Modificación Horst al Coeficiente KR – 20 por Dispersión de la Dificultad de los Ítems

Cesar Merino Soto
Universidad Nacional Federico Villarreal, Lima, Peru
Universidad Privada San Juan Bautista, Lima, Peru
Richard Charter[1]
Gulf of the Farallones National Marine Sanctuary Advisory Council, Washington, USA
Marine Programs for Defenders of Wildlife, Washington, USA

Compendio

La fórmula KR – 20 es una técnica muy conocida de confiabilidad por consistencia interna, y es un caso especial para ítems dicotómicos desde la formulación del coeficiente alfa de Cronbach; sin embargo, es menos común conocer los presupuestos que condicionan su uso y que se basan en el modelo de relación entre las partes. Uno de estos requisitos esenciales de KR – 20 es la igualdad de la dificultad de los ítems, que corresponde al modelo equivalente tau. Dado que en la práctica es usual hallar un amplio rango de dificultad en los ítems de un instrumento, el coeficiente KR – 20 subestimará la consistencia interna. El presente artículo presenta la modificación de Horst (1953) para atenuar el impacto de la heterogeneidad de la dificultad de los ítems. Ilustramos esta técnica mediante un ejemplo, y finalizamos con una discusión sobre su uso, y limitaciones.
Palabras clave: KR-20; Confiabilidad; Psicometría; Error de medición; Cronbach.

Horst's Modification to KR – 20 Coefficient due to Dispersion of Difficulties of Items

Abstract

The KR – 20 equation is a well-known technique for internal consistency reliability, and is a special case for dichotomous items from the formulation of alpha Cronbach's coefficient. However, it is less common knowing the assumptions that determine its use, which are based in the model of relationship between parts. One of the essential requirements for KR – 20 is the equality of items difficulty, which corresponds to the equivalent tau model. Because in practice it is usual to find a wide range of difficulty on the items of an instrument, the KR - 20 coefficient will underestimate the internal consistency. This article presents the modification of Horst (1953) to mitigate the impact of the heterogeneity of the items difficulty. We demonstrate this technique with an example; and we also discuss its use and limitations.
Keywords: KR-20; reliability; Psychometry; Measurement error; Cronbach.

En la estimación de la consistencia interna mediante, por ejemplo, el coeficiente KR-20, y su forma más general que es el coeficiente alfa (Cronbach, 1951), uno de los criterios que se deben considerar para elegir el coeficiente más apropiado es la que corresponde a las características estructurales y funcionales internas de las partes, además de asumir que los datos están completos (Huynh, 1977). En este artículo nos interesará el efecto de las primeras características sobre la magnitud del coeficiente KR-20; es decir, lo relativo al número efectivo y funcional de partes (Feldt & Brennan, 1989), y modelo de medición que subyace a los elementos de la teoría clásica de los test, es decir, el puntaje verdadero y la varianza de error (Feldt & Charter, 2003). Estos modelos son de *paralelismo*, *equivalencia tau* y *congenérico* (Feldt & Brennan, 1989), y representan el funcionamiento matemático entre las partes de un instrumento. Hay una amplia literatura especializada que describe estos modelos, y que el lector puede recurrir a ellos (por ejemplo, Feldt & Brennan, 1989; Feldt & Charter, 2003; Graham, 2006; Hogan, Benjamin, & Brezinski, 2000).

Para iniciar con nuestro tema, recordemos que la fórmula ampliamente conocida como KR-20, un caso especial del coeficiente alfa de Cronbach (Martinez, 2005), es:

$$KR_{20} = \frac{n}{n-1}\left(1 - \frac{\sum p_i q_i}{\sigma^2}\right) \qquad (1)$$

En que n es el número de ítems en el instrumento, p es la proporción de estudiantes que responden correctamente al ítem i (calificado usualmente con 1), q es la

[1] Dirección: Marine Programs for Defenders of Wildlife, 1130, 17th Street, NW, Washington, DC, USA, 20036. *E-mail*: Richc446@aol.com

proporción de estudiantes que responden incorrectamente al ítem *i* (usualmente, 0) y σ^2 es la varianza del puntaje total. Un presupuesto elemental para el uso de la fórmula 20 de Kuder y Richardson (Feldt, 1965; Kuder & Richardson, 1937) es que los ítems que componen la prueba sean de igual dificultad (Guilford & Fruchter, 1978); esto significa el cumplimiento de la condición llamada *equivalente tau* (Feldt & Brennan, 1989), una condición que también condiciona el uso del coeficiente alfa (Cronbach, 1951). Tal condición asume que los puntajes verdaderos de las partes componentes de una prueba están expresados bajo una misma escala, y que son iguales entre sí; adicionalmente, esta condición acepta que la varianza de error sea diferente entre las partes. Cuando esta condición entre las partes (ítems) de un instrumento no se cumple, se subestimará la confiabilidad (consistencia interna). Por lo tanto, la elección apropiada de un coeficiente supone en primer lugar, una inspección conceptual del modelo de medición sobre el que es construido el instrumento, y en segundo lugar, la comprobación empírica del ajuste entre este modelo y las características estadísticas del instrumento.

Coeficiente KR – 20 Corregido por Dispersión en los Ítems

Para resolver el problema del incumplimiento del modelo equivalente tau en los ítems, Horst (1953) derivó un coeficiente que toma en cuenta las diferencias entre los ítems dicotómicos respecto a la dificultad, situación que sugiere una relación congenérica entre los ítems. Este coeficiente lo produjo al examinar la propuesta de Loevinger (1948) para atenuar el efecto de la dispersión de la dificultad de los ítems sobre KR-20, pero esencialmente su propuesta era una estimación de la correlación inter-ítems corregida por la variación en la dificultad en ítems dicotómicos (Horst, 1953). Horst (1953) obtuvo la siguiente fórmula:

$$r_m = \left(\frac{\sigma_t^2 - \sum pq}{\sigma_m^2 - \sum pq} \right) \frac{\sigma_m^2}{\sigma_t^2} \qquad (2)$$

σ_m^2 = varianza del puntaje total, p = dificultad del ítem, $q = 1 - p$, σ_m^2 = varianza máxima del puntaje total, que se obtiene:

$$\sigma_m^2 = 2\sum ip_i - \overline{X}\left(1 + \overline{X}\right) \qquad (3)$$

Para la obtención de la máxima varianza (σ_m^2), se requiere el puntaje promedio, y el orden de ranking (*i*) del ítem considerando su dificultad (p_i). Un paso necesario para obtener *i* es el ordenamiento de la dificultad de los ítems, que se realiza descendentemente, de tal manera que al ítem con mayor frecuencia de respuesta (el ítem más fácil), se le asigna el ranking 1, y asi sucesivamente. El siguiente ejemplo demostrará el impacto de la amplitud de la dispersión de los ítems sobre el coeficiente KR – 20, así como los pasos esenciales para el cálculo de la modificación de Horst con los datos de la Tabla 1.

Tabla 1
Estadísticos para los Ítems y Ranking de la Dificultad de los ítems: un Ejemplo

	Media	SD	Rankings
Item 1	0.80	0.40	4
Item 2	0.80	0.40	4
Item 3	0.60	0.49	6
Item 4	0.95	0.22	1
Item 5	0.50	0.50	10
Item 6	0.55	0.50	7
Item 7	0.55	0.50	7
Item 8	0.32	0.47	11
Item 9	0.92	0.26	2
Item 10	0.90	0.30	3
Item 11	0.52	0.50	9

El puntaje promedio de la prueba total es 7.420, y la varianza del puntaje total (σ_t^2) es 4.708. La suma de los valores *p* y *q* es 2.030; y la suma de los rankings multiplicados con sus respectivas dificultades de los ítems (suma $i^* p_i$) es 36.500. Con estos datos, podemos calcular la varianza máxima (σ_m^2), que es igual a 10.440 (Ecuación 3). Finalmente, el KR-20 no ajustado es igual a 0.640, pero con la modificación de Horst (Ecuación 2), el coeficiente toma un nuevo valor igual a 0.710:

$$r_m = \left(\frac{\sigma_t^2 - \sum pq}{\sigma_m^2 - \sum pq} \right) \frac{\sigma_m^2}{\sigma_t^2} = \left(\frac{4.708 - 2.031}{10.444 - 2.031} \right) \frac{10.444}{4.708} = 0.710$$

Aplicando los niveles de la cualificación de la consistencia interna, el valor original está en la región generalmente considerada por los expertos como pobre o inaceptable (Anastasi & Urbina, 1998; Cicchetti, 1994; Helms, Henze, Sass, & Mifsud, 2006; Nunnally & Bernstein, 1995), mientras que el ajuste de Horst elevó sustancialmente la estimación, y consecuentemente la cualificación de esta nueva estimación ahora aceptable según los autores citados. De este ejemplo, se debe anotar que aún cuando la consistencia interna se incrementó con el ajuste Horst, eso no hace que los puntajes sean válidos, sino que si precisión está mejor

estimada. Con un amplio rango de valores de dificultad (desde 0.320 hasta 0.950), la discrepancia entre el KR-20 modificado y no modificado es perfectamente predecible, ya que una mayor dispersión de las dificultades de los ítems derivará en una infraestimación de la consistencia interna; y lo contrario también es cierto, es decir que una menor dispersión de la dificultad de los ítems se asociará a un estimación menos sesgada por este problema. La subestimación de la confiabilidad por medio de la fórmula KR-20 cuando los ítems tienen un amplio rango de dificultades también produce el problema de que no se puede llegar al máximo nivel de confiabilidad (1.000), y por lo tanto su límite será inferior a este valor. La modificación de Horst al KR-20 permite que el máximo valor posible del coeficiente llegue a la unidad (1.000), sin ser afectado por la magnitud de la dispersión de las dificultades de los ítems. Es claro que las estimaciones de confiabilidad usando el ajuste de Horst serán siempre mayores a KR-20 mientras mayor sea la dispersión de las dificultades de los ítems que componen la prueba.

Específicamente, en una prueba con ítems cuyas dificultades varíen entre 0.100 y 0.800, el KR-20 modificado será mayor que en un instrumento cuya dispersión de sus ítems sea, por ejemplo, entre 0.300 y 0.500. Esta diferencia entre los coeficientes puede servir para a la evaluación *post hoc* indirecta del presupuesto de equivalencia tau entre los ítems, pues el monto de la discrepancia entre ambos coeficientes se incrementa en una relación directa con la dispersión de las dificultades de los ítems. Tomando el ejemplo de Guilford y Frutcher, el KR-20 fue 0.810 y la aplicación de la versión modificada de Horst llevó a un valor de 0.880. Desde un punto de vista práctico, este incremento en la confiabilidad es equivalente a un 12% de disminución en la amplitud del intervalo de confianza alrededor del puntaje obtenido en una prueba, el mismo que se obtiene del error estándar de medición.

El segundo autor del presente artículo revisó 19 libros de texto en inglés sobre teoría psicométrica y evaluación psicológica y educativa, y solo la edición del libro de Guilford y Fruchter (1978) menciona y da un ejemplo del procedimiento de Horst (existe una traducción al español de tal obra en Guilford & Fruchter, 1984, que los lectores interesados pueden acceder, publicado por la misma editorial). Complementando esta búsqueda, en todos los libros en habla hispana revisados por el primer autor, no se hizo alguna mención de la modificación de Horst. En algunas recientes revisiones y comentarios y análisis extendidos sobre el coeficiente alfa de Cronbach en el habla hispana (Barraza, 2007; Celina & Campo-Arias, 2005; Cervantes, 2005; Merino & Lautenschlager, 2003), no se hace alguna referencia a esta técnica, aun considerando que es algebraicamente equivalente al KR-20 (Feldt, 1969). El lector interesado en este procedimiento, puede solicitar una copia de un programa en MS Excel, que ahorrará el tiempo del cálculo manual. Además, se incorpora los intervalos de confianza (Feldt, Woodruff, & Salih, 1987) para este coeficiente en los niveles elegidos por el usuario.

Usos

Desde su primera exposición por Horst (1953) hasta la llamada a su uso por Charter (1995), la técnica no es un competidor en popularidad como sus ancestros (KR-20, alfa), y parece que este tipo de propuestas alternativas no podrían llegar a una popularidad debido a la facilidad de cálculo de KR-20 y/o alfa (Cliff, 1984), pues para ambos coeficientes se necesitan únicamente el número de ítems, la varianza total y la varianza de los ítems, condiciones que no son desvaloradas para el usuario no especializado.

Sin embargo, la modificación de Horst se ha utilizado en algunas investigaciones aplicadas (por ejemplo, Duran, 2005; Perkins, 1988), y en mayor frecuencia por Charter (Charter, 2000, 2001; Charter & Webster, 1997; Lopez, Charter, & Newman, 2000), quien en un breve artículo alertó de la mayor exactitud que se puede alcanzar en la estimación de la consistencia interna por este procedimiento (Charter, 1995). Por otro lado, en un análisis psicométrico de la Prueba de Categorías de Halstead (un componente de la Batería de Pruebas Neurpsicológica Halstead-Reitan, Reitan & Wolfson, 1993) conducido por Lopez et al. (2000), su Tabla 1 reporta las estimaciones de KR-20 con y sin la modificación de Horst, para cada una de las 7 subescalas y para el puntaje total. Ahí, el porcentaje de decremento en el ancho del error estándar de medición de cada puntaje varió entre 0.0% y 9.7% en las subescalas, y de 13% en el puntaje total. Un paso previo a determinar el uso del ajuste para el KR-20 puede ser la comprobación de la diferencia en las dificultades de los ítems, como lo hicieron Charter y Webster (1997), quienes probaron esta condición con un chi cuadrado para porcentajes correlacionados. Debido que las diferencias entre los ítems fueron estadísticamente significativas, Charter y Webster (1997) concluyeron que la desigual dificultad de los ítems llamó por el uso de la modificación de Horst para el coeficiente KR-20. En los estudios de Charter y Webster (1997) y Lopez et al. (2000), las diferencias entre el coeficiente modificado y no modificada pueden ser pequeñas, pero elegir el coeficiente que permite la mayor precisión puede ser un buen criterio de elección en situaciones no extrañas para el profesional o investigador aplicado, como dispone de muestras de tamaño restringido, en la práctica clínica o cuando debe calcular parámetros psicométricos que sean consistentes com la Teoria Clásica de los Test. En el libro de Guilford y Fruchter, el KR-20 calculado fue

0.810, y con la modificación de Horst fue 0.880. El aumento de la magnitud del coeficiente es 9% el tamaño del coeficiente no modificado, pero desde un punto de vista práctico, este incremento en el coeficiente es equivalente a un 21% de disminución en el ancho del intervalo de confianza al 95% para un puntaje obtenido. Aún si el coeficiente ajustado por el método de Horst fuera 0.850 (5% de aumento del coeficiente no modificado), esto representaría un decremento de 11.15% en el intervalo de confianza del puntaje. En contextos en que se deben tomar decisiones individuales basadas en los resultados de las pruebas, hallar más precisión relativa en estas estimaciones numéricas siempre es la mejor opción.

En los análisis inferenciales y descriptivos aplicables al coeficiente alfa (y su forma específica, KR-20) también se pueden usar las estimaciones ajustadas por el procedimiento de Horst, pues favorecen al investigador tomar decisiones apropiadas. Estos análisis incluyen pruebas de hipótesis e intervalos de confianza (Feldt, et al., 1987), estimación del error estándar condicional (Charter, 1996a), evaluación de las diferencias confiables y anormales (Charter, 1996b), correcciones por atenuación en los puntajes (Nunnally & Bernstein, 1995) y en los índices de magnitud del efecto (Hunter & Schmidt, 1990), entre otros.

Anotaciones Finales

El presente artículo se ha enfocado principalmente sobre una técnica de confiabilidad por consistencia interna para ítems dicotómicos, pero no desestima otras estimaciones de confiabilidad que son esenciales para el proceso de la medición que se efectúa. Por ejemplo, la principal fuente de varianza de error en pruebas que requieren el juicio del calificador (como en las las pruebas visomotoras o proyectivas) proviene desde los calificadores; y en esta situación, una correlación intraclase (Shrout & Fleiss, 1979) o coeficiente kappa (Cohen, 1960) serán más apropiados para la estimación de la varianza de error. Se deben examinar también otras fuentes de varianza de error (estabilidad, relación entre los ítems) para deducir apropiadamente las características de confiabilidad. Esta situación es un llamado a usar herramientas más amplias para un análisis de la confiabilidad como la teoría de la generalizabilidad, y que Cronbach últimamente anotó como una elección psicométricamente madura (Cronbach & Shavelson, 2004). Dado que el cálculo de este coeficiente requiere un esfuerzo adicional (la obtención de la varianza máxima y el ranking de los ítems basados en la dificultad), el lector puede solicitar al autor principal una hoja de cálculo en MS Excel que automatiza el proceso, de tal modo que únicamente se necesitarán la dificultad y la desviación estándar de cada ítem, y la desviación estándar del puntaje total.

Dado el acelerado interés y la promoción por el área la psicometría, se requieren más opciones analíticas que no queden únicamente en las oscuridades de las revistas técnicas y valorados por lectores altamente especializados, sino que su conocimiento sea extendido. Pero su divulgación no asegura su adecuado uso, como es el caso del alfa de Cronbach ha sido proclive al abuso de su aplicación (Cortina, 1998; Green, Lissitz, & Mulaik, 1977), y este mal uso estaría fuertemente condicionado por la facilidad de su cálculo y su incorporación masiva en los programas generales de análisis estadísticos. Finalmente, la mejor estimación de la consistencia interna para ítems dicotómicos no resuelve otros problemas que pueden contener los datos o las diferencias en las características de la muestra; estos problemas que pueden disminuir la estimación de la consistencia interna son, por ejemplo, la homogeneidad de la muestra, el número de ítems, la multidimensionalidad (Helms et al., 2006; Nunnally & Bersntein, 1995), o los puntajes desde diferentes poblaciones o *poblaciones mixtas* (Waller, 2006), y deberían ser apropiadamente inspeccionados por el usuario si toma bien en serio su trabajo. Finalmente, debemos señalar que la dispersión de la dificultad de los ítems no es un problema intrínseco en al construcción de una prueba, pues aunque el nivel de dificultad puede ser pragmática y teóricamente homogénea respecto a su dificultad, los autores de las pruebas pueden hacer variar intencionalmente la magnitud de la dificultad de los ítems para que converjan con el uso de la prueba (Anastasi & Urbina, 1998; Rathvon, 2004); por ejemplo, una prueba de despistaje de problemas de aprendizaje, de desarrollo psicomotor, o de acreditación, presentaran diferentes niveles de dificultad para ser más discriminativos entre los que baja habilidad o alta habilidad.

Referencias

Anastasi, A., & Urbina, S. (1998). *Test psicológicos* (7. ed.). México, DF: Pretince Hall.

Barraza, A. (2007). ¿Cómo valorar un coeficiente de confiabilidad? *Investigación Educativa Duranguense, 6*, 6-10.

Celina, H., & Campo-Arias, A. (2005). Aproximación al uso del coeficiente alfa de Cronbach. *Revista Colombiana de Psiquiatría, 34*(4), 572-580.

Cervantes, V. (2005). Interpretaciones del coeficiente alpha de Cronbach. *Avances en Medición, 3*(1), 9-28.

Charter, R. A. (1995). The under-representation of Horst's modification of the KR-20 reliability coefficient. *Perceptual and Motor Skills, 81*(3), 770.

Charter, R. A. (1996a). Revisiting the standard errors of measurement, estimate, and prediction and their application to test scores. *Perceptual and Motor Skills, 82*, 1139-1144.

Charter, R. A. (1996b). Formulas for reliable and abnormal differences in raw test scores. *Perceptual and Motor Skills, 83*, 1017-1018.

Charter, R. A. (2000). An alternate short form of the Speech-Sounds Perception Test. *Perceptual and Motor Skills, 90*(2/3), 1184-1186.

Charter, R. A. (2001). Speech-sounds perception test: Long- and short-forms reliability adjusted for item difficulty. *Perceptual and Motor Skills, 92*(1), 31-34.

Charter, R. A., & Webster, J. S. (1997). Psychometric structure of the Seashore Rhythm Test. *The Clinical Neuropsychology, 11*(2), 167-173.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and estandardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290.

Cliff, N. (1984). An improved internal consistency reliability estimate. *Journal and Educational and Behavioral Statistics, 9*(2), 151-161.

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

Cortina, J. M. (1998). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418.

Duran, A. (2005, November) *Las violencias en las escuelas: Factores que inciden en la aparición y el mantenimiento de esta conducta.* Paper presented at the II Congresso Ibero-Americano sobre Violências nas Escolas, Observatório de Violências nas Escolas-Brasil, Belém, PA.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30*, 357-370.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*, 363-373.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.

Feldt, L. S., & Charter, R. A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Methods, 8*(1), 102-109.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*(1), 93-103.

Graham, J. M. (2006). Congeneric and (Essentially) Tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930-944.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827-838.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in Psychology and Education* (6th ed.) New York: McGraw-Hill.

Guilford, J. P., & Fruchter, B. (1984). *Estadística aplicada a la Psicología y a la Educación*. México, DF: McGraw Hill

Helms, J. E., Henze, K. T., Sass, T. L., & Mifsud, V. A. (2006). Treating Cronbach's alpha reliability coefficients as data in counseling research. *The Counseling Psychologist, 34*, 630-660.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*, 523-531.

Horst, P. (1953). Correcting the Kuder-Richardson reliability formula for dispersion of item difficulties. *Psychological Bulletin, 50*, 371-374.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Huynh, H. (1977, April). *Estimation of the KR20 reliability coefficient when data are incomplete*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Loevinger, J. (1948). The technique of homogenous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45*, 507-530.

Lopez, M. N., Charter, R. A., & Newman, R. J. (2000). Psychometric properties of the Halstead Category Test. *Clinical Neuropsychologist, 14*(2), 157-161.

Martinez, R. (2005). *Psicometría: teoría de los tests psicológicos y educativos*. Madrid, España: Síntesis.

Merino, C., & Lautenschlager, G. J. (2003). Comparación estadística de la confiabilidad alfa de Cronbach: aplicaciones en la medición educacional y psicológica. *Revista de Psicología de la Universidad de Chile, 12*(2), 127-136.

Nunnally, J. C., & Bernstein, I. J. (1995). *Teoría psicométrica* (3. ed.). México, DF: McGraw-Hill.

Perkins, K. (1988). Measuring ESL readers' ability to apply reasoning in reading: A validity study of the TOEFL reading comprehension subtest. *Journal of Research in Reading, 11*(1), 36-49.

Rathvon, N. (2004). *Early reading: A practitioners handbook.* New York: Corwin.

Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychology Battery: Theory and clinical interpretation.* Tucson, AZ: Neuropsychology Press.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Waller, N. G. (2006). Commingled samples: A neglected source of bias in reliability analysis. *Applied Psychological Measurement, 32*, 211-223.

Received 09/09/2009
Accepted 05/01/2010

Cesar Merino Soto. Universidad Nacional Federico Villarreal, Lima, Peru. Universidad Privada San Juan Bautista, Lima, Peru.
Richard Charter. Gulf of the Farallones National Marine Sanctuary Advisory Council, Washington, USA. Marine Programs for Defenders of Wildlife, Washington, USA.