An Empirical Demonstration of the Existence of Measurement Dependence in the Results of a Meta-Analysis

William R. Nugent *University of Tennessee*
Sukyung Yoon *University of Tennessee*
Jayme Walters *University of Tennessee*

ABSTRACT ***Objective:*** Findings from meta-analytic studies that use standardized mean differences (SMDs) may be overly dependent on the original measures that were used to generate SMDs. This may be particularly true when measures have arbitrary metrics or when measures fail to meet measurement equivalence. We test the hypothesis that in such cases, meta-analytic results may vary significantly—statistically and practically—as a function of the measures used to derive SMDs. ***Methods:*** We conducted 5 secondary random-effects meta-analyses of SMDs—each under a different measurement scenario—from a published meta-analysis comparing the efficacy of cognitive–behavioral therapy with that of reminiscence therapy for depression in older adults. In each scenario, SMDs were based on scores from measures with arbitrary metrics, some of which failed to meet measurement equivalence. ***Results:*** Consistent with the hypothesis, meta-analysis results differed significantly—statistically and practically—between the measurement scenarios under conditions of measurement nonequivalence. ***Conclusions:*** Results of meta-analyses involving measures with arbitrary metrics may depend on the measures that the SMDs are based on when measurement equivalence fails to hold. Inferences concerning the relative efficacy of different treatments can be measurement dependent.

KEYWORDS: meta-analysis, measurement, effect sizes, measurement equivalence, arbitrary metrics

doi: 10.1086/699248

As meta-analysis has become a preferred method for identifying evidence-based interventions (e.g., Murad, Asi, Alsawas, & Alahdab, 2016; Rubin & Bellamy, 2012), the number of published meta-analyses has increased rapidly (Borenstein, Hedges, Higgins, & Rothstein, 2009; White, 2009). Often, different studies use different measures. Consequently, the meta-analyst must accumulate and compare effect sizes (ESs) based on scores from different measures that frequently

Journal of the Society for Social Work and Research, Volume 10, Number 1. 2334-2315/2018/1001-0008/$10.00.

have *arbitrary metrics* (Kazdin, 2006). Blanton and Jaccard (2006, p. 27) define a metric as arbitrary "when it is not known where a given score locates an individual on the underlying psychological dimension or how a one-unit change on the observed score reflects the magnitude of change on the underlying dimension." Unique to each scale (Lord & Novick, 1968), arbitrary metrics are problematic not only because of their inherent uncertainties but also because the relationships between these metrics are unknown and thus, the scores cannot be directly compared (Dorans, Pommerich, & Holland, 2010). Trying to directly compare them would be like trying to compare the temperature in Town X (35 °C) with that in Town Y (95 °F) without knowing the relationship between the units on the Fahrenheit and Celsius scales—°F = (1.8 × °C) + 32—that facilitates the direct comparison (35 °C = 95 °F).

The use of different measures in different studies and the associated score comparison problems led Lipsey and Wilson (2001) to frame the following question in their introduction to meta-analysis:

> With these quite different measures yielding different numerical values that are meaningful only in relation to the specific operationalization and scales used, how can their quantitative findings be encoded in a way that allows them to be statistically combined and compared? (p. 4)

Lipsey and Wilson (2001, p. 4) emphasized that the answer to this question "relates to an essential feature of meta-analysis, indeed, *the feature that makes meta-analysis possible and provides the hub around which the entire process revolves* [emphasis added]."

Lipsey and Wilson's (2001) answer to this question is that ESs are *standardized*. Because ESs represent a statistical standardization of study findings, they are presumably interpretable in a consistent manner across all measures involved (Lipsey & Wilson, 2001). In the words of Grissom and Kim (2005, p. 49), this creates "a measure of effect size that *places different dependent variable measures on the same scale* [emphasis added]." This explanation suggests that ESs—such as the standardized mean difference (SMD), which is commonly used in treatment outcome studies—have taken the scores from different measures, with different and arbitrary metrics, and transformed them into ESs expressed on the same numerical scale in much the same way that the temperatures in Town X and Town Y can be converted to the same temperature scale (e.g., °F) using the relationship between the Fahrenheit and Celsius scales and then be compared.

For any construct of interest, a number (m) of measurement procedures (j) exist ($j = 1, 2 \ldots m$; the measures are expressed as $j = 1, j = 2 \ldots j = m$) and produce scores that are inferred to represent the construct of interest. The number m can be rather large; for example Mitchell (2010) identified 50 general depression scales. The true SMD for Study i ($i = 1, 2 \ldots n$) based on scores from Measure j is symbolized by $\delta_i(j)$; for example, the true SMD for Study 1 based on the scores from Mea-

sure $j = 2$ would be $\delta_1(2)$. In some cases, the true SMD for Study i may be based on the average of g true SMDs; the true SMDs are based on scores from g ($g \leq m$) of the m measures (Borenstein et al., 2009). In this case, the true SMD would be symbolized by $\delta_i(1, 2 \ldots g) = \text{avg}[\delta_i(1), \delta_i(2) \ldots \delta_i(g)]$ since this SMD is an average and involves the true SMDs based on scores from the g measures. For example, say the true SMD for Study 2 is the average of the true SMDs based on scores from Measures $j = 1$ and $j = 5$. The true SMD for Study 2 would thus be symbolized by $\delta_2(1, 5) = \text{avg}[\delta_2(1), \delta_2(5)]$. Of course, the meta-analyst will not have the true SMD for any study; rather, they will have an estimate of the true SMD. The estimate of a true SMD will be symbolized the same way as the true SMD but with the lowercase d in place of the lowercase Greek δ. Thus, for example, the estimated true SMD for Study 1 based on the scores from Measure $j = 2$ will be symbolized as $d_1(2)$.

It is important to note that estimated SMDs for any study (i) based on the scores from all measures (m) will almost certainly *not* be available to a meta-analyst; that is, $d_i(1)$, $d_i(2) \ldots d_i(m)$ will not all be available. Researchers select one or more measures for use in their studies based on considerations of their particular research exigencies, and it is unlikely that all m measures will be used (especially if m is large). To illustrate, suppose that a group of three studies is to be meta-analyzed and that the outcome variable in these studies could, with justification, have been measured using any of four measures ($j = 1$ through $j = 4$). This is illustrated in Figure 1. In this figure, the left-hand column shows the measures used in Studies 1–3 (above the line) as well as the estimated SMDs available to a meta-analyst for each study (below the line). Measure $j = 1$ was used in Study 1, Measures $j = 3$ and $j = 4$ were used in Study 2, and $j = 4$ was used in Study 3; Measure $j = 2$ was not used in any of the studies. This collection of measures used in the three studies is called the *measurement scenario* for the three studies and is symbolized as $\{(j = 1)_1, (j = 3, j = 4)_2, (j = 4)_3\}$, where the measures used in the studies are contained within parentheses and the subscript attached to each parenthesis indicates the particular study. This measurement scenario leads to the estimated SMDs—$d_1(1)$ for Study 1; $d_2(3)$, $d_2(4)$, and, possibly, $\text{avg}[d_2(3), d_2(4)]$ for Study 2; and $d_3(4)$ for Study 3—that are available to the meta-analyst. In Figure 1, the right-hand column shows, above the line, the measures *not* used in the studies; the estimated SMDs *not* available to the meta-analyst are shown below the line. Because Measures $j = 2$, $j = 3$, and $j = 4$ were not used in Study 1; $j = 1$ and $j = 2$ were not used in Study 2; and $j = 1$, $j = 2$, and $j = 3$ were not used in Study 3, any estimated SMDs based on scores from these measures are unavailable to a meta-analyst. These inestimable SMDs are a form of what Gilovich (1991) called *absent data*—in this case, estimated SMDs that cannot be computed and hence are unavailable to the meta-analyst. This scenario readily generalizes to circumstances in which there are m reasonable measures and n studies.

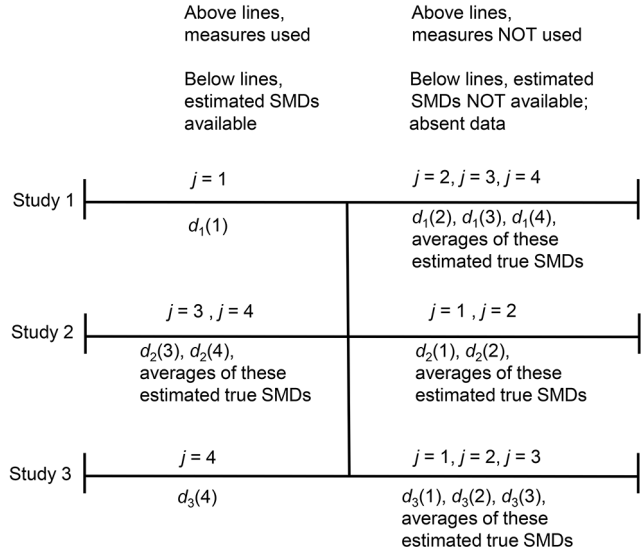

Figure 1. Illustration of hypothetical measurement scenario for three studies. The left-hand column shows measures actually used in the hypothetical studies, and the right-hand column shows the measures *not* used in the hypothetical studies. The estimated SMDs below the line in the right-hand column are absent data. SMD = standardized mean difference.

If arguments by methodologists such as Lipsey and Wilson (2001) and Grissom and Kim (2005)—that the standardization inherent in the SMD places estimated true SMDs on the same scale—are assumed to be correct, then the true SMD for Study i based on scores from Measure $j = 1$, $\delta_i(1)$ will be equal to that for Study i based on the scores from a different measure—$j = 2$, $\delta_i(2)$—even if the metrics of $j = 1$ and $j = 2$ are arbitrary, different, and the relationship between them is unknown. Similarly, the *estimated* true SMDs $d_i(1)$ and $d_i(2)$ will be on the same scale and will therefore be statistically interchangeable since, although they will differ due to sampling variability, they will nonetheless be estimates of the same true SMD. Consequently, all measurement scenarios will be statistically interchangeable and the results of meta-analyses of the n studies will be the same, within limits of sampling variability, across all possible measurement scenarios involving the m measures. It will not matter which measures are used in which studies. This is, theoretically, what makes meta-analysis so informative and valuable for research synthesis.

It has been argued that true SMDs for Study i based on scores from two measures ($j = 1$ and $j = 2$) will be equal *only* when two measurement conditions—*construct equivalence* and *equal reliabilities*—simultaneously hold for the scores from Measures $j = 1$ and $j = 2$ in all subpopulations of a population of interest (Nugent, 2012). The construct equivalence condition is formally expressed by Equation 1,

$$T_1 = (F_{12} \times T_2) + H_{12}, \quad (1)$$

where T_1 and T_2 represent the true scores on Measures $j = 1$ and $j = 2$, respectively, and F_{12} and H_{12} represent real-number parameters defining the relationship between the metrics of T_1 and T_2. Conceptually, this relationship implies that the scores from Measures $j = 1$ and $j = 2$ represent the exact same construct (Joreskog, 1971). The equal reliabilities condition asserts that the reliability coefficients for the scores from Measures $j = 1$ and $j = 2$ are equal. The conjunction of construct equivalence and equal reliabilities defines a form of measurement equivalence, stated conceptually as *Measures j = 1 and j = 2 produce scores representing the exact same construct, with the exact same reliability, in all subpopulations of the population of interest*. For the sake of brevity, we will hereafter refer to the term "the conjunction of construct equivalence and equal reliabilities" as CE (for construct equivalence) and ER (for equal reliabilities), or *CE and ER*.

If CE and ER does *not* hold—particularly if CE fails to hold—the scores from one or more of the m measures will *not* represent the exact same construct with the same reliability (Solow, 2002). Thus, vectors of estimated SMDs input into a meta-analysis can differ statistically. Consequently, meta-analytic results can differ based on which vectors of estimated SMDs are input into the meta-analysis, and thus which measures are used in which studies. Nugent (2012, 2013) argued that the greater the extent to which the estimated SMD for any study (i) in a group of n studies differs between measurement scenarios as a consequence of violations of CE and ER, and the larger the fraction of the n studies for which these large differences exist, the greater the likelihood that results of a meta-analysis of the studies will differ between measurement scenarios and, therefore, be measurement dependent.

The Current Study

The current study augments previous research by empirically demonstrating the measurement dependence of meta-analytic results. Nugent (2013) conducted a model-based simulation showing that violation of the CE condition can lead to large differences between $\delta_i(1)$ and $\delta_i(2)$ for Study i, and that simultaneous violations of CE and ER can lead not only to greater differences between $\delta_i(1)$ and $\delta_i(2)$ for Study i but also to different rank orderings of these true SMDs for the same pair of studies. More recently, Nugent (2017) conducted a Monte Carlo simulation testing the conjecture that meta-analytic results can vary as a consequence of true SMDs based on scores from measures that violate CE. Findings suggested that violations of CE can lead to contradictory results in a meta-analysis of the same set of studies. The results of these model-based simulations can be considered as hypotheses, or predictions, that require empirical verification (Banks, 2009; Harrison, Carroll, & Carly, 2007). However, no empirical studies have investigated whether the results of actual

meta-analyses may be measurement dependent. The current study addresses this absence.

Method

The secondary analyses in this study were of estimated SMDs comparing treatments for depression in older adults; the estimated SMDs were from eight studies of randomized clinical trials that were included in a systematic review conducted by Peng, Huang, Chen, and Lu (2009). The secondary meta-analyses involved 11 comparisons of the efficacy of types of cognitive–behavioral therapy (e.g., problem-solving therapy, cognitive and behavioral bibliotherapy, and cognitive and behavioral self-management)—the various forms of which we will subsequently refer to as "cognitive–behavioral therapy"—and various forms of reminiscence therapy, which we will henceforth refer to as "reminiscence therapy." Seven of the studies compared cognitive–behavioral therapy with waiting-list control or delayed treatment, and four studies compared reminiscence therapy with waiting list control or delayed treatment. We will subsequently refer to the waiting-list control and delayed-treatment conditions as "no treatment." We included a single estimated SMD for the direct comparison of cognitive–behavioral therapy with reminiscence therapy by Arean and colleagues (1993) because this estimated SMD was correlated with two others from this study, and this inclusion brought the information contained in these covariances into analyses (Gleser & Olkin, 2009). Readers are referred to Peng et al. (2009) for details of this systematic review and meta-analysis.

Measures

The measures used in the eight studies included in the secondary meta-analyses were the Beck Depression Inventory (Beck, Steer, & Garbin, 1988); the Center for Epidemiologic Studies Depression scale (Radloff, 1977); the Hamilton Rating Scale (Hamilton, 1960); the Geriatric Depression Scale (Sheikh & Yesavage, 1986); and the Brief Symptom Inventory depression subscale (Derogatis & Melisarotos, 1983). All of these frequently used and well-known measures have arbitrary metrics. Three measures were used in three of the studies, two were used in three of the studies, and one was used in two of the studies. Table 1 displays the measures upon which the estimated SMDs were based in the current study.

Estimated SMDs

Each of the authors independently computed the estimated SMDs for treatment comparisons and then compared results. There was 100% agreement for the SMD estimates. We computed the estimated SMDs from information in the published studies using formulas from Borenstein et al. (2009) and Gleser and Olkin (2009). The estimated true SMD for a treatment/no-treatment comparison from Study i based on scores from Measure j was estimated by

$$d_i(j) = \frac{\bar{Y}(j)_{g1} - \bar{Y}(j)_{g2}}{s_p(j)}, \tag{2}$$

where $\bar{Y}(j)_{g1}$ was the sample mean score on the dependent variable for Group 1 and $\bar{Y}(j)_{g2}$ was the sample mean score on the dependent variable for Group 2, based on scores from Measure j. Also, $s_p(j)$ was the pooled sample standard deviation based on scores from Measure j, given by

$$s_p(j) = \sqrt{\frac{\left[(n_{g1} - 1) \times s^2(j)_{g1}\right] + \left[(n_{g2} - 1) \times s^2(j)_{g2}\right]}{n_{g1} + n_{g2} - 2}}, \tag{3}$$

where n_{g1} was the sample size of Group 1 and n_{g2} was the sample size for Group 2; $s^2(j)_{g1}$ was the sample variance of scores for Group 1 and $s^2(j)_{g2}$ was the sample variance of scores for Group 2, based on the scores from Measure j. These estimated SMDs were corrected for small-sample bias by multiplying them by

$$J = 1 - \frac{3}{4(n_{g1} + n_{g2} - 2) - 1}. \tag{4}$$

We estimated the small-sample bias-corrected variances of sample estimates of $d_i(j)$—symbolized as $\text{var}[d_i(j)]$—by

$$\text{var}[d_i(j)] = J^2 \times \left(\frac{1}{n_{g1}} + \frac{1}{n_{g2}} + \frac{d_i^2(j)}{2 \times n_{total}}\right), \tag{5}$$

where n_{total} was the total number of people in the study (Gleser & Olkin, 2009).

In studies in which multiple treatments were compared with a common no-treatment group, the values of $d_i(j)$ for treatment group versus no-treatment group comparisons were correlated due to sharing the same no-treatment group (Gleser & Olkin, 2009; Higgins & Green, 2011). We computed the covariance between estimates of $d_i(j)$ in these cases—symbolized as $\text{cov}(d_i(j)_{tx1}, d_i(j)_{tx2})$—from

$$\text{cov}(d_i(j)_{tx1}, d_i(j)_{tx2}) = \frac{1}{n_c} + \frac{d_i(j)_{tx1} \times d_i(j)_{tx2}}{2 \times n_{total}}, \tag{6}$$

where n_c was the number of people in the shared no-treatment group, and $d_i(j)_{tx1}$ and $d_i(j)_{tx2}$ were the estimated SMDs based on scores from Measure j for comparing Treatment 1 (tx1) with the common no-treatment group and Treatment 2 (tx2) with the common no-treatment group, respectively (Gleser & Olkin, 2009).

In three of the five measurement scenarios described later, we estimated unweighted mean SMDs for studies in which multiple measures were used. We computed the variances of these estimated SMDs using Equation 24.5 from Borenstein et al. (2009, p. 230). In the interest of clarity, this complex formula is shown here using the symbolism we defined earlier. For an estimated SMD that was the average

Table 1

Studies, Treatment Comparisons, Measures, and Symbolic Representations of Estimated SMDs Included in the Secondary Meta-Analyses

Study and Treatments Compared	Measurement Scenario 1	Measurement Scenario 2A	Measurement Scenario 2B	Measurement Scenario 3A	Measurement Scenario 3B
Arean et al. (1993)—CBT vs. ntx	HRS, GDS, BDI **avg[d_1(HRS), d_1(GDS), d_1(BDI)]**	HRS* **d_1(HRS)**	BDI* **d_1(BDI)**	HRS** **d_1(HRS)**	GDS and BDI** **avg[d_1(GDS), d_1(BDI)]**
Arean et al. (1993)—RT vs. ntx	HRS, GDS, BDI **avg[d_8(HRS), d_8(GDS), d_8(BDI)]**	GDS **d_8(HRS)**	HRS **d_8(BDI)**	HRS, GDS, BDI **d_8(HRS)**	HRS, GDS, BDI **avg[d_8(GDS), d_8(BDI)]**
Floyd, Scogin, McKendree-Smith, Floyd, & Rokke (2004)—CBT vs. ntx	HRS, GDS **avg[d_2(HRS), d_2(GDS)]**	HRS **d_2(HRS)**	GDS **d_2(GDS)**	HRS, GDS **avg[d_2(HRS), d_2(GDS)]**	HRS, GDS **avg[d_2(HRS), d_2(GDS)]**
Mastel-Smith, McFarlane, Sierpina, Malecha, A., & Haile (2007)—RT vs. ntx	BSI-D **d_9(BSI-D)**	BSI-D **d_9(BSI-D)**	BSI-D **d_9(BSI-D)**	BSI-D **d_9(BSI-D)**	BSI-D **d_9(BSI-D)**
Rokke, Tomhave, & Jocic (1999)—CBT vs. ntx	BDI, GDS, HRS **avg[d_3(HRS), d_3(GDS), d_3(BDI)]**	HRS* **d_3(HRS)**	BDI* **d_3(BDI)**	GDS and HRS** **avg[d_3(HRS), d_3(GDS)]**	BDI** **d_3(BDI)**
Rokke et al. (1999)—CBT vs. ntx	BDI, GDS, HRS **avg[d_4(HRS), d_4(GDS), d_4(BDI)]**	HRS* **d_4(HRS)**	BDI* **d_4(BDI)**	GDS, HRS **avg[d_4(HRS), d_4(GDS)]**	BDI **d_4(BDI)**

Scogin, Hamblin, & Beutler (1987)—CBT vs. ntx	HRS, GDS, BDI **avg[d_7(HRS), d_7(GDS), d_7(BDI)]**	HRS* **d_7(HRS)**	BDI* **d_7(BDI)**	HRS** **d_7(HRS)**	GDS and BDI** **avg[d_7(GDS), d_7(BDI)]**
Scogin, Jamison, & Gochneaur (1989)—CBT vs. ntx	HRS, GDS **avg[d_5(GDS), d_5(HRS)]**	HRS* **d_5(HRS)**	GDS* **d_5(GDS)**	HRS** **d_5(HRS)**	GDS** **d_5(GDS)**
Scogin et al. (1989)—CBT vs. ntx	HRS, GDS **avg[d_6(GDS), d_6(HRS)]**	HRS* **d_6(HRS)**	GDS* **d_6(GDS)**	HRS** **d_6(HRS)**	GDS** **d_6(GDS)**
Serrano, Latorre, Gatz, & Montanes (2004)—RT vs. ntx	CES-D **d_{10}(CES-D)**	BDI **d_{10}(CES-D)**	CES-D **d_{10}(CES-D)**	CES-D, BDI **d_{10}(CES-D)**	CES-D, BDI **d_{10}(CES-D)**
Wang, Hsu, & Cheng (2005)—RT vs. ntx	GDS **d_{11}(GDS)**	GDS **d_{11}(GDS)**	GDS **d_{11}(GDS)**	GDS **d_{11}(GDS)**	GDS **d_{11}(GDS)**

Note. Symbolic representations of estimated true standardized mean differences (SMDs) are indicted in **bold type**. For example, **d_1(BDI)** is the estimated true SMD for Study 1 (Arean et al., 1993) based on scores from the Beck Depression Inventory (BDI). BSI-D = Brief Symptom Inventory depression subscale; CES-D = Center for Epidemiologic Studies Depression scale; CBT = cognitive–behavioral therapy; GDS = Geriatric Depression Scale; HRS = Hamilton Rating Scale; ntx = no treatment; RT = reminiscence therapy.

*Estimated SMDs are statistically different between Measurement Scenarios 2A and 2B

**Estimated SMDs are statistically different between Measurement Scenarios 3A and 3B.

of three estimated SMDs—say, $d_i(1)$, $d_i(2)$, and $d_i(3)$—based on scores from Measures $j = 1$, $j = 2$, and $j = 3$, this formula is

$$\begin{aligned}\text{var}\{\text{avg}[d_i(1), d_i(2), d_i(3)]\} = \left(\frac{1}{3}\right)^2 \{&\text{var}[d_i(1)] + \text{var}[d_i(2)] + \text{var}[d_i(3)] \\ &+ 2r_{j=1,\, j=2}\sqrt{\text{var}[d_i(1)]}\sqrt{\text{var}[d_i(2)]} + 2r_{j=1,\, j=3}\sqrt{\text{var}[d_i(1)]}\sqrt{\text{var}[d_i(3)]} \\ &+ 2r_{j=2,\, j=3}\sqrt{\text{var}[d_i(2)]}\sqrt{\text{var}[d_i(3)]}\},\end{aligned} \tag{7}$$

where $r_{j\,=\,1,\,j\,=\,2}$ is the correlation between the scores on Measures $j = 1$ and $j = 2$, and similarly for the correlations $r_{j\,=\,1,\,j\,=\,3}$ and $r_{j\,=\,2,\,j\,=\,3}$. If there were only two measures—for instance, $j = 1$ and $j = 2$—the terms involving $r_{j\,=\,1,\,j\,=\,3}$ and $r_{j\,=\,2,\,j\,=\,3}$ drop out of this equation; the multiplicative term at the front of the right-hand side of this formula would be $(1/2)^2$. Because estimates of these correlations were not reported in any of the studies in the Peng et al. (2009) meta-analysis, per Borenstein et al. (2009), we conducted analyses for upper and lower ends of a range of plausible values of the correlations, specifically .90 and .70. In the interest of brevity, we are reporting only the results assuming the correlations were .90.

Tests of Construct Equivalence and Equal Reliabilities

Hedges and Olkin (1985, pp. 210–212) described a test of the null hypothesis for a study (i) in which g different measures were used to measure the dependent variable:

$$H_0 : \delta_i(1) = \delta_i(2) = \ldots\ldots = \delta_i(g) = \delta_i. \tag{8}$$

Expressed in words, this null hypothesis states: The estimated population of true SMDs for Study i based on the scores from the g measures of presumably the same construct used in Study i is an estimate of a mutual (i.e., the same) SMD (δ_i). The alternate hypothesis (H_1) is that one or more of the g estimated SMDs for Study i based on scores from the g measures is *not* an estimate of the mutual true SMD (δ_i). We used this method to test the plausibility that CE and ER held for the scores from the $g = 2$ or $g = 3$ measures upon which estimated SMDs were based in studies using multiple measures of depression. Statistically nonsignificant results of this test would suggest it was plausible to assume that CE and ER held for the scores from the g measures used in the study; statistically significant results would suggest it was plausible to assume CE and ER did *not* hold.

We computed the chi-square statistic for this test using the matrix equation (the letters in **bold type** indicate matrices; see Equation 7 in Hedges & Olkin, 1985, p. 211),

$$\chi^2(g - 1) = \boldsymbol{d}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{d}, \tag{9}$$

where g was the number of measures (and estimated SMDs) in the study, $\boldsymbol{d}$ was a column vector of estimated SMDs for the given study based on the scores from the g dif-

ferent depression measures, and $\boldsymbol{d}^{\mathrm{T}}$ was the transpose of this column vector. In this equation,

$$\boldsymbol{M} = \boldsymbol{\Lambda} - \left(1/\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{e}\right)\boldsymbol{\Lambda}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Lambda}, \tag{10}$$

$\boldsymbol{\Lambda}$ was the inverse of the variance–covariance matrix for the g estimated SMDs for Study i based on the scores from the different measures, and $\boldsymbol{e}$ was a column vector of ones, the number of which was equal to g; $\boldsymbol{e}^{\mathrm{T}}$ was the transpose of this vector (see Equation 8 in Hedges & Olkin, 1985, p. 211).

The variance–covariance matrix of the g estimated SMDs ($\sum_g$) was estimated by

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{g}} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}, \tag{11}$$

(Hedges & Olkin, 1985, p. 211), where $\boldsymbol{D}$ was a diagonal matrix (the elements of which were the estimated standard deviations of the estimates of the true SMDs given by the square root of Equation 5), and $\boldsymbol{R}$ was the matrix of correlations between the scores from the g different measures used in Study i. Assuming a range of plausible correlations from .70–.90, we computed the chi-square statistics at values of .70 and .90. If the results of the chi-square test produced statistically significant results at a correlation of either .70 or .90, we inferred that it was plausible that the SMDs were based on scores from different measures that failed to meet CE and ER.

We used a two-step procedure. First, we conducted an omnibus test, which tested the null hypothesis that at least one of the g estimated SMDs was based on scores that failed to meet CE and ER. If this test was statistically significant, then we conducted pair-wise tests to determine which estimated SMDs were based on scores failing to meet CE and ER.

Measurement Scenarios

Our general scheme was to create measurement scenarios meeting the conditions that Nugent (2013) identified as most likely to lead to different meta-analytic outcomes as a consequence of measurement nonequivalence. Thus, measurement scenarios were created such that

- sizable percentages of large, statistically significant differences existed between the estimated SMDs for given treatment/no-treatment comparisons between scenarios but based on different measures in the studies included in Peng et al.'s (2009) meta-analysis; and
- if there were multiple treatment/no-treatment comparisons in a study, the estimated SMDs for all comparisons were based on scores from the same measure or average of measures. We considered it unlikely that a researcher would use one measure, or combination of measures, for a treatment/no-

treatment comparison, but then use a different measure, or combination of measures, for a second such comparison within the same study.

Measurement Scenario 1. The estimated SMD for a given treatment/no-treatment comparison in this scenario was the average across the *g* estimated SMDs based on scores from the *g* measures used in the study. Symbolic representations of the estimated SMDs and the measures they were based on in this scenario are shown in Table 1.

Measurement Scenarios 2A and 2B. These scenarios exemplified circumstances in which researchers used only a single measure of depression in their study. The measures that estimated SMDs were based on, and symbolic representations of the estimated SMDs in these scenarios, are shown in Table 1. For comparisons of cognitive–behavioral therapy with no treatment, we created these scenarios as follows:

- If a study included a single comparison of cognitive–behavioral therapy with no treatment, we used the largest estimated SMD in Scenario 2A and the smallest in Scenario 2B.
- If a study compared two forms of cognitive–behavioral therapy with no treatment, we used the largest SMDs for the two treatments *based on the same measure* in Scenario 2A, and we used the smallest based on the same measure in Scenario 2B.

For comparisons of reminiscence therapy with no treatment:

- For any study in which only reminiscence therapy was compared with no treatment, if there was a single estimated SMD based on a single measure, we used that SMD in Scenarios 2A and 2B.
- For the study in which both reminiscence therapy and cognitive–behavioral therapy were compared with no treatment (Arean et al., 1993), the SMD for reminiscence therapy used in Scenario 2A was based on the same measure as that for cognitive–behavioral therapy. The SMD for reminiscence therapy used in Scenario 2B was based on the same measure as that for cognitive–behavioral therapy.

Of the seven estimated SMDs for comparing cognitive–behavioral therapy with no treatment in Scenario 2A, 86% were larger than the corresponding SMDs for the same comparison in Scenario 2B to statistically significant levels (see Table 2). The differences between Scenarios 2A and 2B embodied Nugent's (2012, 2013) measurement condition likely to lead to differing meta-analytic results.

Measurement Scenarios 3A and 3B. The measures that the estimated SMDs were based on, and symbolic representations of the estimated SMDs in these scenarios,

are shown in Table 1. These estimates incorporated all measures by using averages of SMDs based on scores from measures that appeared to meet CE and ER as indicated by results of the Hedges and Olkin (1985) tests. For comparisons of cognitive–behavioral therapy with no treatment, we created these measurement scenarios as follows:

- If there were two estimated SMDs for a cognitive–behavioral therapy/no treatment comparison based on scores from two different measures, and if these two did not differ to a statistically significant degree, we used the average of the two in Scenarios 3A and 3B.
- If there were two estimated SMDs for a cognitive–behavioral therapy/no treatment comparison based on scores from two different measures, and if these two did differ statistically, we used the larger of the two in Scenario 3A, and we used the smaller in Scenario 3B.

We used a slightly more complicated procedure (illustrated in Figure 2) if there were three estimated SMDs for a comparison of cognitive–behavioral therapy with no treatment:

- If one of the SMDs for a cognitive–behavioral therapy/no-treatment comparison differed from the other two to a statistically significant degree and the other two did not, the other two were averaged. We used whichever was larger—the single SMD or the average of the two—in Scenario 3A and used the smaller in Scenario 3B.
- To maintain consistency in the measures used for the cognitive bibliotherapy/no-treatment comparison in the Rokke, Tomhave, and Jocic (1999) study, the SMDs in Scenarios 3A and 3B were based on the same measures that the SMDs for the behavioral bibliotherapy/no-treatment comparison were based on.

For comparisons of reminiscence therapy with no treatment:

- For any study in which only reminiscence therapy was compared with no treatment, if there was a single estimated SMD based on a single measure, we used that SMD in Scenarios 3A and 3B.
- For the study in which both reminiscence therapy and cognitive–behavioral therapy were compared with no treatment (Arean, et al., 1993),
    - the SMD for reminiscence therapy used in Scenario 3A was based on the same single measure as the SMD for cognitive–behavioral therapy in 3A, and

Table 2
Results of Hedges and Olkin (1985) Tests of Construct Equivalence and Equal Reliabilities

Study and Treatment	Results of Tests of Homogeneity of SMDs Based on Scores From Different Measures
Arean et al. (1993)—CBT vs. ntx	**Omnibus test, $\chi^2(2) = 107.6$, $p < .05$** SMDs based on HRS and BDI differ, $\chi^2(1) = 87.5$, $p < .05$; SMDs based on HRS and GDS differ, $\chi^2(1) = 81.1$, $p < .05$; SMDs based on BDI and GDS do not differ, $\chi^2(1) = .006$, $p > .05$; *SMD based on HRS differed from mean SMD based on GDS and BDI, $\chi^2(1) = 85.8$, $p < .05$.*
Arean et al. (1993)—RT vs. ntx	**Omnibus test, $\chi^2(2) = 3.42$, $p > .05$**
Floyd et al. (2004)—CBT vs. ntx	**Omnibus test, $\chi^2(1) = 2.99$, $p > .05$**
Mastel-Smith et al. (2007)—RT vs. ntx	Only a single estimated SMD based on scores from BSI-D.
Rokke et al. (1999)—CBT vs. ntx	**Omnibus test, $\chi^2(2) = 9.44$, $p < .05$** SMDs based on BDI and HRS differ, $\chi^2(1) = 6.69$, $p < .05$; SMDs based on GDS and BDI differ, $\chi^2(1) = 4.46$, $p < .05$; SMDs based on GDS and HRS do not differ, $\chi^2(1) = .02$, $p > .05$; *SMD based on BDI differed from mean SMD based on GDS and HRS, $\chi^2(1) = 41.4$, $p < .05$.*
Rokke et al. (1999)—CBT vs. ntx	**Omnibus test, $\chi^2(2) = 9.86$, $p < .05$** SMDs based on GDS and HRS differ, $\chi^2(1) = 9.86$, $p < .05$; SMDs based on BDI and HRS do not differ, $\chi^2(1) = 2.73$, $p > .05$; SMDs based on GDS and BDI do not differ, $\chi^2(1) = 2.24$, $p > .05$; *SMD based on BDI did not differ from mean SMD based on GDS and HRS, $\chi^2(1) = .007$, $p > .05$.*

Scogin et al. (1987)—CBT vs. ntx	**Omnibus test, $\chi^2(2) = 9.31$, $p < .05$** SMDs based on HRS and GDS differ, $\chi^2(1) = 5.90$, $p < .05$; SMDs based on HRS and BDI differ, $\chi^2(1) = 8.75$, $p < .05$; SMDs based on GDS and BDI do not differ, $\chi^2(1) = 0.24$, $p > .05$; *SMD based on HRS differed from mean SMD based on GDS and BDI, $\chi^2(1) = 6.88$, $p < .05$.*
Scogin et al. (1989)—CBT vs. ntx	**Omnibus test, $\chi^2(1) = 41.55$, $p < .05$** The two estimated SMDs based on different measures differed statistically.
Scogin et al. (1989)—CBT vs. ntx	**Omnibus test, $\chi^2(1) = 38.32$, $p < .05$** The two estimated SMDs based on different measures differed statistically.
Serrano et al. (2004)—RT vs. ntx	Only a single estimated SMD based on scores from CES-D.
Wang et al. (2005)—RT vs. ntx	Only a single estimated SMD based on scores from GDS.

Note. BDI = Beck Depression Inventory; BSI-D = Brief Symptom Inventory depression subscale; CBT = cognitive–behavioral therapy; CES-D = Center for Epidemiologic Studies Depression scale; GDS = Geriatric Depression Scale; HRS = Hamilton Rating Scale; ntx = no treatment; RT = reminiscence therapy.

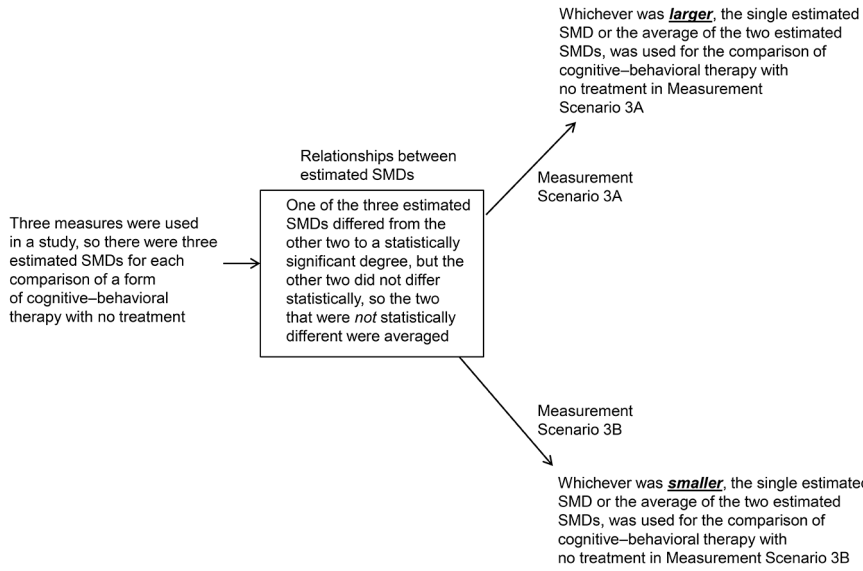

Figure 2. Illustration of the procedure used to create Measurement Scenarios 3A and 3B when there were three measures used in a study and there were three estimated standardized mean differences (SMDs) for each comparison of a form of cognitive–behavioral therapy with no treatment.

- the SMD used in Scenario 3B was the average of two SMDs based on the same two measures that the averaged SMD for cognitive–behavioral therapy was based on.

Of the seven estimated SMDs comparing cognitive–behavioral therapy with no treatment in Scenario 3A, 71% were statistically larger than the corresponding SMDs in Scenario 3B (see Table 2). This condition exemplified Nugent's (2012, 2013) condition likely to lead to differing meta-analytic results between scenarios. Scenarios 3A and 3B were conceived as a generalization of 2A and 2B; the estimated SMDs in 3A and 3B included SMDs that were averages of SMDs based on scores from different measures that appeared to meet CE and ER.

Analysis Methods

We used random effects analysis methods (Borenstein et al., 2009) given the different forms of cognitive–behavioral therapy and reminiscence therapy involved in the studies included in Peng et al.'s (2009) meta-analysis. The weighted regression methods for multiple treatment studies described by Gleser and Olkin (2009) were used as well. The estimated mean SMDs for cognitive–behavioral therapy as com-

pared with no treatment, for reminiscence therapy as compared with no treatment, and the differences between these means, were estimated using the weighted regression model,

$$\widehat{\boldsymbol{\delta}_m} = \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{d}, \tag{12}$$

where $\widehat{\boldsymbol{\delta}_m}$ was the column vector of estimated mean SMDs for the treatment effects, $\boldsymbol{d}$ was the column vector of sample estimated SMDs for the different treatment comparisons, and $\boldsymbol{X}$ was the design matrix composed of dummy variables indicating treatment. The transpose of the design matrix was

$$\boldsymbol{X}^{\mathrm{T}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}, \tag{13}$$

where the ones in the first row in $\boldsymbol{X}^{\mathrm{T}}$ indicated direct comparison of cognitive–behavioral therapy with reminiscence therapy; the ones in the second row indicated comparison of cognitive–behavioral therapy with no treatment; and the ones in the third row indicated comparison of reminiscence therapy with no treatment. As noted earlier, we included this single direct comparison of cognitive–behavioral therapy with reminiscence therapy (Arean et al., 1993) in order to infuse into the analyses the information contained in the covariances between this study's estimated SMD and the two other estimated SMDs. (The results for this single direct comparison are not described here because they were based on a single study.) Finally, we computed the Q-statistics needed for estimating τ^2 (the variance of true SMDs) and I^2 (the ratio of true heterogeneity to total variance of observed SMDs; Borenstein, 2009) for these analyses from

$$Q = \boldsymbol{d}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{d} - \widehat{\boldsymbol{\delta}_m^{\mathrm{T}}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{X}\widehat{\boldsymbol{\delta}_m}, \tag{14}$$

where $\widehat{\boldsymbol{\delta}_m^{\mathrm{T}}}$ was the transpose of the column vector from Equation 12.

Practical Significance

We assessed the practical significance of the magnitude of estimated mean SMDs—and the practical significance of the estimated differences between them—by comparing the estimated mean SMDs to the mean (.47) and standard deviation (*SD* = .28) of Lipsey and Wilson's (1993) "refined" distribution of mean ESs from 156 meta-analyses of 9,400 treatment-effectiveness studies (see Lipsey & Wilson, 1993, Figure 7, p. 1198). Lipsey and Wilson stated that this refined distribution had better statistical properties than a more inclusive distribution of mean ESs from 302 meta-analyses (1993, Figure 1, p. 1192).

Meta-Analytic Research Questions and Hypotheses

In the current study, we conducted analyses to address three meta-analytic questions:

1. What was the efficacy of cognitive–behavioral therapy as compared with no treatment for decreasing depression in older adults, and what was the magnitude of this treatment effect?
2. What was the efficacy of reminiscence therapy as compared with no treatment for decreasing depression in older adults, and what was the magnitude of this treatment effect?
3. What was the efficacy of cognitive–behavioral therapy as compared with no treatment—relative to the efficacy of reminiscence therapy as compared with no treatment—for decreasing depression in older adults, and what was the difference between the magnitudes of these treatment effects?

The latter research question indirectly addressed comparison of the efficacy of cognitive–behavioral therapy with that of reminiscence therapy; if one treatment was more efficacious than the other, then the SMDs representing the efficacy of that treatment relative to no treatment should be larger in magnitude than those representing the efficacy of the other treatment relative to no treatment. Measurement Scenarios 2A and 2B, 3A and 3B, 2A and 3B, and 3A and 2B were not interchangeable; however, 2A and 3A and 2B and 3B were interchangeable.

Results

Results of Tests of CE and ER

The results of Hedges and Olkin (1985) tests, shown in Table 2, suggested that SMDs based on scores from different measures failed to meet CE and ER for six of the seven comparisons of cognitive–behavioral therapy with no treatment. In contrast, the scores from all the measures that the estimated SMDs for comparisons of reminiscence therapy with no treatment were based on appeared to meet CE and ER. For example, as shown in Table 2, the results of the omnibus test for differences between the estimated SMDs in the Arean et al. (1993) study for comparing cognitive–behavioral therapy with no treatment based on scores from the Hamilton Rating Scale, Geriatric Depression Scale, and Beck Depression Inventory was $\chi^2(2) = 107.6$, $p < .05$, suggesting that at least one of the estimated SMDs was not an estimate of a shared true SMD. Results of follow-up pair-wise comparisons indicated that the estimated SMD based on the Hamilton Rating Scale differed from the estimated SMD based on the Geriatric Depression Scale—$\chi^2(1) = 81.1$, $p < .05$—and the estimated SMD based on the Beck Depression Inventory—$\chi^2(1) = 87.5$, $p < .05$. The estimated SMDs based on the Geriatric Depression Scale and the Beck Depression Inventory did not differ beyond what is expected if both were estimates of a mutual

true SMD, $\chi^2(1) = 0.006$, $p > .05$. Finally, the estimated SMD based on the Hamilton Rating Scale differed statistically from the average of the estimated SMDs based on the Geriatric Depression Scale and the Beck Depression Inventory, $\chi^2(1) = 85.8$, $p < .05$.

Results for Different Measurement Scenarios

The results for the different measurement scenarios are illustrated in Figure 3.

Results for Measurement Scenario 1. The estimated variance of true SMDs (τ^2) was .074, $\tau = .272$, and the estimated I^2 was 38.7%. The mean SMD for cognitive–behavioral therapy as compared with no treatment was −.92, with $Z = -4.86$, $p < .05$, 95% CI [−1.29, −0.55]. The mean SMD for reminiscence therapy as compared with no treatment was −0.55, with $Z = -2.77$, $p < .05$, 95% CI [−0.94, −0.16].

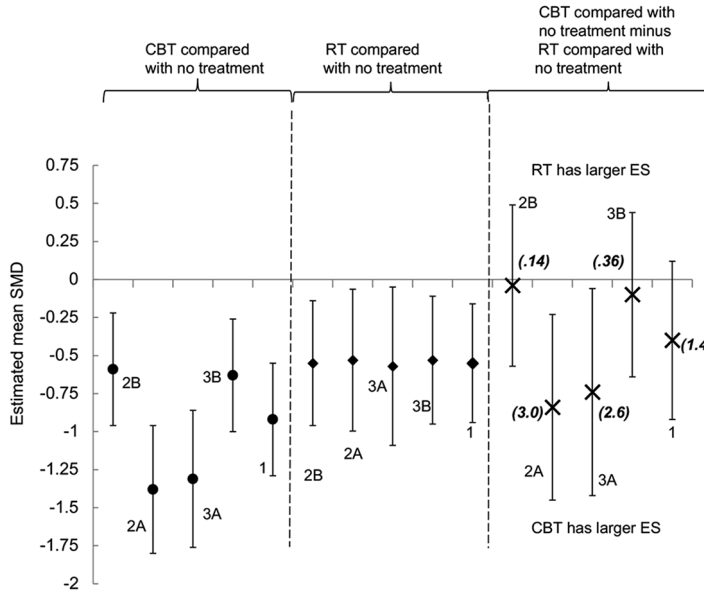

Figure 3. Results for estimated mean standardized mean differences (SMDs) for comparing cognitive–behavioral therapy (CBT) with no treatment (circular points); for comparing reminiscence therapy (RT) with no treatment (diamond-shaped points); and estimated differences between mean SMDs for comparing CBT with no treatment and mean SMDs for comparing RT with no treatment (X markers). The vertical bars show 95% confidence intervals. Numbers in ***italicized and bold type*** in parentheses show the number of *SD*s of the refined Lipsey and Wilson (1993) distribution that a difference between estimated mean SMDs represents. The brackets above each partitioned section of the graph indicate the different comparisons made. The alphanumeric labels in the figure indicate the measurement scenarios that the results are based on. A minus sign on an estimated mean SMD, or difference between estimated mean SMDs, indicates that the treatment reduces depression to a greater degree than no treatment, while a positive sign indicates the reverse. ES = effect size.

The difference between the mean SMD for comparing cognitive–behavioral therapy with no treatment and that for comparing reminiscence therapy with no treatment was −.40, with $Z = -1.50$, $p > .05$, 95% CI [−0.92, 0.12]. This statistically nonsignificant difference covered about 1.4 *SD*s in the Lipsey and Wilson (1993) refined distribution.

Results for Measurement Scenario 2A. The estimated τ^2 was .144, $\tau = .379$, and the estimated I^2 was 55.1%. The estimated mean SMD for cognitive–behavioral therapy as compared with no treatment was −1.38, with $Z = -6.43$, $p < .05$, 95% CI [−1.80, −0.96]. For reminiscence therapy as compared with no treatment, the estimated mean SMD was −0.53, with $Z = -2.24$, $p < .05$, 95% CI [−1.0, −0.07]. The difference between the mean SMD for cognitive–behavioral therapy compared with no treatment and that for reminiscence therapy compared with no treatment was −.84, with $Z = -2.71$, $p < .05$, 95% CI [−1.46, −0.23]. This statistically significant difference covered about 3.0 *SD*s in the Lipsey and Wilson (1993) refined distribution.

Results for Measurement Scenario 2B. The estimated τ^2 was .094, $\tau = .306$, and the estimated I^2 was 46.1%. The estimated mean SMD for cognitive–behavioral therapy as compared with no treatment was −.59, with $Z = -3.16$, $p < .05$, 95% CI [−0.96, −0.23]. For reminiscence therapy as compared with no treatment, the estimated mean SMD was −0.55, with $Z = -2.65$, $p < .05$, 95% CI [−0.96, −0.14]. The estimated difference between the mean SMD for cognitive–behavioral therapy compared with no treatment and that for reminiscence therapy compared with no treatment was −.04, with $Z = 0.15$, $p > .05$, 95% CI [−0.57, 0.49]. This statistically nonsignificant difference covered only about .14 *SD*s in the Lipsey and Wilson (1993) refined distribution.

Results for Measurement Scenario 3A. The estimated τ^2 was .198, $\tau = 0.445$, and the estimated I^2 was 63.0%. The estimated mean SMD for cognitive–behavioral therapy as compared with no treatment was −1.31, with $Z = -5.67$, $p < .05$, 95% CI [−1.76, −0.86]. For reminiscence therapy as compared with no treatment, the estimated mean SMD was −0.57, with $Z = -2.13$, $p < .05$, 90% CI [−1.10, −0.05]. The difference between the estimated mean SMD for comparing cognitive–behavioral therapy with no treatment and that for comparing reminiscence therapy with no treatment (−.74) was statistically significant, $Z = -2.14$, $p < .05$, 95% CI [−1.42, −0.06]. This statistically significant difference covered about 2.6 *SD*s in the Lipsey and Wilson (1993) refined distribution.

Results for Measurement Scenario 3B. In this measurement scenario, we estimated τ^2 to be .103, $\tau = .32$; I^2 was 48.5%. The estimated mean SMD for cognitive–behavioral therapy as compared with no treatment was −.63, with $Z = -3.31$, $p < .05$, 95% CI [−1.01, −0.26]. For reminiscence therapy compared with no treatment, the estimated mean SMD was −.53, with $Z = -2.50$, $p < .05$, 95% CI [−0.95, −0.11]. The difference between the estimated mean SMD for comparing cognitive–behavioral

therapy with no treatment and that for comparing reminiscence therapy with no treatment (−.10) was statistically nonsignificant, $Z = -.36$, $p > .05$, 95% CI [−0.64, 0.44]. This statistically nonsignificant difference covered .36 *SD*s in the Lipsey and Wilson (1993) refined distribution.

Comparison of Results Between Measurement Scenarios

First, in all five measurement scenarios the meta-analytic results suggested cognitive–behavioral therapy as superior to no treatment for decreasing depression in older adults. In terms of magnitude, for the exact same group of comparisons of cognitive–behavioral therapy with no treatment, the magnitude of the estimated mean SMDs ranged from −.59 to −1.38. In terms of the Lipsey and Wilson (1993) refined distribution of mean SMDs, this range covered about 2.8 *SD*s. Referenced to the mean of the Lipsey and Wilson (1993) refined distribution, these mean SMDs ranged from +.43 *SD*s to +3.25 *SD*s above the mean ES.

The smallest differences between estimated mean SMDs for comparing cognitive–behavioral therapy with no treatment were between the interchangeable Scenarios 2A and 3A (−.07; 0.25 *SD* in the Lipsey & Wilson refined distribution), and between 2B and 3B (0.04; 0.14 *SD* in the Lipsey & Wilson distribution). In contrast, the largest differences were between the noninterchangeable scenarios 2A and 2B (−.79; 2.8 *SD* in the Lipsey & Wilson distribution); 2A and 3B (−.75; 2.7 *SD* in the Lipsey & Wilson distribution); 3A and 2B (−.72; 2.6 *SD* in the Lipsey & Wilson distribution); and 3A and 3B (−.68; 2.4 *SD* in the Lipsey & Wilson distribution). The small degree of overlap of the 95% CIs for the estimated mean SMDs for scenarios 2A and 2B, 3A and 2B, 2A and 3B, and 3A and 3B suggested that these differences might be statistically significant (Cumming, 2012), though we could not test this. These findings are consistent with our hypothesis and with Nugent's (2012, 2013) speculation that mean SMDs—for the same group of studies when based on scores from different measures violating CE and ER—could vary substantially.

Second, in all five measurement scenarios, the meta-analytic results suggested that reminiscence therapy was better than no treatment for reducing depression in older adults. For this exact same group of comparisons, the estimated mean SMDs ranged from −.53 to −.57. In terms of the Lipsey and Wilson (1993) refined distribution of mean SMDs, this range covered only about 0.14 *SD*s. Referenced to the mean of the Lipsey and Wilson refined distribution, these mean SMDs ranged from .21 to .36 *SD*s above the mean ES. The 95% CIs for these estimated mean SMDs all overlapped substantially (see Figure 1), suggesting that none differed to a statistically significant degree. These estimated mean SMDs were far less variable across the different measurement scenarios than those for the comparisons of cognitive–behavioral therapy with no treatment. This lower variability is consistent with Nugent's (2012, 2013) conjecture that SMDs based on scores from measures meeting CE and ER will not vary to a significant degree.

Two of the five comparisons of the estimated mean SMDs for comparing cognitive–behavioral therapy and no treatment with the estimated mean SMDs for comparing reminiscence therapy and no treatment—those for measurement scenarios 2A and 3A—suggested that cognitive–behavioral therapy may have statistically significant larger magnitude reductions in depression for older adults than reminiscence therapy. The differences between these mean SMDs covered 3.0 and 2.6 *SD*s, respectively, in the Lipsey and Wilson (1993) refined distribution. These differences were of statistical and practical significance.

Sensitivity Analyses

As mentioned earlier, we conducted secondary analyses using correlations between measures of .70 when estimating variances of estimated SMDs that were averages. We conducted analyses a second time to assess the sensitivity of the previous results, based on these correlations set at .90, to what these correlations were assumed to be. The results of these sensitivity analyses did not differ significantly from the previous results and are not reported here.

Discussion

In a real-data situation for a given set of studies, the results demonstrated the existence of plausible alternate measurement scenarios in which scores from different measures appeared to violate CE and ER; as a consequence, meta-analyses based on the different measurement scenarios produced results that differed to statistically and practically significant degrees. The results shown in Figure 1 reveal significant variability in estimated mean true SMDs for cognitive–behavioral therapy compared with no treatment; results also show significant variability in estimated differences between mean SMDs for comparing cognitive–behavioral therapy with no treatment and mean SMDs for comparing reminiscence therapy with no treatment between different measurement scenarios. The estimated mean true SMDs for comparing reminiscence therapy with no treatment were quite consistent and showed relatively low variability across measurement scenarios. The set of seven estimated SMDs for comparisons of cognitive–behavioral therapy with no treatment were *not* interchangeable between measurement scenarios, but all four estimated SMDs for the comparisons of reminiscence therapy with no treatment *were* interchangeable. Thus, the differences in variability of results appears to be associated with whether the scores from the different measures the estimated SMDs were based on met CE and ER and were interchangeable.

As noted earlier, Nugent (2012, 2013) had speculated that meta-analytic results were most likely to differ between measurement scenarios when large differences existed between the estimated SMDs for any study (*i*) based on scores from different measures as a consequence of violations of CE and ER, and when a substantial per-

centage of the studies in a meta-analysis exhibited these large differences. These circumstances held for the measurement scenarios in this study. Eighty-six percent of the estimated SMDs for comparisons of cognitive–behavioral therapy with no treatment exhibited such large differences between measurement scenarios, and these differences appeared to be associated with violations of CE and ER. Thus, the measurement dependence of the foregoing meta-analytic results appears to have been associated with the conditions Nugent (2012, 2013) hypothesized.

The measurement scenarios were eminently plausible. All the measures used in the studies from the Peng et al. (2009) meta-analysis were well established and commonly used measures of depression. The use of estimated SMDs based on single measures in Scenarios 2A and 2B, and in some cases in 3A and 3B, assumed that researchers chose to use only a single measure of depression. This would be a plausible methodological choice. For example, such a choice might be made when multiple measures of other constructs are going to be used in a planned study and use of a single measure of depression will reduce the likelihood of measurement fatigue. Similarly, the use of averages of estimated SMDs based on multiple measures used in Scenarios 3A and 3B presumed that researchers chose to use multiple measures of depression. Such a choice would also be reasonable. For example, the use of the Hamilton Rating Scale, a clinician rating scale, and concomitant use of either the Geriatric Depression Scale or the Beck Depression Inventory, both self-report scales, could readily be conceptualized as a multimethod measurement strategy.

These findings confirm Nugent's (2013, 2017) model-based simulation results. The results of the current study, as well as the prior research by Nugent (2012, 2013, 2017), extend the literature on the importance of measurement equivalence (e.g., Chen, 2008; Vandenberg & Lance, 2000). The results of the current study suggest that measurement equivalence not only concerns the integrity of results from individual studies in which, for example, the means of different groups are compared, but also the validity of meta-analytic results.

The foregoing considerations concern the existence of counterfactual measurement scenarios involving plausible measures *not* used by researchers in their studies and the possibility that meta-analyses of the same studies, but based on alternate measurement scenarios, might produce different results. The concern is that the results of meta-analyses may be measurement dependent unless the scores from all of the measures in a set that researchers might justifiably use meet the form of measurement equivalence defined by CE and ER. It should also be noted the results of the current study, as well as Nugent's (2013, 2017) previous research, are relevant for situations in which researchers use multiple measures of a construct—so there are multiple estimated ESs for a given treatment comparison—yet meta-analysts select only one, or a subset of the estimated ESs to include in a meta-analysis. Such practices may also lead to measurement-dependent results of meta-analyses.

The findings of the current study pertain to circumstances in which the measures used in studies in a meta-analysis have arbitrary metrics and fail to meet CE and ER. These problems are most likely to be a concern for social work, psychiatric, psychological, and behavioral meta-analyses of studies of treatments for mental health, substance abuse, and behavioral problems because measures with arbitrary metrics are frequently used in such studies (Blanton & Jaccard, 2006; Kazdin, 2006).

Along with Nugent's (2013, 2017) prior work, the results of the current study suggest that the we must use caution in interpreting the results of meta-analyses in which the ESs are based on scores from measures with arbitrary metrics that fail to meet the measurement equivalence defined by CE and ER. The meta-analyst or social work practitioner who is a consumer of meta-analytic results will need to carefully investigate the nature of the measures on which the ESs in the meta-analysis are based. When at least some of the measures that the ESs are based on have arbitrary metrics, and when they do not meet the form of measurement equivalence defined by CE and ER, the results of the meta-analysis may be valid only for the particular measurement scenario upon which the meta-analysis was based. Had the ESs in the included studies been based on scores from different but defensible measures, the results could have been quite different—perhaps with contradictory findings and implications.

The methodology of the current study might be considered as a form of sensitivity analysis of the results of a meta-analysis. As was done in this study, the ESs in a meta-analysis in which at least some of the studies used multiple measures of the same construct can be subjected to secondary analyses based on different measurement scenarios. The results may offer insight into the degree of measurement dependence for meta-analysis results.

Finally, this study is a single empirical demonstration of the measurement dependence of meta-analytic results. These findings do not address the frequency with which measurement-dependent meta-analytic results actually occur. Further research on this topic is needed.

Author Notes

William R. Nugent, PhD, is Associate Dean for Research in the College of Social Work, University of Tennessee-Knoxville.

Sukyung Yoon, MSW, is a PhD candidate in the College of Social Work, University of Tennessee-Knoxville.

Jayme Walters, MSW, is a PhD student in the College of Social Work, University of Tennessee-Knoxville.

Correspondence regarding this article should be directed to William R. Nugent, College of Social Work, Henson Hall, University of Tennessee-Knoxville, Knoxville, TN, 37996-3333, or via e-mail to wnugent@utk.edu

Acknowledgments

We would like to thank the anonymous reviewer who so thoroughly critiqued this article and made numerous suggestions to improve upon this work.

References

*Asterisks indicate studies in the Peng et al. (2009) meta-analysis that were sources of the estimated SMDs for the secondary analyses in the current study.

*Arean P., Perri M., Nezu, A., Schein, R., Christopher, F., & Joseph, T. (1993). Comparative effectiveness of social problem solving therapy and reminiscence therapy as treatments for depression in older adults. *Journal of Consulting & Clinical Psychology*, *61*, 1003–1010. https://doi.org/10.1037/0022-006X.61.6.1003

Banks, C. (2009). What is modeling and simulation? In J. Sokolowski & C. Banks (Eds.), *Principles of modeling and simulation: A multidisciplinary approach* (pp. 3–24). Hoboken, NJ: Wiley.

Beck, A., Steer, R., & Garbin, M. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*, 77–100. https://doi.org/10.1016/0272-7358(88)90050-5

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *6* (61), 27–41. https://doi.org/10.1037/1082-989X.8.4.406

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. West Sussex, UK: John Wiley & Sons.

Chen, F. (2008). What happens if we compare chopsticks and forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*, 1005–1018. http://dx.doi.org/10.1037/e514412014-064

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Derogatis, L., & Melisarotos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine*, *13*, 595–605. https://doi.org/10.1017/S0033291700048017

Dorans, N., Pommerich, M., & Holland, P. (Eds). (2010). Linking and aligning scores and scales. New York, NY: Springer.

*Floyd, M., Scogin, F., McKendree-Smith, N., Floyd, D., & Rokke, P. (2004). Cognitive therapy for depression: A comparison of individual psychotherapy and bibliotherapy for depressed older adults. *Behavior Modification*, *28*, 297–318. https://doi.org/10.1177/0145445503259284

Gilovich, T. (1991). *How we know what isn't so*. New York, NY: The Free Press.

Gleser, L., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). New York, NY: The Russell Sage Foundation.

Grissom, R., & Kim, J. (2005). *Effect sizes for research: A broad practical approach*. New York, NY: Psychology Press.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*, 56–62. https://doi.org/10.1136/jnnp.23.1.56

Harrison, J., Carroll, G., & Carly, K. (2007). Simulation modelling in organizational and management research. *Academy of Management Review*, *32*, 1229–1245. https://doi.org/10.5465/AMR.2007.26586485

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.

Higgins, J., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0). The Cochrane Collaboration. Retrieved from http://handbook.cochrane.org

Joreskog, K. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. https://doi.org/10.1007/BF02291393

Kazdin, A. (2006). Arbitrary metrics. *American Psychologist*, *61*, 42–49. https://doi.org/10.1037/0003-066X.61.1.42

Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209. https://doi.org/10.1037/0003-066X.48.12.1181

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

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

*Mastel-Smith, B., McFarlane, J., Sierpina, M., Malecha, A., & Haile, B. (2007). Improving depressive symptoms in community-dwelling older adults: A psychosocial intervention using life review and writing. *Journal of Gerontological Nursing*, *33*, 13–19. https://www.ncbi.nlm.nih.gov/pubmed/17511331

Mitchell, A. (2010). Overview of depression scales and tools. In A. Mitchell & J. Coyne (Eds.), *Screening for depression in clinical practice: An evidence-based guide* (pp. 29–56). New York, NY: Oxford University Press.

Murad, M., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *Evidence-Based Medicine*, *21*, 125–127. https://doi.org/10.1136/ebmed-2016-110401

Nugent, W. (2012). The interchangeability of scores from different measures and meta-analytic effect size comparability. *Journal of the Society for Social Work and Research*, *3*, 213–232. https://doi.org/10.5243/jsswr.2012.14

Nugent, W. (2013). The interchangeability of scores from different measures and meta-analytic effect size comparability II: A simulation study. *Journal of the Society for Social Work and Research*, *4*, 76–98. https://doi.org/10.5243/jsswr.2013.6

Nugent, W. (2017). Variability in the results of meta-analysis as a function of comparing effect sizes based on scores from noncomparable measures: A simulation study. *Educational and Psychological Measurement*, *77*, 449–470. https://doi.org/10.1177/0013164416654517

Peng, X., Huang, C., Chen, L., & Lu, Z. (2009). Cognitive behavioural therapy and reminiscence techniques for the treatment of depression in the elderly: A systematic review. *The Journal of International Medical Research*, *37*, 975–982. https://doi.org/10.1177/147323000903700401

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401. http://dx.doi.org/10.1177/014662167700100306

*Rokke, P., Tomhave, J., & Jocic, Z. (1999). The role of client choice and target selection in self-management therapy for depression in older adults. *Psychology and Aging*, *14*, 155–169. https://doi.org/10.1037/0882-7974.14.1.155

Rubin, A., & Bellamy, J. (2012). *Practitioner's guide to using research for evidence-based practice* (2nd ed.). Hoboken, NJ: John Wiley.

*Scogin, F., Hamblin, D., & Beutler, L. (1987). Bibliotherapy of depressed older adults: A self-help alternative. *The Gerontologist*, *2*, 383–387. https://doi.org/10.1093/geront/27.3.383

*Scogin, F., Jamison, C., & Gochneaur, K. (1989). Comparative efficacy of cognitive and behavioral bibliotherapy for mildly and moderately depressed older adults. *Journal of Consulting & Clinical Psychology*, *57*, 403–407. https://doi.org/10.1037/0022-006X.57.3.403

*Serrano, J., Latorre, J., Gatz, M., & Montanes, J. (2004). Life review therapy using autobiographical retrieval practice for older adults with depressive symptomatology. *Psychology and Aging*, *19*, 272–277. https://doi.org/10.1037/0882-7974.19.2.272

Sheikh, J., & Yesavage, J. (1986). Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. *Clinical Gerontologist: The Journal of Aging and Mental Health*, *5*, 165–173.

Solow, D. (2002). *How to read and do proofs* (3rd ed.). New York, NY: John Wiley & Sons.

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. https://doi.org/10.1177/109442810031002

*Wang, J., Hsu, Y., & Cheng, S. (2005). The effects of reminiscence in promoting mental health in Taiwanese elderly. *International Journal of Nursing Studies*, *42*, 31–36. https://doi.org/10.1016/j.ijnurstu.2004.05.010

White, H. (2009). Scientific communication and literature retrieval. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*, (2nd ed., pp. 51–72). New York, NY: Russell Sage Foundation.

Manuscript submitted: April 17, 2017
First revision submitted: July 17, 2017
Second revision submitted: September 4, 2017
Third revision submitted: November 1, 2017
Accepted: November 27, 2017
Electronically published: July 27, 2018