Psicothema 2011. Vol. 23, nº 3, pp. 516-522
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Effects of the heterogeneity of the variances on reliability generalization: An example with the Beck Depression Inventory

Juan Botella and Graciela Ponte
Universidad Autónoma de Madrid

The studies of Reliability Generalization (RG) analyze estimates of the reliability of scores from a test provided by a set of studies. As their goals and designs are usually very varied, the sampling of individuals obeys very different schemas. Thus, the variances of the scores might be more heterogeneous than expected from random sampling. Two main problems associated with this potential source of heterogeneity should be taken into account. First, heterogeneity has been usually identified subjectively, not very rigorously. Second, once identified, it has not been taken into account in subsequent analyses. In previous papers, various ways to face both problems have been proposed. The procedures are summarized and applied to a set of 65 independent studies that report estimates of the internal consistency of the Beck Depression Inventory. The results show why any study of RG should take into account the heterogeneity of the variances. In addition to this, the only source that additionally accounts for significant variance in the coefficients is the version of the test employed: the second and third versions of the test involve significant increases in the internal consistency. The consequences of ignoring the heterogeneity of the variances are discussed.

Efectos de la heterogeneidad de las varianzas en la generalización de la fiabilidad: un ejemplo con el Inventario de Depresión de Beck. Los estudios de Generalización de la Fiabilidad (GF) analizan estimaciones de la fiabilidad de las puntuaciones de un test proporcionadas por un conjunto de estudios. Como normalmente sus objetivos y diseños son extremadamente variados, sus muestreos de individuos obedecen a esquemas muy diferentes. Consecuentemente, las varianzas de las puntuaciones serán más heterogéneas de lo esperado por mera fluctuación aleatoria. Hay dos problemas principales asociados a esta fuente potencial de heterogeneidad que se deberían tener presentes. Primero, la heterogeneidad ha sido normalmente identificada subjetivamente, de forma poco rigurosa. Segundo, una vez identificada no se ha tenido en cuenta en los análisis posteriores. En trabajos previos se han propuesto formas de afrontar ambos problemas. Se resumen esos procedimientos y se aplican a un conjunto de 65 estudios independientes que aportan estimaciones de la consistencia interna del Inventario de Depresión de Beck. Los resultados muestran porqué cualquier estudio de GF debería tener en cuenta la heterogeneidad de las varianzas. Aparte de ésta, la única fuente que explica adicionalmente una varianza significativa en los coeficientes es la versión: la segunda y tercera versiones del test implican incrementos significativos de la consistencia. Se discuten las consecuencias de ignorar la heterogeneidad de las varianzas.

The meta-analyses of the reliability of the scores obtained with psychological tests, also called Reliability Generalization (RG), are relatively recent (Vacha-Hasse, 1998). Although most methodological problems that arose with the first studies of RG have been noticed and corrected (Dimitrov, 2002; Rodríguez & Maeda, 2006; Sawilowsky, 2000), some have proven quite elusive. One of the latter has been the presence, often ignored or not taken into account, of the heterogeneity in the scores' variances.

Sometimes has been pointed out its possible role, and even its relationship with the reliability coefficients (the effect sizes in RG studies) has been discussed (e.g., Barnes, Harp, & Jung, 2002; Henson, Kogan, & Vacha-Hasse, 2001; Rodríguez & Maeda, 2006). But two main problems remain unsolved. First of all, the presence of heterogeneity has been diagnosed in a subjective way. Secondly, once made a decision about its presence, it has not been taken into account in the subsequent analyses. For example, Rodriguez and Maeda (2006) have highlighted the importance of variances heterogeneity. However, they decided subjectively that the variances were heterogeneous, taking into account this fact by correcting the coefficients according to the well known formula for correcting reliability coefficients under range restriction conditions (e.g., Crocker & Algina, 1986)

Regarding to the first problem we have proposed (Botella & Suero, in press; Botella, Suero, & Gambara, 2010) that the decision about its presence can be made by analyzing the heterogeneity of the samples' means and variances. Specifically, the presence of a significant heterogeneity (larger than expected from random

Fecha recepción: 30-9-10 • Fecha aceptación: 16-2-11
Correspondencia: Juan Botella
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
e-mail: juan.botella@uam.es

fluctuations) in the samples' means and variances should be interpreted as an indicator of the presence of varied sampling schemes in the studies included in the meta-analysis.

Once the presence of heterogeneity of variances has been diagnosed, the role of other moderator variables can be analyzed, incorporating them as additional moderator variables to the variances. That is, instead of analyzing them isolated, they are added to the variances in models with multiple moderators. Specifically, Botella and Suero (in press) have argued that the formula that allows obtaining, under range restriction conditions, the reliability of a test from the reliability of the original test and its variance (Crocker & Algina, 1986; Feldt & Brennan, 1989; Lord & Novick, 1968) provides values quite approximate when applied to the coefficients of internal consistency. Previous analyses suggest that an appropriate way to analyze the role of potential moderating variables is developing a base model to predict the reliability, with the sample variance as the only predictor. Subsequently, other predictors are added to assess any significant increase in the variance accounted for the model.

Botella and Suero (in press) have also highlighted the plausibility of the linear model of equation 1 (see technical details in the appendix):

$$Log(1-\rho) = Log(\sigma_e^2) - Log(\sigma_x^2) \quad (1)$$

Model (1) has two main advantages. The first one is that it is more credible. Specifically, previous attempts to incorporate samples' variances have assumed, rather arbitrarily, a linear relationship between the reliability coefficients and the variances. On the contrary, the linear relationship between the logarithms conveyed in equation (1) is directly derived from the assumptions of the Classical Test Theory. The second advantage is that the variance of $Log(1-\rho)$ is known (Bonett, 2002). Additional moderator variables can be incorporated into this predictive model.

The purpose of the present paper is to show how this procedure can have a significant effect on the conclusions of reliability generalization studies that employ the Cronbach's alpha coefficient as the effect size for reflecting the internal consistency of the scores obtained with a test. Specifically, we show how it is applied to a set of 65 studies in which the estimates of Cronbach's alpha index of internal consistency (Cronbach, 1951) obtained with the Beck Depression Inventory are reported (BDI, Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). We chose this test because it is well known and some RG studies with it have been already published (e.g., Yin & Fan, 2000).

The BDI

The Beck Depression Inventory (BDI) is one of the most widely used instruments to assess the severity of depressive symptoms, in both clinical and nonclinical populations. It includes a somatic or physical subscale and a psychological or affective subscale.

The somatic or physical subscale includes a variety of elements (e.g., loss of pleasure, crying, loss of energy, and so on). The affective subscale also includes several elements (e.g., pessimism, past failures, guilt feelings, and others) (Beck et al, 1961). The tool consists of 21 items in a 4-point scale (0-3), so that scores range from 0 to 63 (higher total scores indicate more severe depressive symptoms).

Several versions of the test have been proposed. The Beck Depression Inventory (BDI; Beck et al, 1961) was the original version. The respondents are instructed to select the statement that seems to fit him/her the best at the present time. In some items, two alternative statements are presented and are assigned the same weight (increase and decrease of sleep and appetite).

This version was revised in 1978 (BDI-IA; see Beck, Rush, Shaw, & Emery, 1979). In this version the two alternative statements of some items were removed. Furthermore, the respondents are asked to rate how they had been feeling during the preceding two weeks.

The BDI-II (Beck, Steer, & Brown, 1996) replaced the BDI and the BDI-IA; it is based on the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM IV) of the American Psychiatric Association. It also includes 21 items in a 4-point scale. Items related to body image, work difficulty, weight loss, and somatic preoccupation, were replaced with items related to agitation, concentration difficulty, and loss of energy. Additionally, the instructions remained as in the previous version (respondents are asked to rate how they have been feeling for the past two weeks). According to Beck et al (1996), the BDI-II has been found to demonstrate high internal consistency, adequate validity and diagnostic discrimination.

The above are the three versions employed in the studies of the present meta-analysis (table 1). There are other shortened versions, as the Beck Depression Inventory for Primary Care (BDI-PC), also known as a BDI fast screen (BDI-FS), but we have not included studies that estimate the internal consistency with this version.

Method

Study selection

We included 65 studies in which an empirical estimation of the internal consistency of the scores in the BDI is reported. Most of them were also integrated in the study of Yin and Fan (2000), although not all the studies used in their paper are included, because of restricted accessibility or lack of statistical details (e.g., mean and variance of the scores). However, we do not pretend to be exhaustive neither in the inclusion of studies nor in the conclusions. The goal of our study is to give an example of the methodological advantages of the framework for RG analysis proposed by Botella and his colleagues (Botella & Suero, in press; Botella et al., 2010). The studies included in the analysis are marked with an asterisk in the reference section.

Codification of studies

Each study was coded according to several moderator variables, both contextual and methodological (Botella & Gambara, 2002). Several outcome variables were also collected, specially the Cronbach's alpha coefficient and the means and variances of the total scores. Table 1 shows the main variables recorded and summarizes the descriptive statistics for the coded variables.

Statistical analyses

Although the effect size index is the alpha coefficient itself, it was previously transformed to $Log(1-\alpha)$. The variance of this transformation (Bonett, 2002) is approximately equal to $2 \cdot J/(J-1)(N-2)$, where J is the number of items and N is the number of participants. The statistical analyses have been done with SPSS, including the macros provided by Lipsey and Wilson (2001). The

Table 1
Main variables and categories coded or calculated from the primary studies

	k	(%)
Version of the questionnaire		
BDI, original version of 1961	5	(7.7)
BDI-IA, version of 1978	24	(36.9)
BDI-II, revised version of 1996	36	(55.4)
Language		
English	48	(73.8)
Spanish	7	(10.8)
Others	10	(15.4)
Type of sample		
Students	26	(40.0)
Patients (inpatients and outpatients)	32	(49.2)
General population	5	(7.7)
Mixed	2	(3.1)
Type of patients		
Affective disorders	9	(26.5)
Others/unspecified	18	(52.9)
Various (groups with different diagnosis)	7	(20.6)

	Mean	(St. dev)	Min. - max.
Sample size (N)	359	(382)	18 - 2260
Mean age of the sample	33	(14.1)	14.7 - 74.9
Sex (% of women)	62.9	(22.8)	0.0 - 100
Mean of the total scores	14.0	(6.7)	5.9 - 34.4
Variance of the total scores	87.5	(47.2)	25 - 224
Cronbach's alpha	.881	(.038)	.77 - .95

specific component of inter-studies variance has been estimated via the method of moments. The studies have been weighted by $1/S^2$, the inverse of its variance, although here is equivalent to weighting by N, since all versions have the same number of items, J= 21.

Results

We conducted a homogeneity test to check whether the coefficients have larger heterogeneity than is expected from mere random fluctuations (Hedges & Olkin, 1985). The result is statistically significant [$Q(64)$= 3287.7, p<.0001], indicating that the coefficients have a larger variability than expected from mere random fluctuations from a fixed effects model. However, in further analysis we have always used random effects models (Hedges & Vevea, 1998).

Heterogeneity of the variances

As explained above, if the estimates included in a RG study involve very different sampling schemes, the samples' variances of the observed scores would be very heterogeneous. In this particular case it is expected to be so, given the variety of designs and the compositions of the samples in the primary studies. Let's see some examples:

a. In the study of Sun, Hui and Watkins (2006) two extreme groups were selected according to their scores on dysphoria, but the internal consistency was estimated including all participants together. The inclusion of extreme samples in a trait strongly associated with depression tends to significantly increase the variance, as there are few scores from the central part of the population's distribution.
b. The work of Dozois, Dobson and Ahnberg (1998) is a psychometric analysis of the test, conducted with a sample of psychology undergraduates.
c. Weeks and Heimberg (2005) obtained the internal consistency in an application to adult patients with Generalized Anxiety Disorder. It is sampled selectively from high values of anxiety, a variable with a well-known positive correlation with depression.
d. Grothe, Dutton, Jones, Bodenlos, Ancona and Brantley (2005) used the BDI to study a sample of low income and African American medical patients. We do not know in what way the patient selection criteria in this study represent a biased selection from the general population.

Obviously, we should not expect similar scores' variances in those samples. This heterogeneity will generate a greater heterogeneity in the coefficients of reliability than that resultant from a random sampling of the general population in all the primary studies.

The framework we have proposed (Botella & Suero, in press; Botella, Suero, & Gambara, 2010) consists of (a) identify the presence of variations in the sampling schemes employed; (b) incorporate the variation due to these schemes in the explanatory models; and (c) include other moderator variables to assess if they add significant explanatory capacity to that already provided by the sampling schemes.

Let see each of these steps. To identify the presence of a variety of sampling schemes, Botella, Suero and Gambara (2010; Botella & Suero, in press) have proposed analyzing the degree of heterogeneity of the samples' means and variances. In our case, the samples' means show a significant level of heterogeneity [$Q(64)$= 10575.9; p<.0001]; this test is enough to conclude about the presence of a varied sampling schemes. However, we have also analyzed the variances' heterogeneity. The result is convergent, as the variance also shows a significant heterogeneity [$Q(64)$= 1907.5; p<.0001]. However, as this test is vulnerable to violations of its assumptions, given that samples' variances are asymmetrically distributed (chi-square with N-1 degrees of freedom), we have reanalyzed this heterogeneity for only the studies with N≥100, so that their distribution can be considered approximately normal. The results again lead us to the conclusion that the heterogeneity is larger than expected from random fluctuations [$Q(54)$= 1838.2; p<.0001]. The conclusion is clear: the studies have been carried out under different sampling schemes, and this is certainly a source of variation in the alpha coefficients collected from the studies.

Building a model

The next step consists in the development of a base-model for the variability of the coefficients, using only the studies' variances as an explanatory variable. The weighted least squares fit of the model expressed in equation 1 gives us the result that appears in equation 2, where the only predictor is the sample variance, and Y' represents the prediction made by the model for the $Log(1-\alpha)$, as expressed in Equation 1. The model explains a significant proportion of the variance in the coefficients [$Q_R(1)$= 38.357; p<.0001; specific variance component of the model, v= 0.05125; R^2= .348].

$$Y' = -0.691 - 0.344 \cdot Log(S^2) \quad (2)$$

Given the conclusion reached about the sampling schemes, it should not be accepted for this set of coefficients any model that does not include the sampling variance as an explanatory variable. Other models below are developed by including additional moderators to this base-model. Only if the percentage of variance accounted for the model increases significantly, it is acceptable.

We tested a high number of models, arising from the combination of predictors. Our final model (a mixed effects model) includes only the test version, besides the sample variances (the weighted average reliability is .827, .873 and .901 for the studies employing the three versions of the test, respectively). The model is specified in equation 3; the versions of the test are coded with ordered values (1, 2, 3, for the first, second, and third versions, respectively). Naturally, it shows a significant result [$Q_R(1)= 88.918$; $p<.0001$; specific variance component of the model, $v= 0.03202$]; the proportion of variance accounted for the model rises to $R^2= .525$.

$$Y' = -0.422 - 0.284 \cdot Log(S^2) - 0.213 \cdot V_i \quad (3)$$

We have tested whether the increase in the explained variance is significant using a statistic for nested models (Judd & McClelland, 1989; Maxwell & Delaney, 1990). The result is significant [$F(1,62)= 36.53$; $p<.001$], so that we can conclude that the inclusion of the version increases significantly the explained variability. Furthermore, the tests for models that include other additional moderators have not shown significant increments in the explanatory power[1].

In summary, the most parsimonious model is one that includes two predictors. The first one is the samples' variances, which we had already decided that should be included because the homogeneity tests of the means and variances have shown statistically significant. The second one is the version of the test, distinguishing between the 1961 (BDI), the 1978 (BDI-IA), and the 1996 (BDI-II) versions. The model establishes that the Cronbach's alpha increases in the BDI-IA version, as compared to the first version. Specifically, for a sample with variance equal to 75 the predicted alpha increases from .845 to .875; in a similar vein, it increases from .875 to .900 if the BDI-II (third version) is employed instead of the BDI-IA (second version).

Consequences of ignoring the heterogeneity of the variances

Although some previous studies of RG have pointed out a significant association between the reliability coefficients and the samples' variances, it rarely has been taking into account for further analysis. Let's see where this approach would have taken us.

We have fit models with several moderators, and the main significant outcomes are those related to categorical models, as Sample type and Language, and also to the continuous model created with Sex. Regarding to the moderator Sample type, it explains a significant part of the variance [$Q_B(3)= 9.158$; $p<.03$]; the participants selected from the general population show a higher internal consistency (.900), then general patients (.896), and finally the group with «various» (different groups) and the students (.874 and .873, respectively). Language also shows a significant result when the category 'English' (the language of the original version) is compared to the category 'other languages' [$Q_B(1)= 4.710$; $p<.04$], although the decrease is small; the studies with the English version have associated an average coefficient of .891, whereas it is reduced to .873 in the studies with other groups.

Sex (percentage of women) also shows a significant result [$Q_R(1)= 5.542$; $p<.02$]. As the negative slope indicates, a higher percentage of women is associated with a lower internal consistency; however the percentage of explained variance is rather small ($R^2= .075$).

An analysis that ignores the samples' variances lead to conclude that all of the above moderators are relevant to predict (or explain) variations in the coefficients. However, these variables also show significant associations with the samples' variances. The variance heterogeneity explains in a more parsimonious way the apparent association among them and the internal consistency.

Discussion

When in a meta-analysis of RG the variations in the sampling schemes of the studies' samples are ignored, the conclusions can be severely misguided. The presence of varied sampling schemes must be diagnosed by testing for significant heterogeneity on samples' means and variances, not in a subjective way. This has been shown in a practical example with the BDI. The results support the ideas developed in our previous work (Botella & Suero, in press; Botella, Suero, & Gambara, 2010).

Once decided the presence of various sampling schemes, the subsequent analysis should incorporate the samples' variances as an explanatory variable. Any other moderator variable should be added to it, not replace it; otherwise, it could be capitalizing a relationship between the moderator variable and specific sampling schemes.

We have applied this analysis framework to 65 estimates of the internal consistency (Cronbach's alpha) of the Beck Depression Inventory (BDI), concluding that the most parsimonious model to explain the variations in the coefficients is one that includes as predictors (or moderators) the samples' variances and the test version.

These findings mean that the observed variability in the internal consistency coefficients does not imply a differential functioning of the test according to the sample characteristics, such as the patient status, the average age, sex, or other. The main source of variation is associated to the sampling of true scores (in the sense of the Classical Test Theory; Lord & Novick, 1968). According to a well known relationship in psychometrics, the more heterogeneous is the sample, the larger is the coefficient. But this does not mean that the psychometric quality of the individual scores is different. The alpha coefficient is a group statistic and its value depends on a circumstantial characteristic, the heterogeneity of the sample, which has nothing to do with the measurement, but much with the decisions made for the sampling of the participants.

The second source of variation in the data is the test version. Successive versions of the BDI (1961, 1978 & 1996) have been developed with the aim of improving the psychometric quality of the test. Our findings support that this goal has been accomplished in each new version, or at least in the three versions we have included in our study. Each new version has conveyed a significant increment in the internal consistency of the scores.

In general, in the field of meta-analysis are often welcome the studies that show that there are moderator variables associated with effect sizes of different magnitude. These refinements improve our understanding of the phenomena under study, and point out the lines for future research (Borenstein, Hedges, Higgins, & Rothstein,

2009; Cooper, 1998; Cooper & Hedges, 1994; Lipsey & Wilson, 2001). However, in RG studies what is welcome is the opposite result. The absence of moderator variables that explain variations in the coefficients of internal consistency is good news for the test. In this sense, our results point that the BDI is an instrument with high psychometric quality, and that can be generalizable to a variety of sample types and stages of implementation, such as those included in this study. Moreover, successive attempts to improve the quality with new versions have proved fruitful, with each new version resulting in improved internal consistency.

Note

[1] Following the suggestion of one of the reviewers we have reanalyzed the data including the versions of the test as two fictitious coding variables. In the first one the value 1 is associated to version 2, whereas in the second one it is associated to version 3. The main consequence of this alternative way to manage the version is that it is lost one degree of freedom when the model is compared with the base model. However, the conclusion does not change, as the statistic is still statistically significant [$F(2,61)= 18,08$; $p<.001$].

Acknowledgements

This research has been financially supported by the Ministerio de Ciencia y Tecnología of Spain, Project SEJ2006-12546/PSIC. Thanks are due to the two anonymous reviewers of *Psicothema*, that helped to improve significantly the manuscript.

Appendix

Classical Test Theory assumes that the observed score (X) can be defined as the true score (T) plus some random error (E), and several distributional assumptions (Crocker & Algina, 1986). The reliability is defined as,

$$\rho = \frac{\sigma_T^2}{\sigma_X^2} \quad \text{(A1)}$$

but the assumptions allow express it also as,

$$\rho = 1 - \frac{\sigma_E^2}{\sigma_X^2} \quad \text{(A2)}$$

Under range restriction conditions the variance is different (σ'^2_E), but as the variance of the errors is constant, the reliability under those conditions (σ') is,

$$\rho' = 1 - \frac{\sigma_E^2}{\sigma'^2_X} \quad \text{(A3)}$$

When the studies included in a RG meta-analysis involve varied sampling schemes that generate different variances, the expected reliability for each one can be obtained from (A3); this equation can be expressed as,

$$1 - \rho' = \frac{\sigma_E^2}{\sigma'^2_X} \quad \text{(A4)}$$

Taking logarithms we reach equation (1),

$$Log(1\text{-}\rho') = Log(\sigma_E^2) - Log(\sigma'^2_X)$$

References

(References marked with an asterisk indicate studies included in the meta-analysis).

* Aasen, A. (2001). An empirical investigation of depression symptoms: Norms, psychometric characteristics and factor structure of the Beck Depression Inventory II. Hovedfagsoppgave i psykologi. Universitetet i Bergen.

* Al-Musawi, N.M. (2001). Psychometric properties of the Beck Depression Inventory-II with university students in Bahrain. *Journal of Personality Assessment, 77*(3), 568-579.

* Arnau, R.C., Meagher, M.W., Norris, M.P., & Bramson, R. (2001). Psychometric evaluation of the Beck Depression Inventory II with primary care medical patients. *Health Psychology, 20*(2), 112-119.

Barnes, L.L.B., Harp, D., & Jung, W.S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement, 62*, 603-618.

* Beck, A.T., Steer, R.A., Ball, R., & Ranieri, W.F. (1996). Comparison of Beck Depression Inventories-IA and II in psychiatric outpatients. *Journal of Personality Assessment, 67*(3), 588-597.

Beck, A.T., Rush, A.J., Shaw, B.F., & Emery, G. (1979). *Cognitive Therapy of Depression*. Guilford, New York.

Beck, A.T., Steer, R.A., & Brown, G.K. (1996). *Manual for Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 53-63.

* Bernal, G., Bonilla, J., & Santiago, J. (1995). Confiabilidad interna y validez de construcción lógica de dos instrumentos para medir sintomatología psicológica en una muestra clínica: el Inventario de Depresión de Beck y la Lista de Cotejo de Síntomas-36. *Revista Latinoamericana de Psicología*, *27*(2), 207-229.

Bonett, D.G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*, 335-340.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley and Sons.

Botella, J., & Gambara, H. (2002). *Qué es el meta-análisis*. Madrid: Biblioteca Nueva.

Botella, J., & Suero, M. (in press). Managing heterogeneity of variances in studies of internal consistency generalization. *Methodology*.

Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods*, *15*(4), 386-397.

* Buhlmann, U., Etco, N.L., & Wilhelm, S. (2006). Emotion recognition bias for contempt and anger in body dysmorphic disorder. *Journal of Psychiatric Research, 40,* 105-111.

* Cahill, J., Barkham, M., Stiles, W.B., Twigg, E., Hardy, G.E., Rees, A., & Evans, C. (2006). Convergent validity of the CORE measures with measures of depression for clients in cognitive therapy for depression. *Journal of Counseling Psychology*, *53*(2), 253-259.

* Carmody, D.P. (2005). Psychometric characteristics of the Beck Depression Inventory-II with college students of diverse ethnicity. *International Journal of Psychiatry in Clinical Practice, 9*(1), 22-28.

* Coelho, R., Martins, A., & Barros, H. (2001). Clinical profiles relating gender and depressive symptoms among adolescents ascertained by the Beck Depression Inventory II. *European Psychiatry, 17,* 222-226.

* Coles, M.E., Gibb, B.E., & Heimberg, R.G. (2002). Psychometric evaluation of the Beck Depression Inventory in adults with social anxiety disorder. *Depression and Anxiety, 14,* 145-148.

Cooper, H. (1998). *Synthesizing research* (3ª ed.). Thousand Oaks, CA: Sage pub.

Cooper, H., & Hedges, L.V. (1994). *The handbook of research synthesis*. Nueva York: Russell Sage Foundation.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

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

Dimitrov, D.M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62*(5), 783-801.

* Dozois, D., Dobson, K.S., & Ahnberg, J.L. (1998). A psychometric evaluation of the Beck Depression Inventory II. *Psychological Assessment, 10*(2), 83-89.

* Endler, N.S., Rutherford, A., & Denisoff, E. (1999). Beck Depression Inventory: Exploring its dimensionality in a nonclinical population. *Journal of Clinical Psychology, 55*(10), 1307-1312.

* Falck, R.S., Wang, J., Carlsin, R.G., & Siegal, H.A. (2006). Prevalence and correlates of current depressive symptomatology among a community sample of MDMA users in Ohio. *Addictive Behaviors, 31,* 90-101.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (ed.), *Educational measurement* (pp. 105-146). Nueva York: MacMillan.

* Geisner, I.M., Neighbors, C., & Larimer, M.E. (2006). A randomized clinical trial of a brief, mailed intervention for symptoms of depression. *Journal of Consulting and Clinical Psychology*, *74*(2), 393-399.

* Ghassemzadeh, H., Mojtabai, R., Karamghadiri, N., & Ebrahimkhani, N. (2005). Psychometric properties of a Persian language version of the Beck Depression Inventory second edition: BDI II Persian. *Depression and Anxiety, 21,* 185-192.

* Gorenstein, C., Andrade, L., Guerra Vieira Filho, A.H., Tung, T.C., & Artes, R. (1999). Psychometric properties of the Portuguese version of the Beck Depression Inventory on Brazilian college students. *Journal of Clinical Psychology, 55*(5), 553-562.

* Grothe, K.B., Dutton, G.R., Jones, G.N., Bodenlos, J., Ancona, M., & Brantley, P.J. (2005). Validation of the Beck Depression Inventory II in a low income African American sample of medical outpatients. *Psychological Assessment, 17*(1), 110-114.

Hedges L.V., & Olkin, I. (1985). *Statistical methods of meta-analysis*. Orlando, Academic Press.

Hedges, L.V., & Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

* Hee Yoo, S., Matsumoto, D., & LeRoux, J.A. (2006). The influence of emotion recognition and emotion regulation on intercultural adjustment. *International Journal of Intercultural Relations, 30,* 345-363.

Henson, R.K., Kogan, L.R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement*, *61*, 404-420.

* Hewitt, P.L., & Norton, G.R. (1993). The Beck Anxiety Inventory: A psychometric analysis. *Psychological Assessment, 5*(4), 408-412.

Judd, C.M., & McClelland, G.H. (1989). *Data analysis: A model comparison approach*. San Diego, CA: Horcourt Brace Jovanovich.

* Kashdan, T.B., Elhai, J.D., & Frueh, B.C. (2006). Anhedonia and emotional numbing in combat veterans with PTSD. *Behaviour Research and Therapy, 44*, 457-467.

* Kim, Y., Pilkonis, P.A., Frank, E., Thase, M.E., & Reynolds, T.F. (2002). Differential functioning of the Beck Depression Inventory in late-life patients: Use of Item Response Theory. *Psychology and Aging*, *17*(3), 379-391.

* Kojima, M., Furukawa, T.A., Takahashic, H., Kawai, M., Nagaya, T., & Tokudome, S. (2002). Cross-cultural validation of the Beck Depression Inventory-II in Japan. *Psychiatry Research, 110,* 291-299.

* Krefetz, D.G., Steer, R.A., & Gulab, N.A. (2002). Convergent validity of the Beck Depression Inventory-II with the Reynolds Adolescent Depression Scale in psychiatric inpatients. *Journal of Personality Assessment, 78*(3), 451-460.

* Lewandowski, K.E., Barrantes-Vidal, N., Nelson-Gray, R.O., Clancy, C., Kepley, H.O., & Kwapil, T.R. (2006). Anxiety and depression symptoms in psychometrically identified schizotypy. *Schizophrenia Research, 83,* 225-235.

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage pub.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental tests scores*. Reading, Mass.: Addison-Wesley.

* Luciano, J.V., Algarabel, S., Tomás J.M., & Martínez, J.L. (2005). Development and validation of the thought control ability questionnaire. *Personality and Individual Differences, 38,* 997-1008.

* Manne, S., Ostroff, J., Martini, R., Mee, L., Sexson, S., Nereo, N., DuHamel, K., Parsons, S., Williams, S., Lewis, J., Vickberg, S.J., &. Redd, W.H. (2001). Anxiety and depression in mothers of children undergoing bone marrow transplant: Symptom prevalence and use of the Beck Depression and Beck Anxiety Inventories as screening instruments. *Journal of Consulting and Clinical Psychology, 69*(6), 1037-1047.

Maxwell, S.E., & Delaney, H.D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

* McDade-Montez, E.A., Christensen, A.J., Cvengros, J.A., & Lawton, W.J. (2006). The role of depression symptoms in dialysis withdrawal. *Health Psychology, 25*(2), 198-204.

* Miles, A., McManus, C., Feinmann, C., Glover, L., Harrison, S., & Pearce, S. (2001). The factor structure of the BDI in facial pain and other chronic pain patients: A comparison of two models using confirmatory factor analysis. *British Journal of Health Psychology, 6,* 179-196.

* Moulds, M.L., & Kandris, E. (2006). The effect of practice on recall of negative material in dysphoria. *Journal of Affective Disorders, 91*, 269-272.

* Osman, A., Kopper, B.A., Barrios, F., Gutiérrez, P.M., & Bagge, C.L. (2004). Reliability and validity of the Beck Depression Inventory-II with adolescent psychiatric inpatients. *Psychological Assessment, 16*(2), 120-132.

* Otto, K., Boos, A., Dalbert, C. Shöps, D., & Hover, J. (2006). Posttraumatic symptoms, depression and anxiety of flood victims: The impact of the belief in a just world. *Personality and Individual Differences, 40,* 1075-1084.

* Penley, J.A., Wiebe, J.S., & Nwosu, A. (2003). Psychometric properties of the Spanish Beck Depression Inventory-II in a medical sample. *Psychological Assessment, 15*(4), 569-577.

* Radnitz, C.L., McGrath, R.E., Tirch, D.D., Willard, J., Strumolo, L.P., Festa, J., Binks, M., Broderick, C.P., Schlein, I.S., Walczak, S., & Lillian, L.B. (1997). Use of the Beck Depression Inventory in veterans with spinal cord injury. *Rehabilitation Psychology, 42*(2), 93-101.

* Rassin, E., & van Rootselaar, A.F. (2006). From dissociation to trauma? Individual differences in dissociation as predictor of 'trauma 'perception. *Journal of Behavior Therapy and Experimental Psychiatry, 37*, 127-139.

Rodríguez, M.C., & Maeda, Y. (2006). Meta-analysis of coefficient Alpha. *Psychological Methods, 11*(3), 306-322.

* Ruscio, A.M., & Ruscio, J. (2002). The latent structure of analogue depression: Should the Beck Depression Inventory be used to classify groups? *Psychological Assessment, 14*(2), 135-145.

* Sanz, J., & Vázquez, C. (1998). Fiabilidad, validez y datos normativos del Inventario para la Depresión de Beck. *Psicothema, 10*(2), 303-318.

* Sanz, J., Navarro, M.E., & Vázquez, C. (2003). Adaptación española del Inventario para la Depresión de Beck-II (BDI-II): 1. Propiedades psicométricas en estudiantes universitarios. *Análisis y Modificación de Conducta, 29*(124), 241-288.

* Sato, T., & McCann, D. (2000). Sociotropy-autonomy and the Beck Depression Inventory. *European Journal of Psychological Assessment, 16*(1), 66-76.

Sawilowsky, S.S. (2000). Psychometrics versus datametrics: Comment on Vacha-Hasse's «reliability generalization» method and some EPM editorial policies. *Educational and Psychological Measurement, 60*, 157-173.

* Skorikov, V.B., & Vandervoort, D.J. (April, 2003). Relationships between the underlying constructs of the Beck Depression Inventory and the Center for Epidemiological Studies Depression Scale. *Educational and Psychological Measurement, 63*(2), 319-335.

* Sloan, D.M., Marx, B.P., Bradley, M.M., Strauss, C.C., Lang, P.J., & Cuthbert, B.C. (2002). Examining the high-end specificity of the Beck Depression Inventory using an anxiety sample. *Cognitive Therapy and Research, 26*(6), 719-727.

* Sprinkle, S.D., Lurie, D., Insko, S.L., Atkinson, G., Jones, G.L., Logan, A.R., & Bissada, N.N. (2002). Criterion validity, severity cut scores, and test-retest reliability of the Beck Depression Inventory-II in a University Counseling Center sample. *Journal of Counseling Psychology, 49*(3), 381-385.

* Steer, R.A., Ball, R., Ranieri, W.F., & Beck, A.T. (1999). Dimensions of the Beck Depression Inventory-II in clinically depressed outpatients. *Journal of Clinical Psychology, 55*(1), 117-128.

* Steer, R.A., Clark, D.A., Beck, A.T., & Ranieri, W.F. (1998). Common and specific dimensions of self-reported anxiety and depression: The BDI-II versus the BDI-IA. *Behaviour Research and Therapy, 37,* 183-190.

* Steer, R.A., Ranieri, W.F., Kumar, G., & Beck, A.T. (2003). Beck Depression Inventory-II items associated with self-reported symptoms of ADHD in adult psychiatric outpatients. *Journal of Personality Assessment*, *80*(1), 58-63.

* Steer, R.A., Rissmiller, D.J., & Beck, A.T. (2000). Use of the Beck Depression Inventory-II with depressed geriatric inpatients. *Behaviour Research and Therapy, 38,* 311-318.

* Stice, E., Orjada, K., & Tristan, J. (2006). Trial of a psychoeducational eating disturbance intervention for college women: A replication and extension. *International Journal of Eating Disorders, 39*(3), 233-239.

* Storch, E.A., Roberti, J.W., & Roth, D.A. (2004). Factor structure, concurrent validity and internal consistency of the Beck Depression Inventory-second edition in a sample of college students. *Depression and Anxiety, 19,* 187-189.

* Sun, R.C.; Hui, E.K., & Watkins, D. (2006). Towards a model of suicidal ideation for Hong Kong Chinese adolescents. *Journal of Adolescence, 29,* 209-224.

* Tolin, D.F., Worhunsky, P., & Maltby, N. (2006). Are «obsessive» beliefs specific to OCD? A comparison across anxiety disorders. *Behaviour Research and Therapy, 4,* 469-480.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58,* 6-20.

* Vázquez, C., & Sanz, J. (1997). Fiabilidad y valores normativos de la versión española del Inventario para la Depresión de Beck, 1978. *Clínica y Salud, 8*(3), 403-422.

* Vázquez, C., & Sanz, J. (1999). Fiabilidad y validez de la versión española del Inventario para la Depresión de Beck de 1978 en pacientes con trastornos psicológicos. *Clínica y Salud, 10*(1), 59-81.

* Viinamaki, H., Tanskanen, A., Honkalampi, K., Koivumaa-Honkanen, H., Haatainen, K., Kaustio, O., & Hintikka, J. (2004). Is the Beck Depression Inventory suitable for screening major depression in different phases of the disease? *Nordic Journal Psychiatry*, *58,* 49-53.

* Walter, L.J., Maresman, J.F., Kramer, T.L., & Evans, R.B. (2003). The Depression-Arkansas Scale: A validation study of a new brief depression scale in a HMO. *Journal of Clinical Psychology, 59*(4), 465-481.

* Weeks, J.W., & Heimberg, R.G. (2005). Evaluation of the psychometric properties of the Beck Depression Inventory in a non-elderly adult sample of patients with generalized anxiety disorder. *Depression and Anxiety, 22,* 41-44.

* Whisman, M.A., Pérez, J.E., & Ramel, W. (2000). Factor structure of the Beck Depression Inventory-second edition (BDI-II) in a student sample. *Journal of Clinical Psychology, 56*(4), 545-551.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*, 201-223.