© International Journal of Clinical and Health Psychology

ISSN 1697-2600
2009, Vol. 9, Nº 1, pp.143-159

The Hamilton Rating Scale for Depression: A meta-analytic reliability generalization study

José A. López-Pina[1], Julio Sánchez-Meca, and Ana I. Rosa-Alcázar
(*Universidad de Murcia, Spain*)

(Received December 10, 2007 / Recibido 10 de diciembre 2007)
(Accepted July 10, 2008 / Aceptado 10 de julio 2008)

ABSTRACT. Reliability generalization is a meta-analytic approach to study how reliability estimates of a test scores depend on the specific characteristics under which the test is applied and, as a consequence, the risks of inducing score reliability from previous applications of the test. The Hamilton Rating Scale for Depression (HAM-D) is one the most popular measurement instruments in clinical psychology to assess depressive symptoms and several versions of the scale have been designed. The present meta-analytic study provided a reliability generalization (RG) study of the HAM-D scale for estimating the typical measurement reliability, to test the heterogeneity of reliability estimates across studies, to examine the influence of study characteristics and to compare the results with those obtained in previous RG studies on other depression scales. Analyses carried out with 35 alpha coefficients, obtained from 23 published research studies, showed a mean reliability of 79 (*SD* = 14), high heterogeneity across studies and several study characteristics related to score reliability, mainly the number of items, the variability of the test scores and the type of disorder studied in the sample. Implications for researchers and clinicians using the HAM-D scale are discussed.

KEYWORDS. Reliability generalization. Hamilton Depression Rating Scale. Measurement reliability. Internal consistency. Meta-analysis.

[1] Correspondencia: Departamento de Psicología Básica y Metodología. Facultad de Psicología. Campus de Espinardo. Universidad de Murcia. 30100 Murcia (Spain). E-mail: jlpina@um.es

RESUMEN. La generalización de la fiabilidad es una aproximación meta-analítica para estudiar cómo las estimaciones de la fiabilidad a partir de las puntuaciones de los tests dependen de las características específicas en las que el test se aplica y, en consecuencia, el riesgo de inducir la fiabilidad de las puntuaciones a partir de aplicaciones previas del test. La Escala de Evaluación de la Depresión de Hamilton (HAM-D, por su nombre en inglés) es uno de los instrumentos de medida más populares de la psicología clínica para evaluar los síntomas depresivos y de la que se han construido algunas versiones. El presente estudio meta-analítico es un estudio de generalización de la fiabilidad (GF) para probar la heterogeneidad de las estimaciones de la fiabilidad a través de los estudios, examinar la influencia de distintas características y compara los resultados con los obtenidos en estudios previos de GF en otras escalas de depresión. Los análisis llevados a cabo con 35 coeficientes alfa obtenidos a partir de 23 estudios publicados mostraron una fiabilidad media de 0,79 (*DT* = 0,14), elevada heterogeneidad a través de los estudios y que algunas características de los estudios influyeron en la fiabilidad de las puntuaciones, principalmente el número de ítems, la variabilidad de las puntuaciones y el tipo de trastorno estudiado en la muestra. Además, se discuten las implicaciones para investigadores y clínicos cuando se utiliza la escala HAM-D.

PALABRAS CLAVE. Generalización de la fiabilidad. Escala de Evaluación de la Depresión de Hamilton. Fiabilidad de la medida. Consistencia interna. Meta-análisis.

The Hamilton Rating Scale for Depression (HAM-D) is one of the most popular depression assessment instruments among the clinician scales in the field of Clinical and Health Psychology, together with the Beck Depression Inventory and other depression scales (Bentz and Hall, 2008; Cabañero-Martínez, Cabrero-García, Richart-Martínez, Muñoz-Mendoza, and Reig-Ferrer, 2007). The first version was published by Max Hamilton in 1960. He designed the scale as a measure of the severity of depression in previously diagnosed depressed inpatients (Hamilton, 1960). Since then different versions have been developed. Although people usually use the 17-item version, the original version had twenty-one items but Hamilton himself decided that the last four items (diurnal variation, depersonalization/derealization, paranoid symptoms, and obsessional and compulsive symptoms) should not be considered part of the disease, because they are not as frequent as the others and therefore should not contribute to the total score. There is another version in which three new items have been added, to make the 24-item version: helplessness, hopelessness, and worthlessness (Paykel, 1985; Rosenthal and Klerman, 1966). Moreover, there are some derivative scales aimed at expanding or reducing item sets. Some authors have explained that the multidimensionality of the HAM-D limits its use as a precise measure of depression severity (Bech and Allerup, 1981; Bech, Allerup, Reisby, and Gram, 1984; Gibbons, Clark, and Kupfer, 1993). This has led to the development of scales derived from a reduced item set. Moreover, other researchers have expanded the list of HAM-D items to include symptoms seen in atypical depression (Gelenberg *et al*., 1990; Paykel, 1985; Terman, 1988; Thase, Frank, Malinger, Hamer, and Kupfer, 1992; Williams, 1988; Williams, Link, Rosenthal, Amira, and Terman, 2000).

The existence of different versions of the HAM-D scale, with different formats and numbers of items, as well as its wide application to different populations and settings in psychological research, justify the convenience of examining whether its psychometric properties, and in particular the score reliability, can be generalized across studies that have used this scale. To accomplish this objective we carried out a reliability generalization (RG) study. Basically, an RG study is a meta-analysis where reliability estimates are substituted for effect sizes. An RG study requires all information available on a test or specific psychological scale to be gathered over a period of time, which generally would run from the first publication up to a given moment. In an RG study, reliability estimates obtained across studies are used as the dependent variable, the sample and instrument features of the studies are used as predictors, and their relationships are examined to explain the variability exhibited by the reliability coefficients (Beretvas and Pastor, 2003; Botella and Gambara, 2006; Henson and Thompson, 2002; Mason, Allam, and Brannick, 2007; Rodriguez and Maeda, 2006; Thompson, 2003; Vacha-Haase, 1998). Although inducing the reliability from previous applications of the test is a common practice, fortunately there are also researchers that do not follow this practice, instead calculating reliability coefficients from the subject sample itself. This enables the development of RG studies by quantitatively integrating reliability estimates obtained in particular applications of a test.

The purpose of this meta-analytic research (Montero and León, 2007) was to carry out an RG study of the HAM-D scale in order to accomplish the following objectives: a) to estimate the average reliability obtained in a representative sample of studies that have applied the HAM-D in psychological research; b) to test whether the reliability of the HAM-D scores can be generalized across different applications of the scale or if, in contrast, reliability estimates show a variability that cannot be explained only by sampling error; c) to examine how reliability estimates are influenced by the number of items in the scale and by the variability in the sample scores, as psychometric theory predicts; d) to explore relationships between other sample and instrument features of the studies and score reliability; and e) to compare our results with those of other RG studies published on three different depression scales.

Method

Literature search

To identify studies for the RG study, a literature search in the electronic database PsycINFO was carried out to find empirical studies that applied some version of the HAM-D scale. The following key words were combined in the electronic search for the period from 1978 to 2004: 'Hamilton rating scale depression' with 'reliability', 'internal consistency', or 'factor analysis'.

Inclusion and exclusion criteria

To be included in the meta-analysis, the studies had to meet two selection criteria: a) be an empirical study that applied some version of the HAM-D scale to (at least) one subject sample, and b) report sample specific reliability coefficients. The search gave

5,668 references and the reading of the abstracts led to a selection of 206 references that had applied the HAM-D to a subject sample. The remaining references were deleted because they were not empirical studies, but theoretical papers about depression and/or other related disorders, or empirical studies that supposedly did not report reliability estimates. Once the 206 papers were obtained, their reading gave 95 papers (46.1%) which reported some reliability coefficient empirically obtained with the study samples. In particular, 75 articles (78.9%) applied an English version of the scale, whereas the 20 remaining articles (21.1%) applied a translated version (Spanish, Turkish, and Korean). In any case, the 95 articles were written in English, with the exception of one article that was written in Spanish.

To maintain the individual reliability estimates in our RG study, the unit of analysis was the subject sample, not the article. This is because in 42 of the 95 articles reliability estimates were reported for different subject samples. On the other hand, when the study implied pretest and posttest measures, only reliability coefficients obtained at the pretest were included, in order to avoid dependence on the data.

A source of heterogeneity among the articles was the type of reliability coefficient reported. The reliability coefficient most frequently used was the coefficient alpha, with 43 estimates (45.3%). The use of other reliability coefficients (inter-coder, within-class, Loevinger, test-retest, etc.) was very scarce. Different reliability coefficients are based on different assumptions and, if they are included in the same meta-analysis, interpreting the results can be troublesome (Dimitrov, 2002; Sawilowsky, 2000). Only the studies with alpha coefficients were included in the RG study, in order not to mix reliability coefficients proceeding from different definitions of reliability (internal consistency, test-retest, parallel forms, concordance). Moreover, we also excluded 8 of the 43 samples that reported alpha coefficients, because the HAM-D scales applied in those cases were special versions that included additional items measuring disorders other than depression. Therefore, our RG study integrated 35 independent samples obtained from 23 separate sources, with a total sample of 7,395 subjects.

Coding of characteristics

According to psychometric theory, it is expected that score reliability will be affected by such variables as the test length and the standard deviation of the test scores in the group. To examine possible relationships between the reliability estimates and the study features, moderator variables related to the instrument and the subject samples were coded:

1. Test length: 6, 17, and 21 items.
2. Score *SD*: Standard deviation of the test scores in the sample.
3. Language: Language of the HAM-D scale version (1, English; 0, other).
4. Mean age: Mean age of the subject sample (in years).
5. Age *SD*: Standard deviation of the age in the sample (in years).
6. Percentage male: Percentage of men in the sample.
7. Population type: 1, clinic; 0, other (normal population or normal population with any physical disease).

8. Disorder: Main disorder in the sample (1, depression; 0, other).
9. Diagnostic: Diagnostic instrument used to select the sample subjects (1, any version of the DSM; 0, other).
10. Use: Use of the scale (1, to measure severity of symptoms; 0, other).
11. Method: Type of empirical study (1, about psychometric properties; 0, other).
12. Hamilton: 1, the study was focused on the psychometric properties of the HAM-D scale; 0, focused on other depression scales.

A code book with detailed descriptions of how the moderator characteristics of the studies were coded can be requested to the authors. In the Appendix 1 a table with the complete database is presented. In order to examine the reliability of the coding process a random sample of the 23 studies (20%) was coded by two independent coders, showing an acceptable inter-rater agreement (mean agreement: .82). Inconsistencies between the coders were solved by discussion.

Statistical analyses

To carry out the RG study, a coefficient alpha was obtained from every sample. In order to normalize the reliability estimates, the square root of each reliability coefficient (that is, the reliability index) was translated into the Fisher's Z (Feldt and Charter, 2006; Sawilowski, 2000; Thompson and Vacha-Haase, 2000). We applied meta-analytic procedures which weight each reliability estimate according to its precision. This implies giving more weight to reliability estimates obtained from studies with a large sample size in comparison with studies with smaller ones. A fixed-effects model was assumed to obtain average reliability estimates and to test the influence of study characteristics on the variability of the reliability coefficients across different applications of the HAM-D scale. Applying a fixed-effects model implies weighting every reliability estimate according to its inverse-variance, where the variance for each reliability estimate refers to the variability due to sampling error (Hedges, 1994; Mason *et al*., 2007). The reason for applying a fixed-effects model and not a random-effects model was because the sample of studies included in our RG review was not very large and, as a consequence, we decided to generalize our results to only studies with similar characteristics to those included in our review.

Together with a weighted average reliability coefficient and a 95 per cent confidence interval, the Q test was applied to assess whether the reliability estimates of the studies were homogeneous around its mean or if, on the contrary, the variability of the reliability estimates cannot be due to sampling error alone. To complement the result of the Q test the I^2 index was also calculated (Higgins and Thompson, 2002). The I^2 index can be interpreted as the percentage of the total variability in a set of reliability estimates caused by true heterogeneity, that is, to between-studies variability. For example, when $I^2 = 50$ it means that half of the total variability among reliability estimates is caused not by sampling error, but by true heterogeneity between the studies.

To explore the effect of study characteristics on the reliability estimates variability, we applied ANOVAs (for the categorical variables) and regression models (for the continuous variables). Finally, by means of weighted multiple regression a tentative explanatory model was proposed that included the most relevant study characteristics for predicting the score reliability.

Results

Descriptive characteristics of the studies

Focusing on the 35 samples that reported alpha coefficients, 34 (97.1%) were published in peer-review journals, with the remaining sample being reported in a book chapter. In most of the cases the main researcher was a psychiatrist (88.6%) and the HAM-D scale was applied as a clinical interview (65.7%). The HAM-D scale version most frequently used was that of 17 items (71.4%). The mean standard deviation of the test scores was 5.82 (*SD* = 2.19). Most of the test applications were with the original format in English (80%), whereas 7 studies used adaptations to other languages (Spanish, Turkish, and Korean). The sample sizes of the studies were very heterogeneous, with a mean of 211 subjects (*SD* = 213.8). The mean age of the subject samples was 45.7 years (*SD* = 12.4 years), although 6 studies did not report this information. The mean standard deviation of the age in the samples was 10.8 years (*SD* = 3.3 years). All the samples were composed of men and women, with the exception of one study which only included men, while in 6 samples this information was not reported. In total, the mean percentage of men in the samples was 38.8% (*SD* = 17.1%). Most of the test applications included in our RG study were for samples selected from populations with some psychological disorder (25 samples, 71.4%), with depression being the most frequent main disorder (22 samples, 62.9%). The most used diagnostic criteria was the DSM in any of its versions (24 samples, 68.6%), although 7 studies did not report this data. In 11 samples (31.4%) the HAM-D scale was used to assess the seriousness of the symptoms and in 13 samples (37.2%) this information was not available. With respect to the purpose of the studies, in 21 cases (60%) the objective was to assess psychometric properties of the HAM-D scale or of another test, whereas in the 14 remaining samples (40%) the purpose was more substantive. Finally, in 15 samples (42.9%) the focus of the study was the HAM-D scale itself, whereas in the 20 remaining samples (57.1%) the objective of the study was not directly related to this scale. A table with the full data set of the RG study can be consulted in the Appendix 1.

Average reliability estimates of the HAM-D Scale

Reliability estimates, in terms of coefficient alpha, ranged from a low of .41 to a high of .89 (*SD* = .14). Table 1 presents the average reliability estimates for the total sample and for the three versions of the HAM-D scale. Applying Fisher's *r*-to-*Z* transformation on reliability indices and weighting them according to their inverse-variance, the average reliability estimate, in terms of coefficient alpha, was 79, with 95% confidence limits of .78 and .79. Therefore, we can consider that the applications of the HAM-D scale, in general, offer an internal consistency over the critical cut off point of 70 usually accepted as the minimum advisable reliability (Nunnally and Bernstein, 1994). But the *Q* test led to rejecting the homogeneity hypothesis of the reliability estimates around its mean ($Q_{(34)} = 757.11$; $p < .001$), and the I^2 index revealed that 95.5% of the variability was due to true heterogeneity among reliability estimates.

TABLE 1. Average reliability estimates as a function of the test length.

			95% C.I.			
Test length	*K*	*Mean*	*Ll*	*Lu*	*Q*	I^2
All studies	35	.79	.78	.79	757.11**	95.5
All equated at 17 items	35	.80	.79	.81	549.90**	93.8
6 items	6	.51	.45	.55	20.36**	75.4
17 items	25	.81	.80	.81	496.14**	95.2
21 items	4	.82	.80	.84	6.98	56.5
	$Q_{B\,2}$ = 233.71**; $Q_{W\,32}$ = 523.39**; ω^2 = 0.26					

Notes. *k*: Number of reliability estimates; Mean: Weighted average reliability estimate in terms of coefficient a; *Ll* and *Lu*: Lower and upper confidence limits at 95% confidence level around the mean reliability; *Q*: Heterogeneity statistic with k – 1 degrees of freedom; ** $p < .01$; I^2: I squared index; Q_B: Q statistic for testing the influence of the test length (with three categories: 6, 17, and 21 items) on the score reliability estimates; Q_W: Global within-category heterogeneity statistic; w^2: Variance proportion explained by the test length.

In an attempt to homogeneize the reliability estimates, the Spearman-Brown correction was applied to the alpha coefficients obtained with the 6 and 21 item versions to equate them to the 17 item version. Only a very slight increase in the average reliability estimate of 80 was obtained, with confidence limits of .79 and .81 (see Table 1). Although the heterogeneity among the reliability estimates decreased, there remained a high variability to be explained ($Q_{(34)}$ = 549.90; $p < .001$; I^2 = 93.8).

The next analysis consisted in calculating separate average reliability estimates for the varying number of items constituting the different HAM-D versions. As psychometric theory predicts, score reliability increases with the test length. In particular, the average reliability estimates (and confidence limits) obtained for 6, 17, and 21 item versions were, respectively, .51 (.45-55), .81 (.80-.81), and .82 (.80-.84). Only the 6 item version obtained an inadmissibly low reliability estimate (see Table 1). The differences between the three average reliability estimates were statistically significant and explained 26% of the variability ($Q_{B(2)}$ = 233.71; $p < .001$; w^2 = .26), although there remained variability to be explained ($Q_{W(32)}$ = 523.39; $p < .001$). In fact, the heterogeneity tests for 6 and 17 item versions were statistically significant and, although the *Q* test for the 21 item version did not reach statistical significance, its I^2 index was of medium magnitude (56.5%). Therefore, the HAM-D scale exhibits a reliability that depends on the particular applications and, as a consequence, it is not appropriate to generalize the reliability of the HAM-D scale to different contexts.

Relating study characteristics with reliability estimates

In addition to the number of items, other characteristics of the studies were analyzed to explain the high variability found among the reliability estimates. Tables 2 and 3 present the results obtained in the ANOVAs and simple regression analyses for the categorical and continuous moderator variables, respectively. As expected from the psychometric theory, the variability of the test scores (Score *SD*) affected reliability estimates positively (see Table 3), showing the highest explained-variance proportion of all of the moderator variables here tested ($Q_{R(1)}$ = 321.67; $p < .001$; R^2_{adj} = .40). So, the higher the score variability the larger the reliability estimate.

TABLE 2. ANOVAs (by weighted least squares) and weighted average reliability estimates for the categorical moderator variables.

			95% C.I.				
Moderator variable	*K*	*Mean*	*Ll*	*Lu*	Q_W	I^2	ω^2
Population type					$Q_B = 5.20$		.00
1: clinic	25	.79	.78	.80	296.06**	91.9	
0: other	10	.77	.77	.79	455.85**	98.0	
Disorder					$Q_B = 260.92$**		.31
1: depression	22	.82	.81	.83	377.42**	94.4	
0: other	13	.60	.56	.63	118.77**	89.9	
Diagnostic					$Q_B = 23.98$**		.08
1: DSM	24	.81	.80	.82	125.59**	82.2	
0: other	4	.71	.66	.75	39.58**	92.4	
Language					$Q_B = 3.17$		.00
1: English	28	.79	.78	.80	721.64**	96.2	
0: other	7	.77	.75	.79	32.39**	81.4	
Use					$Q_B = 0.47$		.00
1: symptom severity	11	.82	.81	.83	295.18**	96.6	
0: other	11	.8	.81	.84	51.14**	80.4	
Method					$Q_B = 142.88$**		.16
1: psychometric	21	.82	.81	.82	365.71**	94.5	
0: other	14	.68	.66	.71	248.52**	94.8	
Hamilton					$Q_B = 74.06$**		.06
1: yes	15	.82	.81	.83	220.21**	93.6	
0: no	20	.75	.73	.76	462.83**	95.9	

Notes. *k*: Number of reliability estimates; Mean: Weighted average reliability estimate in terms of coefficient a; *Ll* and *Lu*: Lower and upper confidence limits at 95% confidence level around the mean reliability; Q_W: Within-category heterogeneity statistic with k – 1 degrees of freedom; * $p < .05$. **p < .01; I^2: I squared index; Q_B: Q statistic for testing the influence of the moderator variables on the score reliability estimates; w^2: Variance proportion explained by the moderator variables.

TABLE 3. Simple regression models (by weighted least squares) for the continuous moderator variables.

Moderator variable	*K*	*b*	Q_R	Q_E	R^2_{adj}
Score SD	35	.08	321.67**	435.43**	.40
Mean age	29	-.01	73.85**	519.09**	.09
Age SD	15	.01	7.46**	151.09**	.00
Percent male	29	.001	2.88	620.89**	.00

Notes. *SD*: Standard Deviation; *k*: Number of studies; *b*: Unstandardized regression coefficient; Q_R: Weighted regression sum of squares with 1 degree of freedom to assess the model fitting; Q_E: Weighted error sum of squares with k - 2 degrees of freedom to assess the model misspecification; ** $p < .01$; R^2_{adj}: Variance proportion explained by the moderator variables.

The next study feature that showed a high explained-variance proportion was whether the main disorder studied in the sample was depression or another (see Table 2). In particular, the studies whose samples were composed mainly of subjects with any type of depression obtained a higher average reliability coefficient ($M = .82$) than those

composed by subjects with other disorders (M = .60) ($Q_{B(1)}$ = 260.92; p < .001; ω^2 = .31). In fact, the samples composed of individuals with other disorders showed a mean reliability coefficient and confidence limits (.56 and .63) below the 70 value, which is the one typically assumed as the minimum advisable reliability coefficient.

Another moderator variable that achieved a strong relationship with the reliability estimates was whether the objective of the study was to examine psychometric properties of the test or something else ($Q_{B(1)}$ = 142.88; p < .001; ω^2 = 16) (see Table 2). In this case, a higher average reliability coefficient was obtained when the purpose of the study was psychometric (M = .82; confidence limits: 81 and 82) than when the objective was substantive, mainly clinical applications of the HAM-D scale (M = 68; confidence limits: .66 and .71).

Other study characteristics also reached a statistically significant relationship (p < .05) with the reliability estimates, but their explained-variance proportions were so small (all of them under 10%) that they can be considered negligible. This was the case of such study characteristics as: a) the mean age of the individuals in the sample, which showed a negative relationship with the reliability coefficients (R^2_{adj} = .09); b) the diagnostic instrument applied in the study, with better reliability estimates obtained by the studies that applied some version of the DSM (M = .81) than those that used other diagnostic instruments (M = .71; ω^2 = .08), and c) the purpose of the study, with a higher average reliability coefficient for the studies that were focused on the properties of the HAM-D scale (M = 82) than those centered on other measurement instruments (M = 75; ω^2 = .06). Another two moderator variables that reached statistical significance but with a null explained-variance proportion were the standard deviation of the age in the samples and the population that the samples represented (clinical versus other). Finally, there were three moderator variables that showed no statistically significant relationship with the reliability coefficients and a null explained-variance proportion: the language of the HAM-D version applied, the percentage of men in the samples, and the use of the HAM-D (to assess symptom severity versus other uses).

Although most of the moderator variables tested here showed a statistically significant relationship with the reliability estimates, in all of the cases there also remained variance to be explained, as is evidenced by the results obtained with the misspecification tests, Q_W and Q_E for the ANOVAs and regression analyses, respectively (see Tables 2 and 3). Therefore, none of the moderator variables, by itself, was able to explain all of the variability in the reliability estimates.

A predictive model

So far the analyses presented here only assessed bivariate relationships between each moderator variable and reliability estimates found in the samples. Due to the collinearity among the study characteristics, it is possible that some of the statistical relationships commented above were spurious. Therefore, a tentative predictive model was proposed that included the most relevant moderator variables, on both a substantive and a statistical basis, to better explain the variability of the reliability coefficients found in the different applications of the HAM-D scale. However, the low number of samples included (only 35 reliability coefficients) limited the number of predictors that might be

introduced in the multiple regression model. Thus, the model proposed here only included the three most relevant moderator variables analyzed in our RG study: the number of items of the HAM-D version, the variability of the test scores, and the disorder studied in the samples (1: Depression; 0: other disorders).

TABLE 4. Results of the multiple regression analysis by weighted least squares.

Moderator variable	*b*	*Z*	*p*	ΔR^2
Test length	.02	5.45	< .0001	.03
Score *SD*	.05	8.93	< .0001	.10
Disorder (1: Yes; 0: No)	.24	7.16	< .0001	.06
$Q_{R(31)}$ = 426.23**; $Q_{E(31)}$ = 330.35**; R^2_{adj} = .52				
Predictive equation:				
Z' = .55 + .02 xTest length + .05xGroup *SD* + .24xDisorder				

Notes. *b*: Partial unstandardized regression coefficient; *z*: Partial z test for each moderator variable; *p*: Probability level for the z test; ΔR^2: Proportion of the variance in reliability estimates accounted for when adding the moderator variable, once the other two variables have already been included in the multiple regression model (i.e., ΔR^2 is the squared semi-partial correlation coefficient); Q_R: Weighted regression sum of squares to assess the model fitting; Q_E: Weighted error sum of squares to assess the model misspecification; R^2_{adj}: Variance proportion explained by the three moderator variables; ** $p < .01$; *Z'*: Predicted Fisher's Z by the regression model.

Table 4 presents the results of the multiple regression model, by weighted least squares, applied for the three moderator variables on the Fisher's *r*-to-*Z* transformation of reliability estimates. Each of the three moderator variables achieved a statistically significant relationship with the reliability estimates, once the influence of the remaining two predictors had been partialized and, as a consequence, the global model fitting was also statistically significant ($Q_{R(3)}$ = 426.23; $p < .001$), with an explained-variance proportion of 52.1%. In particular, the score standard deviation was the moderator variable in the model that explained the largest variance proportion, with an increase in R^2 of 10.7% (see Table 4). Therefore, these results showed that the number of items, the variability of the test scores, and the disorder studied in the samples are study features that strongly affect the reliability estimates obtained when the HAM-D scale is applied to a particular sample.

The predictive model shown in Table 4 could be used to predict reliability estimates of future HAM-D applications in a given context, as a function of the three moderator variables. Thus, for example, in a future HAM-D application with the 17-item version on a sample to study depression and with a standard deviation for the test scores of 5.82 (the mean score *SD* obtained in our RG study), the predicted Fisher's *Z* obtained with the predictive equation was: *Z*' = .55 + .02x17 + .05x5.82 + .24x1 = 1.45. In the metric of coefficient alpha and applying equation (5), this implies a predicted reliability estimate of *r*' = 80. However, if the main disorder studied in the sample was not depression but another, then we can expect a lower reliability, as *Z*' = .55 + .02x17 + .05x5.82 + .24x0 = 1.20, and the predicted coefficient alpha is *r*' = .69. However, predictions should be interpreted cautiously, because the multiple regression model also showed that there is still variance in the reliability estimates to be explained ($Q_{E(31)}$ = 330.35; $p < .001$) and,

as a consequence, the model is misspecified. This implies that there were other study characteristics not included in the model that might also affect the reliability coefficients of the HAM-D applications.

Discussion

The mean reliability and the sources for the variability of reliability estimates across a representative sample of studies in the HAM-D scores were examined by means of the reliability generalization approach. Of the 206 papers selected to be included in the meta-analysis, only 95 (46.1%) reported reliability estimates obtained for the samples in the studies. This implies that the remaining papers were not concerned about measurement reliability exhibited by the scores obtained with the application of the HAM-D scale.

In order not to mix reliability coefficients calculated from different conceptions of measurement reliability (internal consistency, test-retest, etc.), our RG study focused on 35 internal consistency reliability estimates (alpha coefficients) obtained from 23 papers. The mean coefficient alpha obtained across the 35 reliability estimates was .79 and was in the range of an acceptable reliability (Nunnally and Bernstein, 1994). However, the reliability estimates showed a high heterogeneity (95.5%) across studies that sampling error might not explain by itself. Therefore, the score reliability of the HAM-D scale cannot be generalized across their applications, and it is very plausible to think that different characteristics of the studies are influencing reliability estimates.

The first two variables whose influence on reliability estimates was tested were the number of items and the variability of test scores. The HAM-D version most frequently used was that of 17-items, with a mean reliability of .81, the 21-item version exhibiting a mean reliability only slightly higher than the previous one (.83). Only the 6-item version showed a mean reliability below the minimum advisable value (.51). The number of items showed a statistically significant, positive relationship with measurement reliability, due mainly to the low reliability exhibited by the 6-item version. Therefore, the HAM-D version that we recommend is that of 17 items, because adding four items produces a negligible reliability improvement. Together with the number of items in the HAM-D version, and as psychometric theory predicts, the score variability (defined as the standard deviation of the set of scores in the sample) was also positively related with reliability estimates.

Other study characteristics that reached a statistically significant relationship with reliability estimates and a proportion of explained variance over 10% were: a) the disorder studied (depression versus other disorders), with better reliability estimates when the study focused on depression, and b) the purpose of the study, with better reliability estimates achieved by studies of psychometric properties of the HAM-D scale in comparison with studies focused on clinical applications of the scale. Elsewhere, there were other variables that were statistically related to reliability estimates, but with a negligible proportion of explained variance. Therefore, this evidence shows a considerable number of sources of variability that can explain, in part, the heterogeneity found among reliability coefficients across studies, as well as the dependence that measurement reliability exhibits from specific applications of the instrument. Along the same line, a tentative model was proposed that enables the prediction of measurement

reliability in future applications of the scale, taking into account the three most relevant moderator variables here analyzed: the number of items of the HAM-D version, the standard deviation of the test scores in the sample, and whether the study focused on depression or on another disorder.

Our results confirmed that, at least with respect to the HAM-D scale, reliability is a property of the test scores and not of the test itself and, as a consequence, it is more suitable to speak of HAM-D «score reliability» than of HAM-D «reliability». Researchers and applicants of measurement instruments should not fall into the mistake of inducing reliability from previous applications of the instrument (Vacha-Haase, Kogan, and Thompson, 2000). To assess reliability it is important to take into account context, population studied, group homogeneity, and many other characteristics that can affect score reliability. We encourage researchers to report score reliability estimates for the data in hand, and clinicians to take into account the application context in evaluating the measurement reliability of an instrument when they are assessing individuals.

Another purpose of our study was to compare the typical measurement reliability exhibited by the HAM-D scores with that obtained by the three other depression scales that have been subjected to an RG study so far: The Beck Depression Inventory, BDI (Yin and Fan, 2000), the Geriatric Depression Scale (GDS) (Kieffer and Reese, 2002), and the Center for Epidemiologic Studies-Depression Scale (CES-D), applied to care providers (O'Rourke, 2004). Focusing on internal consistency reliability estimates, the three RG studies showed average reliability estimates very similar to that obtained for the HAM-D (M = .79; SD = .14; k = 35); .84 for BDI (SD = .07; k = .142); .80 for GDS (SD = .14; k = 215), and .88 for CES-D scale (SD = .05; k = 130). On the other hand, in all of the RG studies the number of items and variability of test scores were positively related to reliability estimates and similar conclusions were reached about the need to estimate measurement reliability for each application of the scale.

The RG study presented here has some limitations. First, the search strategy to find papers that applied the HAM-D scale only included the electronic database PsycINFO. Although this circumstance is shared with the RG studies for the BDI (Yin and Fan, 2000) and for the GDS (Kieffer and Reese, 2002), our results could be confirmed by extending the search for studies to other databases. On the other hand, the number of reliability estimates included in our RG study was small (k = 35) in comparison with those of the other three RG studies on depression scales. The small number of reliability estimates limited the number of moderator variables included in the multiple regression model and this is probably the reason for obtaining a misspecified model. Extending the search for studies could lead to fitting a more complete regression model.

Finally, it is important to note that all our meta-analytic calculations were made by transforming reliability indices into Fisher's Z. There is some debate about whether to use Fisher's Z transformation on reliability coefficients (Feldt and Charter, 2006; Henson and Thompson, 2002; Leach, Henson, Odom, and Cagle, 2006; Mason *et al*., 2007; Sawilowsky, 2000; Thompson, 2003; Thompson and Vacha-Haase, 2000). In order to test if our results were affected by using Fisher's Z, we carried out a sensitivity analysis that implied repeating the meta-analytic calculations using alpha coefficients without translating them into Fisher's Z, and weighting them by sample size. The results were very similar

to those obtained with Fisher's *Z* transformation: a mean reliability coefficient of 76, high heterogeneity (I^2 = 80.1%), and the same study characteristics showing a statistically significant relationship with reliability estimates, mainly the number of items in the test version, variability of test scores, and type of disorder studied in the sample. Therefore, our results were robust to changes in the meta-analytic technique.

References[2]

*Addington, D., Addington, J., and Atkinson, M. (1996). A psychometric comparison of the Calgary Depression Scale for Schizophrenia and the Hamilton Depression Rating Scale. *Schizophrenia Research*, *19*, 205-212.

*Addington, D., Addington, J., Maticka-Tyndale, E., and Joyce, J. (1992). Reliability and validity of a depression rating scale for schizophrenics. *Schizophrenia Research*, *6*, 201-208.

*Addington, D., Addington, J., and Schissel, B. (1990). A depression rating scale for schizophrenics. *Schizophrenia Research*, *3*, 247-251.

*Akdemir, A., Türkçapar, M.H., Örsel, S.D., Demirergi, N., Dag, I., and Özbay, M.H. (2001). Reliability ad validity of the Turkish Version of the Hamilton Depression Rating Scale. *Comprehensive Psychiatry*, *42*, 161-165.

Bech, P. and Allerup, P. (1981). The Hamilton Depression Scale: Evaluation and selectivity using logistic models. *Acta Psychiatrica Scandinavica*, *63*, 290-299.

Bech, P., Allerup, P., Reisby, N., and Gram, L.F. (1984). Assessment of symptom change from improvement curves on the Hamilton Depression Scale in trials with antidepressants. *Psychopharmacology*, *84*, 276-281.

*Bech, P., Lunde, M., and Undén, M. (2002). Social Adaptation Self-evaluation Scale (SASS): Psychometric analysis as outcome measure in the treatment of patients with major depression in the remission phase. *International Journal of Psychiatry in Clinical Practice*, *6*, 141-146.

*Bech, P., Stage, K.B., Nair, N.P.V., Larsen, J.K., Kragh-Sørensen, P., and Gjerris, A. (1997). The Major Depression Rating Scale (MDS). Inter-rater reliability and validity across different settings in randomized moclobemide trials. *Journal of Affective Disorders*, *42*, 39-48.

*Bent-Hamsen, J., Lunde, M., Klysner, R., Andersen, M., Tanghøj, P., Solstad, K., and Bech, P. (2003). The validity of the Depression Rating Scale in discriminating between citalopram and placebo in depression recurrence in the maintenance therapy of elderly unipolar patients with major depression. *Pharmacopsychiatry*, *36*, 313-316.

Bentz, B.G. and Hall, J.R. (2008). Assessment of depression in a geriatric inpatient cohort: A comparison of the BDI and GDS. *International Journal of Clinical and Health Psychology*, *8*, 93-104.

Beretvas, S.N. and Pastor, D.A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement, 63*, 75-95.

*Bobes, J., Bulbena, A., Luque, A., Dal-Ré, R., Ballesteros, J., Ibarra, N., and Grupo de Validación en Español de Escalas Psicométricas (GVEEP) (2003). Evaluación psicométrica comparativa de las versiones en español de 6, 17, 21 ítems de la Escala de valoración de Hamilton para la evaluación de la depresión. *Medicina Clínica*, *120*, 693-700.

Botella, J. and Gambara, H. (2006). Doing and reporting a meta-analysis. *International Journal of Clinical and Health Psychology, 6*, 425-440.

[2] References marked with an asterisk indicate studies included in the meta-analysis.

Cabañero-Martínez, M.J., Cabrero-García, J., Richart-Martínez, M., Muñoz-Mendoza, C.L., and Reig-Ferrer, A. (2007). Revisión estructurada de las escalas de depresión en personas mayores. *International Journal of Clinical and Health Psychology, 7*, 823-846.

*Cole, J.C., Motivala, S.J., Dang, J., Lucko, A., Lang, N., Levin, M.J., Oxman, M.N., and Irwin, M.R. (2004). Structural validation of the Hamilton Depression Rating Scale. *Journal of Psychopathology and Behavioral Assessment*, *26*, 241-254.

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

Feldt, L.S. and Charter, R.A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement, 66*, 215-227.

*Fuglum, E., Rosenberg, C., Damsbo, N., Stage, K., Lauritzen, L., Bech, P., and Danish University Antidepressant Group (1996). Screening and treating depressed patients. A comparison of two controlled citalopram trials across treatment settings: Hospitalized patients vs. patients treated by their family doctors. *Acta Psychiatrica Scandinavica*, *94*, 18-25.

Gelenberg, A.J., Wojcik, J.D., Falk, W.E., Baldessarini, R.J., Zeisel, S.H., Schoenfeld, D., and Mok, G.S. (1990). Tyrosine for depression: A double-blind trial. *Journal of Affective Disorders*, *19*, 125-132.

Gibbons R.D., Clark, D.C., and Kupfer, D.J. (1993). Exactly what does the Hamilton Depression Rating Scale measure. *Journal of Psychiatry Research, 27*, 259-273.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry*, *23*, 56-62.

*Hammond, M. (1998). Rating depression severity in the elderly physically ill patient: Reliability and factor structure of the Hamilton and the Montgomery-Asberg depression rating scales. *International Journal of Geriatric Psychiatry*, *13*, 257-261.

Hedges, L.V. (1994). Fixed effects models. In H. Cooper and L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285-299). New York: Russell Sage Foundation.

Henson, R.K. and Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting «reliability generalization» studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-127.

Higgins, J.P.T. and Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558.

Kieffer, K.M. and Reese, R.J. (2002). A reliability generalization study of the Geriatric Depression Scale. *Educational and Psychological Measurement, 62*, 969-994.

*Kobak, K.A. and Reynolds, W.M. (2000). The Hamilton Depression Inventory. In M. E. Maruish (Ed.), *Handbook of psychological assessment in primary care settings* (pp. 423-461). Mahwah, NJ: Lawrence Erlbaum Associates.

Leach, L.F., Henson, R.K., Odom, L.R., and Cagle, L.S. (2006). A reliability generalization study of the Self-Description Questionnaire. *Educational and Psychological Measurement, 66*, 285-304.

* Leidy, N.K., Palmer, C., Murray, M., Robb, J., and Revicki, D.A. (1998). Health-related quality of life assessment in euthymic and depressed patients with bipolar disorder. *Journal of Affective Disorders*, *48*, 207-214.

*Maier, W., Philipp, M., Heuser, I., Schlegel, S., Buller, R., and Wetzel, H. (1988). Improving depression severity assessment –I. Reliability, internal validity and sensitivity to change of three observer depression scales. *Journal of Psychiatry Research*, *22*, 3-12.

Mason, C., Allam, R., and Brannick, M.T. (2007). How to meta-analyze coefficient-of-stability estimates: Some recommendations based on Monte Carlo studies. *Educational and Psychological Measurement, 67*, 765-783.

Montero, I. and León, O.G. (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology, 7*, 847-862.

Nunnally, J. and Bernstein, I. (1994). *Psychometric theory*. New York: McGraw-Hill.

O´Rourke, N. (2004). Reliability generalization of responses by care providers to the Center for Epidemiologic Studies-Depression Scale. *Educational and Psychological Measurement, 64*, 973-990.

Paykel, E.S. (1985). The Clinical Interview for Depression: Development, reliability and validity. *Journal of Affective Disorders*, *9*, 85-96.

*Potts, M.K., Daniels, M., Burnam, A., and Wells, K.B. (1990). A structured interview version of the Hamilton Depression Rating Scale: Evidence of reliability and versatility of administration. *Journal of Psychiatry Research*, *24*, 335-350.

*Ramos, J.A. and Cordero, A. (1988). A new validation of the Hamilton Rating Scale for depression. *Journal of Psychiatry Research*, *22*, 21-28.

*Rapp, S.R., Smith, S.S., and Britt, M. (1990). Identifying comorbid depression in elderly medical patients: Use of extracted Hamilton Depression Rating Scale. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*, 243-247.

*Reynolds, W.M. and Mazza, J.J. (1998). Reliability and validity of the Reynolds Adolescent Depression Scale with young adolescents. *Journal of School Psychology*, *36*, 295-312.

*Riskind, J.H., Beck, A.T., Brown, G., and Steer, R.A. (1987). Taking the measure of anxiety and depression. *The Journal of Nervous and Mental Disease*. *175*, 474-479.

Rodriguez, M.C. and Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306-322.

Rosenthal, S.H. and Klerman, G.L. (1966). Endogenous features of depression in women. *Canadian Psychiatric Association Journal*, *11*, 11-16.

*Rush, A.J., Giles, D.E., Schlesser, M.A., Fulton, C.L., Weissenburger, J., and Burns, C. (1986). The Inventory for Depressive Symptomatology (IDS): Preliminary findings. *Psychiatry Research*, *18*, 65-87.

*Rush, A.J., Triverdi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstein, S., Manber, R., Thase, M.E., Kocsis, J.H., and Keller, M.B. (2003). The 16-item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report(QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, *54*, 573-583.

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

*Stage, K.B., Middelboe, T., and Pisinger, C. (2003). Measurement of depression in patients with chronic obstructive pulmonary disease (COPD). *Nordic Journal of Psychiatry, 57*, 297-301.

Terman, M. (1988). On the question of mechanism in phototherapy for seasonal affective disorder: Consideration of clinical efficacy and epidemiology. *Journal of Biological Rhytms, 3*, 155-172.

Thase, M.E., Frank, E., Malinger, A.G., Hamer, T., and Kupfer, D.J. (1992). Treatment of imipramina-resistant recurrent depression, III: Efficacy of monoamine oxidase inhibitors. *Journal of Clinical Psychiatry*, *53*, 5-11.

Thompson, B. (Ed.) (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Thompson, B. and Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195.

*Thunedborg, K., Black, C.H., and Bech, P. (1995). Beyond the Hamilton Depression scores in long-term treatment of manic-melancholic patients: Prediction of recurrence of depression by quality of life measurements. *Psychotherapy and Psychosomatics*, *64*, 131-140.

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

Vacha-Haase, T., Kogan, L.R., and Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals. *Educational and Psychological Measurement, 60*, 509-522.

Williams, J.B.W. (1988). A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry*, *45*, 742-747.

Williams, J.B.W., Link, M., Rosenthal, N.E., Amira, L., and Terman, M. (2000). *Structured Interview Guide for the Hamilton Depression Rating Scale-Seasonal Affective Disorder Version* (SIGH-SAD). New York: New York State Psychiatric Institute.

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

APPENDIX 1. Relation of data for each study included in the RG study (a is the estimated alpha coefficient).

Study	*α*	*Length*	*Score SD*	*Disorder*	*Method*	*Hamilton*	*Mean age*	*Age SD*	*Language*	*% male*	*Population*	*DSM*	*Scale use*
Addington *et al.* (1996)	.73	17	5.46	0	0	0	--	--	1	59.80	1	1	0
Addington *et al.* (1992a)	.77	17	7.10	0	0	0	43.0	--	1	59.00	1	1	0
Addington *et al.* (1992b)	.66	17	4.38	0	0	0	43.0	--	1	59.00	1	1	0
Addington *et al.* (1992c)	.66	17	4.34	0	0	0	30.9	--	1	--	1	1	1
Addington *et al.* (1990)	.78	17	6.71	0	0	0	30.9	--	1	--	1	1	1
Akdemir *et al.* (2001)	.75	17	6.89	1	1	1	34.9	10.6	0	32.00	1	1	1
Bech *et al.* (2002)	.81	6	7.62	1	0	0	41.3	--	1	26.67	1	1	0
Bech-Hanson *et al.* (2003)	.64	21	3.70	1	0	0	41.3	--	1	26.67	1	1	0
Bech *et al.* (1997a)	.48	17	2.84	0	0	0	53.0	--	1	34.20	0	--	--
Bech *et al.* (1997b)	.59	17	3.60	0	0	0	53.0	--	1	34.20	0	--	--
Bech *et al.* (1997c)	.43	21	2.95	0	0	0	53.0	--	1	34.20	0	--	--
Bech *et al.* (1997d)	.41	21	2.85	0	0	0	53.0	--	1	34.20	0	--	--
Bent et al. (2003)	.42	17	3.00	1	1	0	--	--	1	--	1	--	1
Bobes *et al.* (2003a)	.59	21	2.70	1	1	1	45.2	12.8	0	18.10	1	1	--
Bobes *et al.* (2003b)	.74	17	5.60	1	1	1	45.2	12.8	0	18.10	1	1	--
Bobes *et al.* (2003c)	.78	6	6.80	1	1	1	45.2	12.8	0	18.10	1	1	--
Cole *et al.* (2004)	.82	17	8.19	1	1	1	68.3	--	1	45.80	0	1	--
Fuglum *et al.* (1996)	.86	17	10.54	1	0	1	47.9	11.0	1	--	1	1	1
Hammond (1998a)	.46	17	2.73	1	1	1	--	--	1	26.00	1	0	1
Hammond (1998b)	.60	21	4.20	1	1	1	--	--	1	26.00	1	0	1
Kobak and Reynolds (2000)	.89	17	7.51	1	1	1	38.3	15.4	1	43.00	0	--	1
Leidy *et al.* (1998)	.86	17	6.32	0	1	0	--	--	1	--	0	1	0
Maier *et al.* (1988)	.88	6	10.52	1	1	1	--	--	1	--	1	1	1
Potts et al. (1990)	.82	17	8.19	1	1	1	42.2	--	1	25.80	1	1	1
Ramos and Cordero (1988)	.77	17	5.10	1	1	1	46.5	11.6	0	29.63	1	1	1
Rapp *et al.* (1990)	.83	17	6.01	1	0	1	69.3	--	1	100.00	0	0	0
Reynolds and Mazza (1998)	.85	17	5.74	0	1	0	12.5	0.9	1	40.40	0	--	--
Riskind *et al.* (1987)	.73	17	6.84	1	1	1	37.5	13.5	1	46.00	1	0	0
Rush *et al.* (2003a)	.83	6	7.42	1	1	0	43.6	10.7	1	35.60	1	--	--
Rush *et al.* (2003b)	.84	17	9.22	1	1	0	43.6	10.7	1	35.60	1	--	--
Rush *et al.* (1986a)	.80	17	7.10	1	1	0	38.2	12.1	1	46.00	1	--	--
Rush *et al.* (1986b)	.88	17	3.00	1	1	0	35.1	10.7	1	11.40	1	0	0
Stage *et al.* (2003)	.85	17	7.10	1	0	1	71.0	--	0	33.00	0	0	--
Thunedborg *et al.* (1995a)	.83	17	5.74	0	1	0	59.4	8.4	1	34.78	1	1	0
Thunedborg *et al.* (1995b)	.84	21	5.74	0	1	0	59.4	8.4	1	34.78	1	1	0

Note: a, b, c, and d are different samples in the same study.