A RELIABILITY GENERALIZATION STUDY OF THE MARLOWE-CROWNE SOCIAL DESIRABILITY SCALE

S. NATASHA BERETVAS, JASON L. MEYERS,
AND WALTER L. LEITE
University of Texas at Austin

A reliability generalization (RG) study was conducted for the Marlowe-Crowne Social Desirability Scale (MCSDS). The MCSDS is the most commonly used tool designed to assess social desirability bias (SDB). Several short forms, consisting of items from the original 33-item version, are in use by researchers investigating the potential for SDB in responses to other scales. These forms have been used to measure a wide array of populations. Using a mixed-effects model analysis, the predicted score reliability for male adolescents was .53 and the reliability for men's responses was lower than that for women's. Suggestions are made concerning the necessity for further psychometric evaluations of the MCSDS.

Response bias to items on psychological surveys has long been a focus of concern. According to Paulhus (1991), "A response bias is a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content" (p. 17). In particular, social desirability bias (SDB), which is defined as the inclination to respond in a way that will make the respondent look good, has been studied since the 1950s. Several scales have been constructed specifically to assess this tendency. Twelve such scales were already in use by 1984, and others have been developed since (e.g., Paulhus, 1984).

We would like to thank the reviewers and editor for their thoughtful and helpful suggestions to improve this article. Correspondence regarding this article should be addressed to S. Natasha Beretvas, SZB 504 Mail station D5800, Department of Educational Psychology, UT, Austin, TX 78712; e-mail: tasha.beretvas@mail.utexas.edu.

Educational and Psychological Measurement, Vol. 62 No. 4, August 2002 570-589

Social Desirability Bias Scales and Their Uses

SDB scales are primarily used to help provide evidence supporting the validity of responses to psychological surveys. The most common use of these scales involves the calculation of correlations between scores on the SDB scale and scores on the focal psychological instrument. In the majority of such analyses, the researcher hopes that such correlations are not substantial, thereby providing discriminant validity evidence for responses to the focal scale and therefore indicating that scores on the scale are not confounded by a respondent's tendency to respond in a socially desirable way. It should be noted that sometimes the definition of the construct being assessed by the focal scale might coincide somewhat with the construct purportedly being measured by the SDB scales, resulting in a stronger yet meaningful correlation between the two.

Another use for the SDB scales has involved factor analysis of scores on these scales concurrently with scores on the psychological scales of interest. Again, if a researcher believes that the factor explaining responses to SDB items differs from the construct underlying the other scales, then it would be hoped that discrete factors would be extracted. A final use of SDB scales for the validation of scores on focal surveys has involved deletion of responses made by participants with elevated scores on the SDB measures.

The most commonly used SDB scales include Edwards's (1957) and Crowne and Marlowe's (1960) different, though similarly named, Social Desirability Scales, as well as Paulhus's (1984) Balanced Inventory of Desirable Responding. Edward's scale was the preferred scale until the development of the Marlowe-Crowne Social Desirability Scale (MCSDS). Crowne and Marlowe's survey was created as a response to the possible confound with psychopathology detected for responses to Edwards's scale. Paulhus more recently created yet another social desirability scale, the Balanced Inventory of Desirable Responding, that he argued better assessed the two dimensions underlying performance on measures of SDB. Of these three, the MCSDS has continued to be the most frequently used survey for assessment of SDB.

The Marlowe-Crowne Social Desirability Scale

The authors of the MCSDS initially considered the construct being assessed by the scale to be "social desirability in terms of the need for subjects to respond in culturally sanctioned ways" (Crowne & Marlowe, 1960, p. 354). Crowne and Marlowe (1964) later modified the construct to become "need for social approval." The scale contains 33 forced-choice, true-false items concerning everyday behaviors. Eighteen of these items are considered attribution items where selection of the "true" response will award a respondent one point, thereby indicating a stronger tendency to respond in a socially desirable way than someone who had responded with "false." Two examples of attribution items include "Before voting I thoroughly investigate the qualifications of all the candidates" and "I have never intensely disliked anyone." These attribution items refer to socially approved but uncommon behaviors (Cramer, 2000). The remaining 15 items are considered denial items for which a "false" response is assigned one point. These items refer to socially disapproved but common behaviors. An example of such an item is "I like to gossip at times."

In addition to the three common uses of SDB scales, scores on the MCSDS have also been used to provide a measure of repressive coping style. Participants with scores above the median on the MCSDS but below the median for trait anxiety scores as measured by the Taylor Manifest Anxiety Scale are categorized as repressors (e.g., Kraft, 1999; Ringel, 2000).

Several shortened versions of the MCSDS resulted from factor analytic studies of the original 33-item form. The most commonly used short forms include three by Strahan and Gerbasi (1972)—one 20-item version and two differing 10-item forms. Reynolds (1982) developed three forms termed A, B, and C, each containing 11, 12, and 13 items, respectively. Ballard (1992) came up with four more forms, two with 11 and 12 items as well as two other forms, each consisting of a different combination of 13 items. All of these shortened versions contain items not modified from the original form.

A few researchers have selected their own subset of items from the original MCSDS. These studies have not included investigations of the construct validity of the scores on changed forms, although some of them have provided reliability estimates for scores on these additional forms. In addition, the MCSDS, as is found with many psychological measures, was initially evaluated using a sample of college students. It is still most typically used with adult participants. Yet, some researchers have used the scale with adolescent populations. Due to these somewhat indiscriminate uses of the MCSDS, it seems important to reevaluate the psychometric properties of scores on the MCSDS when used by different populations in the multitude of its forms that are currently in use. As a starting point for such analyses, this study will describe a reliability generalization investigation of scores on the various versions of the MCSDS with differing populations. See Vacha-Haase (1998) and other articles in this special issue of *Educational and Psychological Measurement* for discussions of reliability generalization. This investigation evaluated the internal consistency and the test-retest reliability of scores on MCSDS items using study descriptors to help explain potential variability in the reliability estimates.

Method

Data Source

Several databases were searched for the words *Marlowe* and *Crowne* or *Crown* (due to frequent misspellings) to obtain references for articles mentioning use of the MCSDS since 1960. The databases included PsycINFO, ERIC, Sociological Abstracts, and Social Sciences Abstracts. A total of 1,069 articles and dissertations that used the MCSDS were investigated. The vast majority of these articles did not provide reliability estimates specific to the samples to whom the MCSDS was given. Of the 1,069 articles, only 3.93% cited other authors' reported reliability estimates to substantiate the reliability of their participants' scores on the MCSDS. Most of these cited Crowne and Marlowe's (1960) original internal consistency and test-retest reliability estimates based on the responses of a small sample of college students. (The authors' internal consistency estimate, r = .88, was calculated using responses of 39 undergraduates, whereas the test-retest value, r = .89, was computed using scores of 31 students.) In all, only 93 studies (8.70%) reported sample-specific reliability estimates. These 93 studies contributed 149 Cronbach's alpha estimates, 3 Spearman-Brown corrected split-half reliability estimates, 9 KR-20 estimates, and 21 test-retest estimates, for a total of 182 coefficients.

The vast majority of the estimates were based on the English (n = 164) version, with 1 French, 1 Bengali, 1 Hindi, 9 German, 1 Norwegian, and 4 Dutch forms. The size of the samples whose scores were analyzed ranged from 16 to 11,315. Only 147 of the samples provided a gender frequency breakdown. Every study provided the number of items contained in the MCSDS version used for the coefficient. The number of items per form ranged from 5 to 33, providing a total of 23 different versions (including the original 33-item MCSDS).

Descriptors of the ethnicity of a sample's participants were only provided in 46 samples, so this variable was not included in further analyses. The average age of samples' participants was only provided in 83 samples. It was possible to categorize samples' participants into one of three age range categories, namely adolescents, adults younger than 50, and adults older than 50, for all 182 samples. Only one sample provided score reliability information for a sample older than 50; therefore, a dummy-coded age variable contrasting adolescents with adults was created. On this age variable, adolescents were coded with a zero and adults with a value of 1.

The majority of the samples (n = 97) consisted of college students. Only 6 of the samples contained clinical adults, whereas 57 of the samples' participants were nonclinical adults. Ten reliability estimates were based on scores of adolescents (younger than 18 years of age) and only 2 inmate samples contributed reliability estimates. Ten samples consisted of vague mixtures, 9 of which referred to a mix of “students and adults” and 1 that combined as one group the scores of normal adults along with men who physically abused their partners and men who had committed incest. These last 10 samples were deleted from further analysis. Unfortunately, the low number of psychiatric samples prohibited the use of a variable contrasting patients with nonpatients. For the test-retest estimates, the length of time in between testing occasions was coded on a scale ranging from 0 (*up to 1 week*), 1 (*1 to 2 weeks*), 2 (*2 weeks to 1 month*), to 3 (*longer than 1 month*).

Analysis

Use of Fisher’s r*-to-*z *transformation*. Reliability generalization (RG) studies have traditionally involved using fixed-effects models. One of the assumptions underlying these models includes the normality of the distribution of the criterion variable, here the reliability estimate. The sampling distribution of correlation estimates (for estimates not based on large samples) has been found to be skewed (e.g., Field, 2001). This problem is further aggravated when the true correlation is large (Hedges & Olkin, 1985), as will typically be the case for score reliability. Due to this potential for nonnormality, we used Fisher’s r-to-z transformation equation:

$$z_r = \left(\frac{1}{2}\right)\ln\left(\frac{1+r}{1-r}\right). \tag{1}$$

The z_r can then be used as the criterion variable because its distribution tends to be more normal than that of r. In addition to normalizing the associated sampling distribution, this transformation also stabilizes the variance. The asymptotic known variance for z_r is $1/(n-3)$, where n is the sample size associated with the original correlation estimate. Once the analyses are complete, the coefficients can be transformed back into the original correlation coefficient metric using the retransformation formula:

$$r = \frac{\exp(2z_r)-1}{\exp(2z_r)+1}. \tag{2}$$

As noted by Thompson and Vacha-Haase (2000), score reliability coefficients provide estimates of “variance-accounted-for universe values” (p. 186), implying that internal consistency reliability estimates are really in a squared correlation form. The above-mentioned transformation applies to the distribution of unsquared correlation estimates. In this study, we followed

Thompson and Vacha-Haase's recommendations to use the square root of each reliability estimate as the correlation estimate, *r*, that was initially transformed in Equation 1. Once the analyses were run and the coefficients were obtained, they were converted into the correlation form using Equation 2, and then the values were squared to revert to the corresponding reliability estimate metric.

Use of mixed-effects modeling. The conventional fixed-effects model was not used with this study's data set due to violation of the assumption of homoscedasticity. Although statistical tests under the fixed-effects models have shown to be robust to violations of this assumption, meta-analyses tend to involve a far more substantial degree of heterogeneous variances than those typically investigated. In this study, the ratio of the smallest to the largest sample sizes was 1 to 707, providing a degree of heterogeneity far exceeding the typically acceptable ratio for which robustness has been defined.

The nestedness of samples within studies also supported the use of a mixed- over fixed-effects model for analysis of this MCSDS data set. As evidenced by the number of internal consistency reliability estimates almost doubling the number of studies, several studies (27.78%) contained several reliability estimates based on the scores of multiple samples (see Table 1). It was anticipated that there would be some inherent dependency between scores of samples gathered by the same researchers in a single study. This dependency should be modeled in such a way that the variability in reliability estimates can be partitioned into the component that might result within studies from that part resulting from variability between studies (see Beretvas & Pastor, in press).

The partitioning of variability at these additional clustering levels not only addresses more appropriately the violations of the assumptions of homogeneous variances and independence but also provides a better conceptual fit to the intent of RG studies. It is expected that score reliability estimates will vary within and between studies. The focus of interest in RG studies entails the explanation of this variability using sample and study characteristics. Mixed-effects modeling is designed specifically for this type of exploration.

The next section provides a general outline of the steps that were taken for the mixed-effects analysis of transformed reliability estimates (for a fuller description supporting the use of mixed-effects models for RG studies, see Beretvas and Pastor, in press). It should be noted that internal consistency reliability estimates (including Cronbach's alphas, KR-20 coefficients, and Spearman-Brown corrected split-half reliability coefficients) were analyzed separately from the set of test-retest coefficients. This was done because score consistency over time and the internal consistency of scores from items on a scale are different facets of reliability (Henson, 2001; Henson & Hwang, 2002).

Table 1
Frequencies and Percentages Describing the Number of Score Reliability Estimates per Study

	Internal Consistency Estimates		Test-Retest Estimates	
Number of Estimates per Study	Frequency	% of Studies	Frequency	% of Studies
1	52	72.22	11	84.62
2	10	13.89	1	7.69
3	3	4.17	0	0
4	4	5.56	0	0
7	1	1.39	0	0
8	1	1.39	1	7.69
11	1	1.39	0	0
Total	72	100	13	100

Mixed-Effects Models for RG Studies

In hierarchical linear modeling terms (Raudenbush & Bryk, 1985), Level 1 modeled the variability of a sample's transformed reliability estimate, $z_{r_{ij}}$, around its predicted true reliability, $\varsigma_{p_{ij}}$:

$$z_{r_{ij}} = \varsigma_{p_{ij}} + e_{ij}, \tag{3}$$

where *ij* provided the index for sample *i* falling within study *j*. The variability of the error terms, e_{ij}, is treated as “known” in meta-analytic studies.

The variability between reliability estimates within studies was modeled at Level 2 describing the variability of a sample's predicted reliability, $\varsigma_{p_{ij}}$, around its study's predicted reliability, β_{0j}, using Equation 4:

$$\varsigma_{p_{ij}} = \beta_{0j} + r_{0ij}, \tag{4}$$

where β_{0j} represented the expected value of the parameter for study *j* and r_{0ij} is the within-study error term. The more variability between reliabilities within studies, the larger will be the variance of these error terms. The statistical test of this variance evaluates the homogeneity of the transformed correlations. If the statistical test is not significant, this indicates that the variability between score reliabilities within studies is not systematic and appears to result from random variability. This test is equivalent to the meta-analytic *Q*-test statistic (Hedges & Olkin, 1985). Like the *Q*-test statistic, this test can lack the statistical power to identify heterogeneity when based on too few samples.

Last, at Level 3, variability between studies in the reliability estimates was modeled based on the between-studies equation:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{5}$$

with γ_{00} representing the mean value for the population transformed reliability coefficient across studies, and u_{0j} represents the sampling variability between studies. A similar test of the homogeneity of correlations can be conducted at this between-studies level. Again, the more variability there is in reliability estimates between studies, the larger will be the variance component representing the variability of the u_{0j}s.

If it is inferred from the results of the unconditional model, as described in Equations 3, 4, and 5, that there is a substantial amount of variability within studies, then descriptors of the samples can be used to help explain that variability. If a Level 2, within-studies descriptor, X_1, is added to the Level 2 equation (Equation 4), then it becomes

$$\zeta_{\rho ij} = \beta_{0j} + \beta_{1j}X_{1ij} + r_{0ij}, \tag{6}$$

where the intercept, β_{0j}, now represents the predicted value of the transformed reliability coefficient for study *j* given the study is assigned a zero on predictor X_1. The value of the coefficient for the X_1 variable can be tested to assess the strength of the relationship between the predictor and the score reliability. The variability in reliability estimates within studies can be tested to determine whether a substantial degree of variability remains even with predictor X_1 in the equation.

If it is found that there is a substantial amount of heterogeneity of correlations between studies, then predictors at Level 3 can also be added to the model to explain this variability. If a Level 3 predictor, Z_1, is added to the model to explain variability in the expected score reliability for a study, then Equation 5 becomes

$$\beta_{0j} = \gamma_{00} + \gamma_{1j}Z_{1j} + u_{0j}, \text{ and} \tag{7}$$

$$\beta_{1j} = \gamma_{10}. \tag{8}$$

The intercept, γ_{00}, now designates the predicted score reliability for a study associated with samples with values of zero on predictors X_1 and Z_1. Thus, it is very important that the Level 2 predictor is centered appropriately to ensure that a value of zero is meaningful. For example, if X_1 represents a dummy-coded gender variable with zero assigned for men and 1 for women, then at Level 3 the intercept would represent the predicted score reliability for men. Variability in men's estimates could be explained with the addition of Level 3 predictors, but the model would not provide information about the explanation of variability in women's estimates.

In Equation 8, it can be seen that there is no error term associated with the coefficient for the relationship between the Level 2 predictor and the criterion variable. This indicates that if such a relationship is detected, it will be assumed fixed across studies. The coefficient γ_{10} provides the predicted reliability coefficient for a study given a value of 1 on the Level 2 predictor X_1. Last, as with the Level 2 model, it is possible to determine whether the study level predictor, Z_1, sufficiently explains the variability between studies or whether a large amount of variability remains unexplained.

Choice of Level 2 and Level 3 Predictors

The choice of predictors that can be used in any meta-analytic study, including RG studies, is restricted by the participant and form descriptors offered in each study. Fortunately, some important descriptors were provided in those studies and samples that also contained score reliability estimates. It was expected that the reliability coefficients would be positively related to the number of items on the form used. Because it was found that differing forms were used to assess SDB for samples within a study, a variable representing the number of items was used as a Level 2 variable in an attempt to explain the variability found in reliability coefficients within studies. In addition, the average number of items on the MCSDS forms varied across studies, so this was used as a Level 3 variable to explain variability between studies.

Although there was little variability in the age range of the participants, with 10 samples' estimates based on adolescents and the remaining samples using adults, the size of 1 of the adolescent samples (n = 11,315) was large enough that age range was included as a Level 3 predictor. Last, it was hypothesized that there might be gender differences in reliability estimates so the proportion of men in a sample was used as a Level 2 variable, and the proportion of male participants across samples constituting each study provided another Level 3 predictor.

Fixed-Effects Model Analysis

As described in Table 1, only 27.78% of the studies contributed more than one sample estimate to the analysis, so it was hypothesized that there might not be a substantial amount of variability within studies. In addition, because the use of mixed-effects modeling for RG analyses is still relatively new, it was of interest to conduct a comparison of the results from a traditional multiple regression analysis with the results from the mixed-effects model. For this reason, the same predictors as used in the mixed-effects model were used in a fixed-effects model with known variances. As is commonly done in meta-analyses (Field, 2001), the formula $v_i = 1/(n_i - 3)$ was used to estimate the variance, v_i, for study i.

It was expected that the coefficients' standard errors estimated in the fixed-effects model would be smaller than those resulting from the mixed-effects model. This results from the mixed-effects estimates including an additional component of variance attributed to the higher clustering level (here the within- and between-studies variability). The negative bias (Kreft & de Leeuw, 1998) of the fixed-effects model (in the presence of heterogeneous correlations) can result in improved power for the fixed-effects model at the expense of an inflated Type I error rate. Despite this improved power, ignoring the nestedness of data has also been found to confound results. It is sometimes possible to detect significant relationships under a mixed-effects model that can be masked when using fixed-effects models (Osborne, 2000). This can result from confounding, for example, Level 2 effects with Level 1 effects.

Results

Before presenting the results, it should be emphasized that the mixed-effects model was conducted using only complete data provided by 72 studies associated with a total of 123 samples. As can be seen in Table 1, there was a small proportion of studies that consisted of multiple samples. It has been found that the more groups (here, studies) and the more units (here, samples) within those groups, the more stable will be the estimation of the variability at Levels 2 and 3 (Kreft & de Leeuw, 1998). Thus, in this study, the estimation of the random effects (in RG studies, typically the intercept and variances) should be interpreted with caution, although the estimation of the fixed effects in a mixed-effects model have been found to be relatively robust even for small numbers of Level 2 and 3 units (Newsom & Nishishiba, 2002).

Internal Consistency Score Reliability

Mixed-effects model: Unconditional model. The dependent variable used was the transformed square root of the reliability estimate substituted as the correlation, *r*, into Equation 1. A significant amount of variability in reliability estimates was found both between ($z = 3.69$, $p < .0001$) and within studies ($z = 3.78$, $p < .0001$). As would be expected, the variance component within studies, with a value of .01377, was less than (approximately half of) the variance component of .02633 found between studies. In the unconditional model, the intercept ($\gamma_{00} = 1.2643$, $SE = .0248$) represented the overall mean transformed reliability. Using Equation 2 to transform this value back to the original square root of the reliability metric and then squaring the correlation to obtain the reliability estimate, the predicted internal consistency reliability coefficient, across forms and participants, was .726.

Mixed-effects model: Final model including Level 2 and Level 3 predictors. The Level 2 and Level 3 predictors were entered simultaneously into the mixed-effects model. At Level 3, in addition to the age variable, two predictors summarizing two variables across samples within a study were used. One variable was the average number of items on the forms used in a study. The second variable was the overall proportion of men in the samples contained within a study. It should be noted that the two Level 2 predictors (number of items on the form and proportion of men in the sample) were group-mean centered. Because the model of interest (see Equations 3 and 6 through 8) constrains the Level 2 predictors' coefficients to be fixed at Level 3 with a random intercept term, and group means are used as predictors at Level 3, the group-mean centering model will be equivalent to using the original raw values of the predictors (Kreft & de Leeuw, 1998). However, the group-mean centering facilitates interpretation of the intercept term at Level 3. The results of the final model are contained in Table 2, including the unstandardized coefficient and the variance component estimates.

With the addition of the within-studies explanatory variable, the Level 2 variance component was reduced from .01377 to .005308, a reduction of 61.45% of the variability within studies. With the three Level 3 predictors, the variance component between studies was reduced by 24.53%, from .02633 to .01987. There still remained a statistically significant amount of unexplained variability both within ($z = 2.75$, $p < .05$) and between studies ($z = 4.03$, $p < .0001$). At Level 1, the estimates of known variance associated with each reliability estimate ranged from .05263 down to .00038 with a mean of .0109.

The value of the coefficient for the Level 2 percentage of men in a sample predictor was small and not statistically significant, $t(49) = -1.26$, $p > .05$. The other Level 2 predictor, number of items on a test form, seemed to explain the bulk of the variability that was explained within studies. The positive value indicates that, as expected, the longer the test form, the larger the reliability coefficient. Although the value of this coefficient looks small, it should be remembered that it represents the change in the untransformed reliability coefficient for scores on test forms that differ by only one item. The Level 3 age predictor indicated that the reliability for scores on the MCSDS for adults is higher than for adolescents. The negative sign of the coefficient for the Level 3 proportion of male participants per study variable indicates that scores of women show higher internal consistency than do those of men.

Because the Level 2 number of items variable is group-mean centered, the intercept represents the predicted reliability for a form with the average number of items of the forms used within a study, controlling for the three Level 3 predictors. More specifically, the intercept represents the predicted score reliability for female adolescents responding to a form with the average number of items for a study. To interpret the values of the resulting coefficients, the coefficient for the intercept was added to 33 times the value for the Level 3 average number of items variable's coefficient. This was then used to provide predicted values for varying participants on the original 33-item MCSDS when that was the form used for all samples within a study.

Table 2
Mixed-Effects Final Model for the Meta-Analysis of Internal Consistency Reliability Coefficients

Fixed Effect	Coefficient	*SE*	*df*	*t*	*p* Value
Intercept, γ_{00}	.9364	.1014	68	9.24	< .0001
Age range, γ_{01}	.2994	.0852	68	3.52	.0008
Proportion of men, γ_{02}	–.2185	.0800	68	–2.74	.0078
Mean number of items, γ_{03}	.0061	.0022	68	2.70	.0087
Number of items, β_{01}	.0126	.0019	49	6.44	< .0001
Percentage male, β_{02}	–.0006	.0005	49	–1.26	.2123

Random Effect	Variance Component	*SE*	*z*	*p* Value
Between studies	.01987	.0049	4.03	< .0001
Within studies	.00531	.0019	2.75	.0003

Note. Predictors associated with γs are Level 3 predictors, and predictors associated with βs are used at Level 2.

The score reliability for women on the 33-item form was predicted to be .797, whereas for men it was .704. For the scores of female adolescents on the 33-item form, the internal consistency reliability was predicted to be .661, whereas it was expected to be .526 for male adolescents. The predicted score reliabilities for studies with an average number of items per form lower than 33 would be predicted to be even lower.

Fixed-effects model: Unconditional model. A fixed-effects model with known variances without predictors was evaluated to provide a comparison with the mixed-effects unconditional model results. The intercept estimate was 1.1713, $t(122) = 198.22$, $p < .0001$. The transformed value for this intercept indicated that the average score reliability was predicted to be .680. As expected, the standard error under the fixed-effects model was smaller ($SE = .0059$) than under the corresponding mixed-effects model ($SE = .0248$).

Fixed-effects model: Final model. A fixed-effects model with known variances, including the same set of five predictors used in the mixed-effects final model, was also evaluated. Table 3 contains the results for this analysis. The value of the intercept indicated that the predicted reliability for scores of female adolescents would be .677. Across the six coefficients, the fixed-effects standard error estimates were smaller than under the mixed-effects model. As expected, the coefficients that were statistically significant under the mixed-effects model were also significant under the fixed-effects model. In addition, the Level 2 percentage male variable was statistically significant under the fixed-effects model, $t(117) = –2.09$, $p < .05$, although not in the

Table 3
Fixed-Effects Final Model for the Meta-Analysis of Internal Consistency Reliability Coefficients

Fixed Effect	Coefficient	*SE*	*df*	*t*	*p* Value
Intercept, γ_{00}	1.1822	.0424	117	27.90	< .0001
Age range, γ_{01}	.3076	.0383	117	8.04	< .0001
Proportion of men, γ_{02}	–.5522	.0252	117	–21.88	< .0001
Mean number of items, γ_{03}	–.0005	.0007	117	–0.71	.4796
Number of items, β_{01}	.0124	.0011	117	11.39	< .0001
Percentage male, β_{02}	–.0008	.0004	117	–2.09	.0385

mixed-effects analysis. However, it should be emphasized that the value of the coefficient seemed small even when taking into consideration the metric underlying this variable. For example, for a study with two samples, if one sample consisted of 20% more men than the other sample from the same study, the predicted reliability for the first would be .669 and .685 for the second.

The only other reversal occurred for the Level 3 average items per form variable, which was not found to be statistically significant in the fixed-effects model, $t(117) = –.71$, $p > .05$. This happened despite the smaller standard error estimated under the fixed- versus the mixed-effects model.

Test-Retest Score Reliability

There were only 13 studies reporting the 21 test-retest reliability coefficients gathered in this study. Only two studies consisted of more than one sample that had contributed test-retest score reliability estimates (see Table 1). With this very small number of Level 2 and Level 3 units, the estimation of random effects would be a serious concern. The stability of the results for a mixed-effects model for this nested data would be questionable. In addition, it was determined that the addition of test-retest time as a Level 3 predictor would result in the deletion of 4 of the 13 studies and their 4 associated samples, further reducing the data set to 9 studies and 17 test-retest coefficients. The values of the test-retest coefficients ranged from a very low .38 (associated with a test-retest time interval of 2 to 4 weeks) to .86 (test-retest interval of more than 1 month).

Discussion

This study provided a meta-analysis of the transformed reliability coefficients based on responses to varying forms of the MCSDS by differing populations. It should be emphasized that, as has been found with the majority of the RG studies conducted, there is a pathetic lack in the reporting of sample-specific reliability estimates. For the MCSDS, only 8.7% of the studies using the scale actually reported a sample-specific reliability estimate. It is particularly important when untried versions of a scale are used, as is commonly done with the MCSDS, that psychometric evidence supporting both the construct validity and the reliability of scores on differing subsets of items be provided.

This RG study employed mixed-effects modeling to compensate for violations of the independence and homogeneous error variances assumptions made in more typically used fixed-effects models. Variability within and between studies was detected and only partly explained by available sample and study characteristic variables. A fixed-effects model was analyzed using known variances to provide an optimal approximation to the mixed-effects model. The results of the two models were compared. The standard error estimates in the fixed-effects model were consistently lower, resulting in the detection of one additional significant effect over those found in the mixed-effects model. However, this pattern was reversed for the average item number per study predictor found significant in the mixed- and not in the fixed-effects model. This probably resulted from using two similar variables representing the number of items on a form concurrently in the fixed-effects model. Although the group-mean centering of the Level 2 number of items variable and the average number of items per study variable at Level 3 makes sense in a mixed-effects model, the use of both variables in a fixed-effects model could be redundant. But this was done to provide as fair a comparison as possible across the two kinds of models.

From the analysis of the internal consistency estimates that were reported, an important caveat concerning the use of the MCSDS can be gleaned. It appears that the reliability of adolescents' scores on the scale is unacceptable. Perhaps some of the content of the items is not relevant to participants under 18 (such as, "Before voting I thoroughly investigate the qualifications of all the candidates" and "I never make a long trip without checking the safety of my car"). And if the irrelevant items are not used as part of the form given to adolescents, then responses to the changed form must be reevaluated psychometrically.

For the scores on the longest form (the original 33-item version), the predicted internal consistency reliability for male adolescents was .53. For a shorter form, the score reliability would be predicted to be even lower. Furthermore, these low predicted values are based on optimal reliability estimates—those actually presented in the papers and dissertations gathered for this study. It can be assumed that Rosenthal's (1979) "file drawer" problem is particularly pertinent to the reporting of reliability estimates because researchers are not obligated to report these coefficients every time groups' scores on a scale are analyzed. Researchers not conducting psychometric analyses do not tend to reevaluate the reliability of their samples' scores. If sample-specific scores are not adequately reliable, then alternate measures should be used to compare or describe groups. But sometimes the realization of poor score reliability comes too late, and the researcher has the option to ignore the scores' low reliability and mask this fact by not reporting the coefficients' low values.

Another warning indicated by the results of this RG study concerns the difference in the reliability of scores for men versus women on the MCSDS. Women's scores tend to show stronger internal consistency reliability. Further analyses are therefore recommended to evaluate the dimensionality of the two genders' responses to the MCSDS.

As noted, score reliability was related to the number of items on the MCSDS used to assess the sample. Although not always true, the longer a test form, the more reliable its scores tend to be. This finding was therefore not surprising, although the practice of ignoring such obvious variables can result in model misspecification. In addition, although the magnitude of the difference in predicted score reliabilities is not substantial for MCSDS versions with differing numbers of items (e.g., the expected reliability of .80 for responses of women to the 33-item test vs. the expected reliability of .75 for women on a 13-item form), researchers must still evaluate the construct validity of scores from differing forms.

Messick (1989) described construct-underrepresentation as a potential threat to the validity of scores. The use of a subset of items taken from a survey threatens construct validity because the subset may not be measuring the same full construct as the original form. Similarly, the finding of no substantial decrease in score reliability for a shorter form does not provide evidence supporting the equivalence of the construct being assessed by the shorter form to that measured by the original longer form. Researchers must continue to investigate the psychometric properties, both reliability and validity, of scores on long-established scales, especially on changed versions of such scales used with differing populations.

One final caution should be made concerning the use of mixed-effects modeling for RG analyses. The underreporting of reliability coefficients and the inconsistency of how samples are described greatly reduces the number of studies and samples that can be analyzed in RG studies. The resulting small sample sizes at the within- and between-studies levels can negatively affect the estimation of the random effects in mixed-effects modeling. However, the use of mixed-effects modeling does provide a better fit with the underlying assumption that score reliability varies and the primary intent of RG studies is to investigate potential sources of this variability.

References

Abreu, J. M. (2000). Counseling expectations among Mexican American college students: The role of counselor ethnicity. *Journal of Multicultural Counseling and Development*, *28*, 30-43.

Abreu, J. M., & Gabarain, G. (2000). Social desirability and Mexican American counselor preferences: Statistical control for a potential confound. *Journal of Counseling Psychology*, *47*, 165-176.

Adams, G. R., Ryan, H. H., Hoffman, J. J., Dobson, W. R., & Nielsen, E. C. (1984). Ego identity status, conformity behavior, and personality in late adolescence. *Journal of Personality & Social Psychology*, *47*, 1091-1104.

Arnold, H. J., & Feldman, D. C. (1981). Social desirability response bias in self-report choice situations. *Academy of Management Journal*, *24*, 377-385.

Arnold, H. J., Feldman, D. C., & Purbhoo, M. (1985). The role of social-desirability response bias in turnover research. *Academy of Management Journal*, *28*, 955-966.

Arrindell, W. A., Hafkenscheid, A. J., & Emmelkamp, P. M. (1984). The hostility and direction of hostility questionnaire (HDHQ): A psychometric evaluation in psychiatric outpatients. *Personality and Individual Differences*, *5*, 221-231.

Ballard, R. (1992). Short forms of the Marlowe-Crowne social desirability scale. *Psychological Reports*, *71*, 1155-1160.

Beatty, M. J., & Payne, S. K. (1983). Speech anxiety as a multiplicative function of size of audience and social desirability. *Perceptual and Motor Skills*, *56*, 792-794.

Becker, G., & Cherny, S. S. (1992). A five-factor nuclear model of socially desirable responding. *Social Behavior and Personality*, *20*, 163-191.

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

Brook, J. S., Lukoff, I. F., & Whiteman, M. (1977). Peer, family, and personality domains as related to adolescents' drug behavior. *Psychological Reports*, *41*, 1095-1102.

Buhrke, R. A. (1988). Factor dimensions across different measures of sex role ideology. *Sex Roles*, *18*, 309-321.

Buras, A. R. (2000). The relationship between family functioning and family roles among college students. *Dissertation Abstracts International*, *60*(9), 4948B. (UMI No. 9943461)

Carr, J. G., Gilroy, F. D., & Sherman, M. F. (1996). Silencing the self and depression among women: The moderating role of race. *Psychology of Women Quarterly*, *20*, 375-392.

*Cramer, D. (2000). Social desirability, adequacy of social support and mental health. *Journal of Community and Applied Social Psychology*, *10*, 465-474.

Crino, M. D., Rubenfeld, S. A., & Willoughby, F. W. (1985). The random response technique as an indicator of questionnaire item social desirability/personal sensitivity. *Educational and Psychological Measurement*, *45*, 453-468.

Crino, M. D., Svoboda, M., Rubenfeld, S., & White, M. C. (1983). Data on the Marlowe-Crowne and Edwards Social Desirability Scales. *Psychological Reports*, *53*, 963-968.

Crowl, T. K. (1984). Grading behavior and teachers' need for social approval. *Education*, *104*, 291-295.

*Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349-354.

*Crowne, D. P., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: Wiley.

*Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden Press.

Eisenberg, N., & Okun, M. A. (1996). The relations of dispositional regulation and emotionality to elders' empathy-related responding and affect while volunteering. *Journal of Personality*, *64*, 157-183.

Eisler, J., Wolfer, J. A., & Diers, D. (1972). Relationship between need for social approval and postoperative recovery and welfare. *Nursing Research*, *21*, 520-525.

Ermann, M. (1984). A specific and taxonomic differentiation between psychovegetative disorders and psychoneuroses. *Psychotherapy and Psychosomatics*, *41*, 116-124.

*Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*, 161-180.

Fisher, G. (1967). Normative and reliability data for the standard and cross-validated Marlowe-Crowne Social Desirability Scale. *Psychological Reports*, *20*, 174.

Fraboni, M., & Cooper, D. (1989). Further validation of the three short forms of the Marlowe-Crowne Scale of Social Desirability. *Psychological Reports*, *65*, 595-600.

Gaines, L. S., Fretz, B. R., & Helweg, G. C. (1975). Self-referent language and need for approval. *Psychological Reports*, *37*, 107-111.

Gisi, T. M. (1999). Evaluating the relationship between traumatic brain injury, anger, and forgiveness. *Dissertation Abstracts International*, *59*(8), 4463B. (UMI No. 9901931)

Gisi, T. M., & D'Amato, R. C. (2000). What factors should be considered in rehabilitation: Are anger, social desirability, and forgiveness related in adults with traumatic brain injuries? *International Journal of Neuroscience*, *105*, 121-133.

Gray, S. (2001). Spiritual well-being and reasons for living: Assessing the connections. *Dissertation Abstracts International*, *61*(8), 4405B. (UMI No. 9982752)

Greenwald, H. J., & O'Connell, S. M. (1970). Comparison of dichotomous and Likert formats. *Psychological Reports*, *27*, 481-482.

Hansen, G. L. (1981). Marital adjustment and conventionalization: A reexamination. *Journal of Marriage & the Family*, *43*, 855-863.

Hanson, R. K., Gizzarelli, R., & Scott, H. (1994). The attitudes of incest offenders: Sexual entitlement and acceptance of sex with children. *Criminal Justice and Behavior*, *21*, 187-202.

*Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Helmcamp, A. M. (1998). Sociocultural and psychological correlates of eating disorder behavior in nonclinical adolescent females. *Dissertation Abstracts International*, *58*(7), 3913B. (UMI No. 9801381)

*Henson, R. K. (2001). Understanding internal consistency estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

*Henson, R. K., & Hwang, D. (2002). Variability and prediction of measurement error in Kolb's Learning Style Inventory scores: A reliability generalization study. *Educational and Psychological Measurement*, *62*, 712-727.

Holden, R. R., & Fekken, G. C. (1989). Three common social desirability scales: Friends, acquaintances, or strangers? *Journal of Research in Personality*, *23*, 180-191.

Johnson, A. L., Luthans, F., & Hennessey, H. W. (1984). The role of locus of control in leader influence behavior. *Personnel Psychology*, *37*, 61-75.

Jome, L. M. (2000). Construct validity of the White Racial Identity Attitude Scale. *Dissertation Abstracts International*, *61*(2), 1133(B). (UMI No. 9963487)

Joubert, C. E. (1991). Self-esteem and social desirability in relation to college students' retrospective perceptions of parental fairness and disciplinary practices. *Psychological Reports*, *69*, 115-120.

Kameoka, V. A. (1986). Reliabilities and concurrent validities of popular self-report measures of depression, anxiety, and social desirability. *Journal of Consulting & Clinical Psychology*, *54*, 328-333.

Kisler, V. A. (1997). Perceptions and metaperceptions of same-sex social interactions in college women with troubled eating patterns. *Dissertation Abstracts International*, *58*(9), 5124B. (UMI No. 9808828)

Kohn, P. M., Cowles, M. P., & Dzinas, K. (1989). Arousability, need for approval, and situational context as factors in pain tolerance. *Journal of Research in Personality*, *23*, 214-224.

Koski, M. J. (1998). Burnout in gay men caring for a partner with aids: The relationship between social support, coping with HIV, and level of intimacy. *Dissertation Abstracts International*, *58*(12), 6813B. (UMI No. 9820186)

Kraft, M. E. (1999). The relationship between the repressive coping style and breast cancer incidence. *Dissertation Abstracts International*, *59*(9-B), 5154. (UMI No. 9907424)

Krasnoff, A. (1973). Self-reported attitudes toward drinking among alcoholics before and after treatment. *Quarterly Journal of Studies on Alcohol*, *34*, 947-950.

*Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Kummel, P. E. (1999). Bringing your family to work: Attachment in the workplace. *Dissertation Abstracts International*, *59*(7), 3747B. (UMI No. 9838962)

Kusyszyn, I., & Greenwood, D. E. (1970). Marlowe-Crowne defensiveness and personality scale faking. *Proceedings of the Annual Convention of the American Psychological Association*, *5*, 343-344.

Lapsley, D. K., & Enright, R. D. (1979). The effects of social desirability, intelligence, and milieu on an American validation of the Conservatism scale. *Journal of Social Psychology*, *107*, 9-14.

Leisen, M. B. (2000). Development and validation of the Adolescent Partner Aggression Scale (APAS). *Dissertation Abstracts International*, *61*(4), 2207B. (UMI No. 9968040)

Leister, K. D. (1999). The relationship between sexual aggression and moral development in a sample of college males. *Dissertation Abstracts International*, *60*(6), 2949B. (UMI No. 9934446)

Martinez, R.D.P. (1997). Development of a Psychotherapist Multicultural Attitudes Inventory. *Dissertation Abstracts International*, *58*(6), 3363B. (UMI No. 9723737)

Mathisen, J. H. (2000). Stigma busting: Does strategic contact with individuals with severe mental illness reduce negative attitudes in an adolescent population? *Dissertation Abstracts International*, *60*(7), 3572B. (UMI No. 9937659)

McCarrey, M. W., Dayhaw, L. T., & Chagnon, G. P. (1971). Attitude shift, approval need, and extent of psychological differentiation. *Journal of Social Psychology*, *84*, 141-149.

McFarland, S. G., & Sparks, C. M. (1985). Age, education, and the internal consistency of personality scales. *Journal of Personality & Social Psychology*, *49*, 1692-1702.

Mckenzie, J. S. (1998). An exploration of the relationship between forgiveness, anger, and depression with prisoners in a drug abuse treatment program. *Dissertation Abstracts International*, *58*(8), 4507B. (UMI No. 9806130)

Mcman, J. C. (1998). The relationship between adult attachment styles of nursing aides and nursing aides' attitudes toward caregiving. *Dissertation Abstracts International*, *58*(11), 6282B. (UMI No. 9815803)

Melamed, S. (1996). Emotional reactivity, defensiveness, and ambulatory cardiovascular response at work. *Psychosomatic Medicine*, *58*, 500-507.

*Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York: American Council on Education.

Middleton, K. L., & Jones, J. L. (2000). Socially desirable response sets: The impact of country culture. *Psychology and Marketing*, *17*, 149-163.

Movius, H. L. (2000). Cardiac vagal tone as a predictor of defensiveness, openness, and self-regulatory style. *Dissertation Abstracts International*, *61*(3), 1686B. (UMI No. 9965881)

*Newsom, J. T., & Nishishiba, M. (2002). *Non-convergence and sample bias in hierarchical linear modeling of dyadic data*. Manuscript submitted for publication.

Newsom, W. S. (1999). Measuring interpersonal violation. *Dissertation Abstracts International*, *59*(7), 3766B. (UMI No. 9841988)

Nordholm, L. A. (1974). A note on the reliability and validity of the Marlowe-Crowne Scale of social desirability. *Journal of Social Psychology*, *93*, 139-140.

O'Gorman, J. G. (1974). Limits to the generality of the Marlowe-Crowne measure of social desirability. *Journal of Clinical Psychology*, *30*, 81.

O'Grady, K. E. (1988). The Marlowe-Crowne and Edwards Social Desirability scales: A psychometric perspective. *Multivariate Behavioral Research*, *23*, 87-101.

Olson, L. M. (1999). The assessment of moral integrity among adolescents and adults. *Dissertation Abstracts International*, *60*(6), 2989B. (UMI No. 9910484)

*Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, *7*(1). Retrieved March 1, 2002, from http://ericae.net/pare/getvn.asp?v=7&n=1

Park, C. L., Cohen, L. H., & Murch, R. L. (1996). Assessment and prediction of stress-related growth. *Journal of Personality*, *64*, 71-105.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality & Social Psychology*, *46*, 598-609.

*Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Pettit, F. A. (2000). Response sets in World Wide Web and paper-and-pencil personality questionnaires. *Dissertation Abstracts International*, *60*(8), 4287B. (UMI No. NQ39301)

Platow, M. J. (1994). An evaluation of the social desirability of prosocial self-other allocation choices. *Journal of Social Psychology*, *134*, 61-68.

Prasad, M. B., & Sinha, B. P. (1980). Need-achievement and defence-orientation. *Psychological Studies*, *25*, 66-75.

Ramanaiah, N. V., & Martin, H. J. (1980). On the two-dimensional nature of the Marlowe-Crowne Social Desirability Scale. *Journal of Personality Assessment*, *44*, 507-514.

*Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*, 75-98.

Ray, J. J. (1979). The authoritarian as measured by a personality scale: Solid citizen or misfit? *Journal of Clinical Psychology*, *35*, 744-747.

Ray, J. J. (1984). The reliability of short social desirability scales. *Journal of Social Psychology*, *123*, 133-134.

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, *38*, 119-125.

Ringel, P. Z. (2000). Effects of optimism and repressive coping on self-report versus behavioral outcomes. *Dissertation Abstracts International*, *60*(12), 6351B. (UMI No. 9955172)

*Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Roy, M. (1982). Social desirability responses to tribal and non tribal subjects. *Psychological Studies*, *27*, 17-19.

Rusbult, C. E., Johnson, D. J., & Morrow, G. D. (1986). Impact of couple patterns of problem solving on distress and nondistress in dating relationships. *Journal of Personality & Social Psychology*, *50*, 744-753.

Schriesheim, C. A. (1979). Social desirability and leader effectiveness. *Journal of Social Psychology*, *108*, 89-94.

Schumm, W. R., Nichols, C. W., Schectman, K. L., & Grigsby, C. C. (1983). Characteristics of responses to the Kansas Marital Satisfaction Scale by a sample of 84 married mothers. *Psychological Reports*, *53*, 567-572.

Schumm, W. R., Scanlon, E. D., Crow, C. L., Green, D. M., & Buckler, D. L. (1983). Characteristics of the Kansas Marital Satisfaction Scale in a sample of 79 married couples. *Psychological Reports*, *53*, 583-588.

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, *28*, 191-193.

Strickland, L. H. (1968). Changes in self-presentation in need for approval scores. *Perceptual and Motor Skills*, *27*, 335-337.

St.-Yves, A., Contant, F., Freeston, M. H., Huard, J., & Lemieux, B. (1989). Locus of control in women occupying middle-management and nonmanagement positions. *Psychological Reports*, *65*, 483-486.

Thomas, C. B. (1979). Evaluation apprehension, social desirability, and the interpretation of test correlations. *Social Behavior and Personality*, *7*, 193-197.

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

Towbes, L. C., & Cohen, L. H. (1996). Chronic stress in the lives of college students: Scale development and prospective prediction of distress. *Journal of Youth and Adolescence*, *25*, 199-217.

Turner, C. (1971). Effects of race of tester and need for approval on children's learning. *Journal of Educational Psychology*, *62*, 240-244.

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

Venable, W. M. (1997). Caretaker psychological factors predicting compliance with children's psychotherapy. *Dissertation Abstracts International*, *57*(8), 5348B. (UMI No. 9701730)

Vleeming, R. G. (1979). Machiavellianism: Some problems with a Dutch Mach V scale. *Psychological Reports*, *45*, 715-718.

White, G. L. (1984). Comparison of four jealousy scales. *Journal of Research in Personality*, *18*, 115-130.

Wichstrom, L. (1995). Harter's Self-Perception Profile for Adolescents: Reliability, validity, and evaluation of the question format. *Journal of Personality Assessment*, *65*, 100-116.

Williams, R. L. (1999). Social-information processing in lower and higher aggressive Black adolescents. *Dissertation Abstracts International*, *59*(7), 3742B. (UMI No. 9840741)

Wilson, J. F. (1982). Recovery from surgery and scores on the Defense Mechanisms Inventory. *Journal of Personality Assessment*, *46*, 312-319.

Wong, F. Y., McCreary, D. R., & Duffy, K. G. (1990). A further validation of the Bem Sex Role Inventory: A multitrait-multimethod study. *Sex Roles*, *22*, 249-259.

Wortley, R. K., & Homel, R. J. (1995). Police prejudice as a function of training and outgroup contact: A longitudinal investigation. *Law and Human Behavior*, *19*, 305-317.

Zook, A., & Sipps, G. J. (1985). Cross-validation of a short form of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, *41*, 236-238.

References marked with an asterisk indicate studies not included in this meta-analysis.