la CES-D obtenidos para las cuatro modalidades de puntuación, en la muestra normativa (triángulo inferior) y clínica (triángulo superior)

	Ordinal	Presencia	Persistencia	Semanal
Ordinal	-	0.76	0.83	0.99
Presencia	0.91	-	0.32	0.67
Persistencia	0.77	0.49	-	0.89
Semanal	0.99	0. 86	0.82	-

Nota:
Todas las correlaciones son significativas ($p<0.05$)

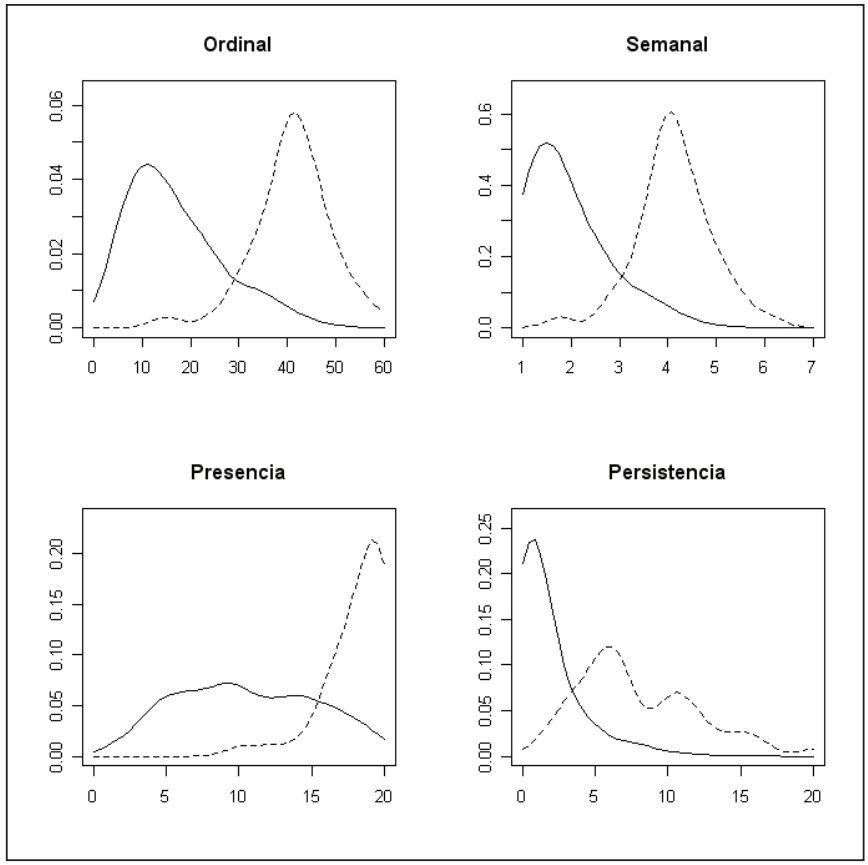

Figura 1. Distribución de frecuencias de las muestras normativa (línea continua) y clínica (línea segmentada) para las cuatro modalidades de puntuación

normativa y clínica los métodos "*ordinal" y "semanal*" correlacionan *r*=0.99, es decir, son casi idénticos en términos del ordenamiento relativo de los sujetos, tal como lo aseveran McArdle et al., (2001). Esto significa que, para dos individuos cualesquiera de la muestra, ambas modalidades concordarán entre sí sobre cuál de ellos tiene mayor riesgo de depresión. En cambio, las modalidades por "*presencia" y "persistencia*" correlacionan *r*=0.49 en la muestra normativa y sólo *r*=0.32 en la muestra clínica, indicando que un mismo sujeto podría ser identificado como más o menos depresivo que otro, dependiendo de cómo se puntúe la escala. En general, en la muestra normativa la puntuación por "*presencia*" correlaciona más alto con los métodos "*ordinal*" y "*semanal*" que la modalidad por "*persistencia*", mientras en la muestra clínica sucede exactamente lo contrario. Esto es relativamente lógico considerando que en un caso se valora la aparición del síntoma y en el otro su manifestación reiterada.

Las correlaciones observadas sólo indican que algunas modalidades de puntuación ordenan a los respondientes de manera diferente, pero no bastan, por sí solas, para concluir que alguna es superior a otra. La única manera de encontrar evidencia al respecto es utilizar un estándar externo a la CES-D para comparar el desempeño diagnóstico de la escala.

Con ese propósito, se determinó la validez concurrente y puntos de corte óptimo para cada modalidad de puntuación mediante Curvas ROC (Swetss y Pickett, 1982), utilizando como estándar la muestra clínica. Las gráficas resultantes se presentan en la Figura 2. En el eje horizontal se representa la fracción de falsos positivos (1-especificidad), y en el eje vertical la fracción de verdaderos positivos, que equivale a la sensibilidad (Swets, 1988).

En una curva ROC, el área bajo la curva (AUC) equivale a la probabilidad de que el test identifique correctamente a dos sujetos como normales o depresivos si uno de ellos fuera extraído aleatoriamente desde la muestra normal y el otro aleatoriamente de la muestra clínica. Por esta razón, el AUC es útil para cuantificar la eficacia diagnóstica de un test. En una prueba sin valor diagnóstico el AUC correspondería al área bajo la diagonal (0.50), mientras a medida que la Curva ROC se aleja de la recta diagonal, aumenta el AUC y, por lo tanto, el valor diagnóstico del test. Una prueba con AUC=1 entregaría una clasificación perfecta. Gracias a estas características, el AUC puede utilizarse como medida de validez concurrente: mientras más cercano a 1 el valor del AUC, mayor será la validez del test (Zhou, Obuchowski y McClish, 2002). Por otro lado, el punto de máxima inflexión de la Curva ROC, es decir, aquel donde la curva se acerca

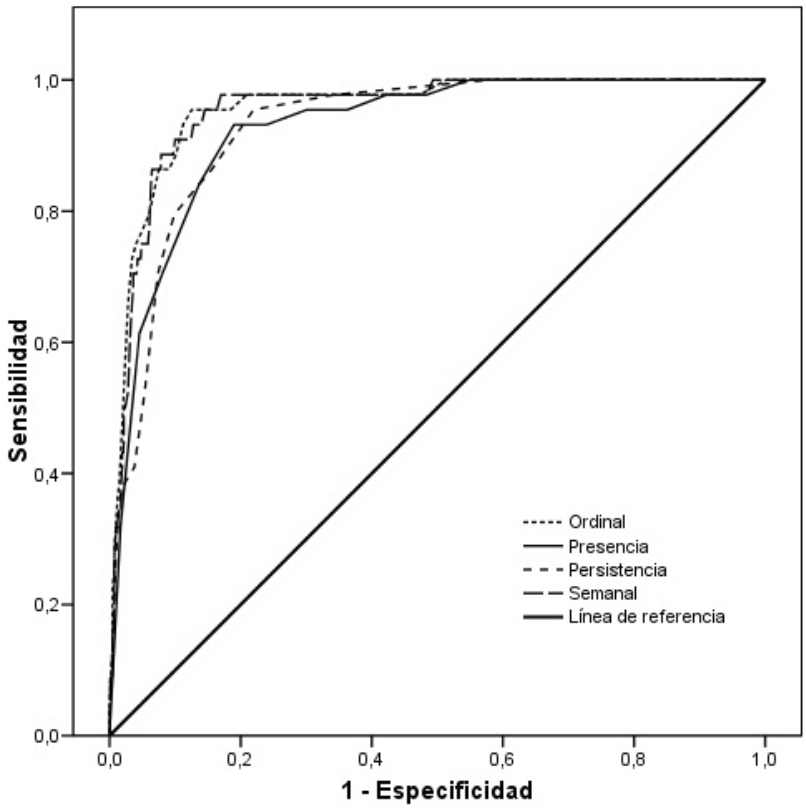

Figura 2: Curvas ROC para las cuatro modalidades de puntuación

Tabla 4. Validez concurrente, punto de corte, sensibilidad, especificidad y fiabilidad de clasificación para cada modalidad de puntuación

Modalidad	*AUC*	*P. Corte*	*Sensibilidad*	*Especificidad*	*Phi(Lambda)*
Ordinal	0.96	24	98%	79%	0.93
Presencia	0.93	12	95%	64%	0.85
Persistencia	0.93	3	86%	84%	0.84
Semanal	0.96	2.74	98%	83%	0.90

más al vértice superior izquierdo del gráfico, corresponde al valor de máxima sensibilidad y especificidad que el test puede alcanzar simultáneamente Utilizando este valor es posible identificar un punto de corte óptimo para la escala. En la Tabla 4 se presentan los coeficientes AUC para cada modalidad de puntuación, estimados con una técnica no paramétrica. En todos los casos, los errores de estimación son de 0.01. Al comparar los AUC entre sí (De Long, De Long & Clarke-Pearson, 1988), se observa que la validez concurrente de las modalidades *"ordinal"* y "*semanal"* es estadísticamente similar (Chi^2=2.21; p=0.13), y significativamente superior a la observada para los métodos de puntuación por "*presencia"* y *"persistencia*" (Chi^2=65.16; p<0.001). En la misma tabla se presentan los puntos de corte óptimos para cada modalidad, con sus correspondientes valores de sensibilidad y especificidad. Como es lógico, dados los resultados anteriores, los puntos de corte correspondientes a las modalidades "*ordinal"* y *"semanal"* muestran mayor sensibilidad diagnóstica. Cabe notar que el punto de corte según "*persistencia*" es el más específico pero menos sensible de todas las modalidades de puntuación.

Finalmente, para estimar la fiabilidad de la clasificación, se utilizó el coeficiente *Phi (Lambda)* desarrollado por Brennan y Kane (1977), en el marco de la *Teoría de la Generalizabilidad*, para estimar la fiabilidad de un punto de corte. En la práctica, *Phi (Lambda)* se interpreta como un coeficiente de fiabilidad convencional, sólo que en lugar de informar sobre la fiabilidad de las puntuaciones, indica la fiabilidad de la clasificación diagnóstica[1]. En este caso, se observa que, nuevamente, las modalidades "*ordinal*" y "*semanal"* superan a las demás, entregando clasificaciones más fiables.

Discusión

Los resultados obtenidos muestran que, a diferencia de lo reportado por López-Pina (2005), el uso de distintas modalidades de puntuación para la CES-D sí tiene un efecto importante sobre la fiabilidad, validez y utilidad diagnóstica de la escala. En concreto, los métodos "*ordinal"* y "*semanal*" exhiben un desempeño óptimo, mientras las puntuaciones por "*presencia*" y "*persistencia"* resultan menos recomendables, aún cuando sus propiedades psicométricas puedan considerarse aceptables de acuerdo a criterios convencionales.

Con respecto a la fiabilidad, se observa que ésta resulta admisible para todas las modalidades, aunque en el caso de "*persistencia*" disminuye notablemente respecto a las demás, para la muestra normativa. Esto es lógico dado que, tal como se discutió previamente, en esta modalidad se valora sólo la gravedad de la sintomatología depresiva, que se supone baja para muestras no clínicas. En ese sentido, resulta tentador argumentar que la puntuación según "*persistencia"* es más fiable en muestras clínicas, considerando que obtiene el coeficiente más elevado en esos datos. Sin embargo, si se comparan los errores estándar de medida (Gempp, 2006) de "*persistencia*" en las muestras clínica y normativa, se obtienen valores de *EEM*=1.94 y *EEM*=1.12, respectivamente, indicando que aún en la muestra clínica la puntuación por "*persistencia*" del síntoma resulta menos precisa que en la muestra normativa y que con las demás modalidades. Desde este punto de vista, y considerando que la fiabilidad es un requisito necesario para la validez y todos los análisis posteriores, la puntuación por "*persistencia*" resulta la menos recomendable de todas.

En cuanto a la validez concurrente de las dos puntuaciones dicotómicas, "*presencia*" y "*persistencia*", ambas muestran buena eficacia diagnóstica, aunque en el primer caso se logra un diagnóstico más sensible pero menos específico, y en el segundo una detección menos sensible pero más específica que con la primera modalidad. ¿Cuál criterio priorizar? ¿Sensibilidad o especificidad? Desde una perspectiva utilitarista, la respuesta depende de los costos asociados a uno u otro tipo de decisión incorrecta. En contextos en que la identificación errada de un caso positivo (i.e. baja especificidad) tiene un costo elevado para los individuos (e.g. diagnóstico incorrecto de déficit intelectual), puede resultar deseable priorizar la especificidad diagnóstica. Por el contrario, en situaciones en que la no detección de un caso verdadero (i.e. baja sensibilidad) acarrea mayores costos (e.g. no detectar riesgo suicida) es preferible priorizar la sensibilidad, aún a riesgo de aumentar la proporción de falsos positivos. En el caso de la CES-D, nos parece que los riesgos asociados a la no detección temprana de un cuadro depresivo ameritan optar por una modalidad de puntuación que garantice la máxima sensibilidad posible, con la mayor especificidad asociada. Desde el punto de vista de la validez, por lo tanto, la modalidad según "*persistencia*" nuevamente resulta menos recomendable que la puntuación por "*presencia*" del síntoma, para el tamizado de depresión. Esta conclusión es consistente con la obtenida en otros cuatro estudios en que se han comparado ambas modalidades (Cho et al., 1993; Furukawa et al., 1997; Gelin & Zumbo, 2003; Myers & Weissman, 1980).

Por otro lado, cuando se compara la fiabilidad y validez concurrente de las modalidades "*ordinal", "semanal*" y por "*presencia*", los resultados favorecen levemente a las dos primeras, lo que también es consistente con los estudios en que se ha comparado el método "*ordinal*" con la puntuación por "*presencia*" (Furukawa et al., 1997; Gelin & Zumbo, 2003; Roberts & Vernon, 1983).

Finalmente, un hallazgo novedoso de la presente investigación es que el método "*semanal*" entrega resultados prácticamente equivalentes e incluso levemente superiores al "*ordinal*", en todos los criterios considerados. ¿Cuál de ellos preferir? Nuestros datos no son conclusivos, así que una recomendación razonable es guiarse por un principio de parsimonia. Cuando se aplica la escala a unos pocos individuos resulta más cómodo y sencillo utilizar la modalidad de puntuación "*ordinal*", mientras que para una investigación propiamente tal, en que una muestra importante de casos es analizada, puede resultar interesante intentar la modalidad "semanal", especialmente si se pretende utilizar modelos de ecuaciones estructurales, que son la razón de ser de este método de puntuación (McArdle et al., 2001).

En suma, puede concluirse que la decisión de corregir la CES-D utilizando una modalidad de puntuación distinta a la tradicional tiene efectos no triviales sobre las propiedades psicométricas de la escala y, consecuentemente, sobre la calidad de la información diagnóstica que entregue. En concreto, hemos demostrado que las modalidades de puntuación "*ordinal*" y "*semanal*" son más recomendables que la puntuación por "*presencia*" del síntoma, y que la modalidad según "*persistencia*" es la menos recomendable de todas. Para cualquiera de estas modalidades, los resultados aquí presentados aportan la información psicométrica necesaria para guiar el uso de la escala y constituyen, de este modo, un complemento al estudio de estandarización original (Gempp et al., 2004). Esperamos que estos hallazgos puedan estimular otras investigaciones sobre la CES-D y sus propiedades psicométricas en nuestro país, así como el estudio y uso racional de distintas modalidades de puntuación para escalas clínicas.

Referencias

Aguilera-Guzmán, R.M., Carreño, M.S., & Juárez, F. (2004). Características psicométricas de la CES-D en una muestra de adolescentes rurales mexicanos de zonas con alta tradición migratoria. *Salud Mental*, *27*, 57-66.

Bowman, A.W., & Azzalini, A. (2004). *Applied smoothing techniques for data analysis. The kernel approach with S-plus illustrations.* New York: Oxford University Press.

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

Cho, M.C., Moscicki, E.K., Narrow, W.E., Rae, D.S., Locke, B.Z., & Regier, D.A. (1993). Concordance between two measures of depression in the Hispanic Health and Nutrition Examination Survey. *Social Psychiatry and Psychiatric Epidemiology*, *28*, 156-163.

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

Craig, T.J., & Van Natta, P.A (1979). Influence of demographic characteristics on tow measures of depressive symptoms. *Archives of General Psychiatry*, *35*, 149-154.

De Long, E.R., De Long E.M., & Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. *Biometrics, 44*, 837-845.

Furukawa, T., Anraku, K., Hiroe, T., Takahashi, K., Kitamura, T., Hirai, T., Takahashi, K., & Iida, M. (1997). Screening for depression among first-visit psychiatric patients: Comparison of different scoring methods for the Center of Epidemiologic Studies Depression Scale using operating characteristic analysis. *Psychiatry and Clinical Neurosciences, 51,* 71-78.

Gelin, M. N., & Zumbo, B.D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement*, *63*, 65-74.

Gempp, F. (2006). El error estándar de medida y la puntuación verdadera de los tests psicológicos: Algunas recomendaciones prácticas. *Terapia Psicológica, 24*, 117-130.

Gempp, R., Avendaño, C., & Muñoz, C. (2004). Normas y punto de corte para la Escala de Depresión del Centro para Estudios Epidemiológicos (CES-D) en población juvenil chilena. *Terapia Psicológica*, *22*, 146-156.

López Pina, J.A. (2005). Items politómicos vs dicotómicos: Un estudio metodológico. *Anales de Psicología*, *21*, 339-344.

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*, 108-149.

Myers, J.K., & Weissman, M.M. (1980). Use of a self-report symptom scale to detect depression in a community sample. *American Journal of Psychiatry*, *137*, 1081-1084.

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

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

Roberts, R.E., & Vernon, S.W. (1983). The Center for Epidemiologic Studies Depression Scale: Its use in a community sample. *American Journal of Psychiatry*, *140*, 41-46.

Salgado de Snyder, V.N., & Maldonado, M. (1994). Características psicométricas de la Escala de Depresión del Centro para Estudios Epidemiológicos en mujeres mexicanas adultas de áreas rurales. *Salud Pública de México*, *36*, 200-209.

Santor, D.A., Gregus, M., & Welch, A. (2006). Eight years decades of measurement in depression. *Measurement*, *4*, 135-155.

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

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

Zhou, X.H., Obuchowski, N.A. & McClish, D.K. (2002). *Statistical methods in diagnostic medicine*. NY: Wiley Interscience.