Personality and Social Psychology Bulletin
37(4) 570–583

DOI: 10.1177/0146167211400619
http://pspb.sagepub.com
SAGE

Sexy But Often Unreliable: The Impact of Unreliability on the Replicability of Experimental Findings With Implicit Measures

Etienne P. LeBel[1] and Sampo V. Paunonen[1]

Abstract

Implicit measures have contributed to important insights in almost every area of psychology. However, various issues and challenges remain concerning their use, one of which is their considerable variation in reliability, with many implicit measures having questionable reliability. The goal of the present investigation was to examine an overlooked consequence of this liability with respect to replication, when such implicit measures are used as dependent variables in experimental studies. Using a Monte Carlo simulation, the authors demonstrate that a higher level of unreliability in such dependent variables is associated with substantially lower levels of replicability. The results imply that this overlooked consequence can have far-reaching repercussions for the development of a cumulative science. The authors recommend the routine assessment and reporting of the reliability of implicit measures and also urge the improvement of implicit measures with low reliability.

Keywords

implicit measures, replicability, replication, reliability, measurement error

Received May 21, 2010; revision accepted November 13, 2010

> Science is concerned with repeatable experiments. If data obtained from experiments are influenced by random errors of measurement, the results are not exactly repeatable. Thus, science is limited by the reliability of measuring instruments and the reliability with which scientists use them.
>
> Nunnally (1982, p. 1589)

A significant achievement of psychological science in recent times has been the development and refinement of a new class of measurement procedures for assessing mental representations and processes. This new class of measurement procedures—collectively referred to as *implicit measures*—derives its name from the fact that psychological constructs are measured relatively indirectly compared to other, more explicit measures. For example, an explicit measure of racial attitudes might ask respondents to report directly about their racial beliefs using a verbal questionnaire. In contrast, an implicit measure of such attitudes might require respondents to perform a speeded categorization task involving a mix of racial and evaluative stimuli (for reviews, see Fazio & Olson, 2003; Wittenbrink & Schwarz, 2007).

Implicit measures hold great promise for our science because they have the potential to provide a fuller understanding of the cognitive and affective mechanisms underlying psychological phenomena than do explicit measures. This promise lies in the fact that implicit measures (a) do not require verbal self-report of the construct of interest and hence may be free of social desirability contamination and other biases and (b) do not require introspection and thus may tap into mental states that are beyond individuals' self-awareness (but see Gawronski, LeBel, & Peters, 2007).

Despite the promise of implicit measures, some challenges have arisen concerning their use. One of the most salient of these is that many of them show lower levels of reliability compared to explicit measures (Fazio & Olson, 2003; Greenwald & Banaji, 1995). Lower levels of reliability imply higher amounts of random measurement error contaminating the measure's scores. Although it is generally known that higher levels of random error in a dependent variable decrease observed effect sizes (Schmidt & Hunter, 1996), noise in these measures also introduces nonrepeatability in detecting

[1]University of Western Ontario, London, ON, Canada

Corresponding Author:
Etienne P. LeBel, University of Western Ontario, Department of Psychology, Social Science Centre, London, Ontario N6A 5C2, Canada
Email: elebel@uwo.ca

Table 1. Examples of Implicit Measures Used in Different Areas of Psychology

Area of psychology	Implicit measure	Authors
Traditional		
Personality	Go/No-go Association Task	Borkenau & Mauer, 2007
Developmental	Perceptual picture identification task	Perez, Peynircio, & Blaxton, 1998
Social	Semantic priming task	Wittenbrink, Judd, & Park, 1997
Educational	Repetition priming task	Woltz & Shute, 1993
Neuropsychology	Word stem completion task	Schott et al., 2005
Cognitive	Emotional Stroop task	Algom, Chajut, & Shlomo, 2004
Industrial/ organizational	Word stem completion task	Johnson & Steinman, 2009
Clinical	Extrinsic Affective Simon Task	De Raedt, Schacht, Franck, & De Houwer, 2006
Nontraditional		
Political	Implicit Association Test	Galdi, Arcuri, & Gawronski, 2008
Health	Affective priming task	Papies, Stroebe, & Aarts, 2009
Consumer	Picture-picture naming task	Spruyt, Hermans, De Houwer, Vandekerckhove, & Eelen, 2007
Positive	Name-letter task	Schimmack & Diener, 2003
Forensic	Implicit Association Test	Gray, McCulloch, Smith, Morris, & Snowden, 2003

experimental effects. This issue of replicability seems to be a growing concern with respect to findings involving implicit measures, an observation based on our own personal experience, on lively discussions we have witnessed among some researchers at conferences, and on the published lamentations of some social psychologists (e.g., Blascovich et al., 2002).

The primary goal of the present article was to investigate the reliability issue with regard to implicit measures from a novel perspective—by connecting reliability to replicability. Although the lower reliability of many implicit measures has been acknowledged, its effect on experimental replication has been overlooked. Moreover, no systematic investigation has been executed to examine the degree of impact of such unreliability on replicability. Hence, the present study sought to determine the consequences of this unreliability for the replicability of experimental findings when such implicit measures are used as dependent variables.

As articulated in our opening quote, random measurement error, which contributes to the unreliability of measures, can prevent an experiment from being exactly repeatable. What is not known, however, and what we sought to quantify in this investigation, is the degree to which such random measurement error affects the replicability of real experimental findings. We used Monte Carlo simulation methodology to examine the effect of different levels of reliability on the replicability of experimental findings in the context of implicit measures. Although our conclusions apply equally well to any measure that has questionable reliability (implicit or otherwise), we position our investigation with respect to implicit measures given that such measure typically have lower reliability than do explicit measures.

Growing Popularity of Implicit Measures

Evidence of the enthusiasm for implicit measures is apparent in their widespread application across disparate areas of psychology. As we show in Table 1, implicit measures have been used in all traditional areas including personality psychology, social psychology, developmental psychology, educational psychology, neuropsychology, cognitive psychology, industrial/organizational psychology, and clinical psychology. In addition, implicit measures have also become popular in less traditional areas such as political psychology, health psychology, consumer psychology, positive psychology, and forensic psychology.

The use of implicit measures has led to many important findings. For example, in the context of understanding the cognitive precursors of depression, De Raedt, Schacht, Franck, and De Houwer (2006) found that, paradoxically, clinically depressed patients exhibited more positive self-associations as compared to nondepressed controls (as assessed using the Extrinsic Affective Simon Task; De Houwer, 2003). Inconsistent with traditional cognitive theories of depression, this finding led to the important insight that self-schemas of depressed individuals do contain some positive content. What differentiates depressed from nondepressed individuals, however, is how that positive content is processed and organized.

In the political arena, Galdi, Arcuri, and Gawronski (2008) found that automatic associations to an important political issue, assessed using the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), predicted future political positions for undecided individuals. For decided individuals, however, conscious political beliefs about the issue predicted future political position. Another example of the utility of implicit measures is in the study of prejudice. Dovidio, Kawakami, and Gaertner (2002) found that automatic racial associations, assessed using an affective priming task, predicted nonverbal friendliness in actual interracial interactions, whereas explicit racial attitudes predicted self-perceived verbal friendliness. These empirical examples reveal just a few of the important insights provided by implicit measures into various psychological phenomena.

Table 2. Reported Reliability Estimates of Different Implicit Measures

Implicit measure	Reported reliability	Reliability cited in
Affective priming task	$r_{sh} < .20$	Fazio, Jackson, Dunton, & Williams, 1995; Olson & Fazio, 2003
Go/No-go Association Task	$r_{sh} < .20$	Nosek & Banaji, 2001
Extrinsic Affective Simon Task	$\alpha < .30$	De Houwer & De Bruycker, 2007; Teige, Schnabel, Banse, & Asendorpf, 2004
Word stem completion task	$r_{sh} < .30$	Buchner & Wippich, 2000
Dot probe task	$\alpha < .30$	Schmukle, 2005
Lexical decision task	$\alpha < .50$	Borkenau, Paelecke, & Yu, 2009
Name letter task	$\alpha = .35–.50$	LeBel & Gawronski, 2009
Implicit Association Test	$\alpha = .60–.90$	Gawronski, Deutsch, & Banse, in press
Affect Misattribution Procedure	$\alpha = .70–.90$	Payne, Cheng, Govorun, & Stewart, 2005

Note: r_{sh} = split-half correlation; α = Cronbach's (1951) coefficient alpha.

Psychometric Issues With Implicit Measures

Despite the widespread use and apparent utility of implicit measures, various issues and challenges have arisen concerning these instruments. For example, researchers have debated the precise meaning and interpretation of some of the measures' scores (Arkes & Tetlock, 2004; Blanton & Jaccard, 2006). Vigorous disagreements concerning the theoretical conceptualization of the core psychological constructs assessed by some of the implicit measures have also emerged. In particular, the question of whether implicit and explicit constructs are (Greenwald, Nosek, & Banaji, 2003) or are not (Greenwald et al., 1998) the same has created much confusion (Blanton, Jaccard, Gonzales, & Christie, 2006).

A related challenge to the construct definition issue mentioned above has been the question of whether to validate variables assessed with implicit measures against corresponding variables assessed with explicit measures or against other variables also assessed with implicit measures. For example, Bosson, Swann, and Pennebaker (2000) found, in the context of evaluating implicit self-esteem measures, that none of the implicit measures they examined predicted explicit measures typically used as self-esteem criteria. This led Bosson et al. to the awkward conclusion that researchers should focus on "indirect or nonconscious criterion measures" when validating implicit measures, at least for those measures related to self-esteem (p. 641).

A somewhat less examined issue with implicit measures concerns their relatively poor psychometric properties in general (Fazio & Olson, 2003). For example, it has been observed that implicit measures generally suffer from a lack of convergent validity even with other implicit measures (Bosson et al., 2000; Wittenbrink, 2007). There is also the issue, already mentioned, that implicit measures show considerable variation in reliability, in terms of both internal consistency and test–retest correlations (Fazio & Olson, 2003). In line with this concern, Gawronski, Deutsch, and Banse (in press) provide a summary of published reliability estimates for various implicit measures ranging from .00 to .90, with the typical reliability of many of these measures being at a level that is clearly unsatisfactory from a psychometric point of view. And there is the problem we and some colleagues have sometimes experienced related to the difficulty in replicating experimental findings. Because the sometimes low reliability of implicit measures is of particular concern to the present study, we discuss that issue in more detail below.

The Reliability of Implicit Measures

In their seminal review, Fazio and Olson (2003) observed that various implicit measures suffered from "rather low reliability" (p. 311), which poses certain theoretical challenges. They reported that many implicit measures typically yield reliability estimates that range from abysmally low (Bosson et al., 2000) to moderate (Kawakami & Dovidio, 2001). Nosek and Banaji (2001) stated very bluntly that the "reliability of implicit measures is far below typical standards for their explicit counterparts" (p. 660). Gawronski et al. (2007) similarly noted the fact that implicit measures have shown relatively low estimates of internal consistency, citing several studies (e.g., Cunningham, Preacher, & Banaji, 2001; Gawronski, 2002). More recently, Uhlmann, Pizarro, and Bloom (2008) reported that the "reliability of standard priming measures averages about .30 across studies" (p. 307; also see Wittenbrink, 2007). In the context of the IAT, Blanton and Jaccard (2008) mentioned that the challenge of creating reliable indices for implicit measures may be greater than that for explicit measures. Finally, from the cognitive psychology literature, both Meier and Perrig (2000) and Buchner and Wippich (2000) have demonstrated and bemoaned the fact that the reliability of most implicit memory measures (e.g., word stem completion task, perceptual clarification task) is typically much lower than the reliability of explicit memory measures (e.g., recall and recognition measures).

Specific examples of the relatively low reliability of implicit measures can be cited, and we have listed some of these in Table 2. As can be seen in the table, low reliability values (estimated using split-half correlations or coefficient alphas) have

been reported for the standard affective priming task, the Go/No-go Association Task (Nosek & Banaji, 2001), the Extrinsic Affective Simon Task (De Houwer, 2003), the word stem completion task, the dot probe task, the lexical decision task, and the name-letter task (Nuttin, 1985), among others. But the situation is not intractable. The IAT has generally exhibited relatively high levels of internal consistency compared to other implicit measures. With the new IAT scoring algorithm (Greenwald et al., 2003), IAT scores typically yield internal consistencies in the range of .60 to .90. Another implicit measure that has typically exhibited relatively high levels of reliability is the Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005), although the number of published studies reporting reliability estimates for the AMP is relatively small.

Unfortunately, our brief review offers only a rough indication of the typical reliability of implicit measures because, as Gawronski et al. (in press) mention, reliability estimates of such measures often go unreported in published articles (also see Schmukle, 2005). Nonetheless, based on general claims made by independent researchers, and on particular examples of reliability found in the literature, it is clear that implicit measures often produce scores with significantly lower internal consistency and test–retest reliability than do corresponding explicit measures.

An important question that follows from our review above is whether—and, if so, to what extent—low reliability of implicit measures affects the replicability of experimental findings when using such measures as dependent variables. An actual example from the literature may serve to clarify the relevance of this question. A widely cited finding in self-esteem research is that individuals who had relatively positive self-beliefs (as measured with the Rosenberg Self-Esteem Scale) but low levels of implicit self-esteem (as measured with the self-esteem IAT) showed the highest levels of self-reported narcissism (Jordan, Spencer, Zanna, Hoshino-Browne, & Correll, 2003). A common interpretation of this finding is that, for individuals with positive self-beliefs, their negative implicit self-feelings may at times seep into their consciousness and be experienced as "nagging doubts," leading to defensive behavior, such as narcissism or in-group bias. In subsequent studies, however, this effect proved difficult to replicate (for a review, see Bosson et al., 2008). For example, Campbell, Bosson, Goheen, Lakey, and Kernis (2007) failed to replicate the pattern even though exactly the same measures were used as in Jordan et al.'s (2003) study. Several unpublished datasets also failed to replicate the effect (Bosson et al., 2000; Gregg & Sedikides, 2008; Rosenthal, 2007; Zeigler-Hill, 2007; all cited by Bosson et al., 2008), although Zeigler-Hill (2006) did replicate the pattern using different self-esteem IAT stimuli. We acknowledge that many factors can contribute to the failure to replicate an experimental effect (e.g., sampling differences, methodological or measurement parameter differences, Type I errors). In the present context, however, we argue that a salient, but seldom recognized, factor underlying such failures at replication may be the low reliability of the measures involved.

The Link Between Reliability and Replicability

Sutcliffe (1958) showed algebraically that increasing amounts of random error in a dependent measure decreases the statistical power of the *F* test to detect differences among group means on that measure. That is, lower levels of reliability are associated with decreasing probabilities of detecting a statistically significant effect, given one exists in the population. In a similar vein, Hopkins and Hopkins (1979) and Rogers and Hopkins (1988) both showed that low reliability of dependent variable scores decreases statistical power, and both provided formulas for adjusting Cohen's *f* observed effect sizes to estimate statistical power given varying levels of reliability. In addition, Kopriva and Shaw (1991) developed tables to estimate statistical power for ANOVA designs depending on reliability, effect size, and sample size (also see Williams & Zimmerman, 1989). Given that the probability of replication is simply a special case of statistical power (i.e., probability of replication is the probability of detecting a statistically significant effect given one exists in the population *and* that the effect has already been found in at least one sample), it follows that decreasing levels of reliability should be associated with reduced likelihood of replication (also see Baugh, 2002; Schmidt & Hunter, 1996). However, the magnitude of the effect of unreliability on replicability is unknown, and shedding light on this issue was one of the primary goals of the present research.

Although the notion that unreliability decreases statistical power is common knowledge in many areas of psychology, the related, but distinct, notion of unreliability decreasing replicability is not well established. Hence, the primary focus and contribution of our investigation is to argue, and to demonstrate empirically, that score unreliability can have broader ranging effects than simply decreasing statistical power. It can also decrease the probability of replicating an experimental effect, which has much more potent metascientific implications in terms of the development of a cumulative science.

To investigate the degree of impact of unreliability on the probability of replicating an experimental effect, we conducted a Monte Carlo simulation as described below. That method is a general procedure that simulates real-life systems that may be too complex or costly to explore directly (Robie & Komar, 2007). The method relies on the repeated sampling of computer-generated data points from predefined populations, having characteristics that replicate known parameters found in typical research situations. After the data are generated, various models of data manipulation can be introduced, and the impact of the factors built into the models can be estimated. In the present case, for example, the Monte Carlo

simulation method provides an ideal way to manipulate precisely varying degrees of reliability in a dependent variable and to observe the resultant effects on the probability of replication.

Our Monte Carlo study has the potential to spell out in very concrete terms some of the abstract issues concerning random measurement error and replicability. As mentioned earlier, we connect random measurement error to replicability specifically in the context of implicit measures. This is because of considerable variation in the reliability of such measures and their known problems with replication of experimental outcomes.

Method

Design

A Monte Carlo simulation was designed to examine the impact of unreliability in a dependent variable on the replicability of results for a simple two-group between-subjects test of means. In this study, we used a three-way factorial design ($10 \times 5 \times 3$) to vary reliability of the dependent variable, group sample size, and population effect size, in evaluating their effects on the likelihood of concluding that the two groups differ on the dependent variable. Those three independent variables represented 10 levels of measurement reliability (ρ_{xx}), essentially covering the entire range ($\rho_{xx} = 0$ to $\rho_{xx} = 1$ in .10 increments), five levels of sample size (N) typical of research in this area ($N = 10$ to $N = 50$ per group in increments of 10), and three levels of population effect size (d) corresponding to Cohen's (1988) small, medium, and large effects ($d = .2$, $d = .5$, and $d = .8$). The medium effect size we chose is consistent with Lipsey and Wilson's (1993) finding that the mean d of more than 300 meta-analyses for a wide variety of psychological treatments was about .50.

Data Generation

The dependent variable data, simulating scores on a hypothetical implicit measure, were generated by repeatedly drawing two samples of N observations from prespecified populations having the reliabilities ρ_{xx} and effect sizes d as specified above. Thus, for our two-group (between-subjects) simulation, this involved drawing a control group and a treatment group sample, both of size N, from their own normally distributed populations, each having a dependent variable mean of μ and a standard deviation of 1. In the case of the control group, $\mu = 0$, whereas for the treatment group, $\mu = .2$, .5, or .8. Importantly, however, the data were configured to reflect varying levels of random measurement error contamination in the dependent variable, as described below, before performing an independent groups t test on the two drawn samples.

To add random error to our dependent variables, we followed the logic from past simulation studies (Charles, 2005; Jaccard & Wan, 1995) and manipulated the ratio of true score variance to total observed variance in our simulated data. This procedure is based on classical test theory, which conceives of an observed score on a variable (X) as the sum of the true score on the attribute in question (T) and a random error score (E; Lord & Novick, 1968).[1] This leads to the well-known variance decomposition formula,

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2,$$

where σ_X^2, σ_T^2, and σ_E^2 represent the variance of the observed scores, true scores, and random error scores, respectively.[2] From this equation, it can be shown that,

$$\rho_{xx} = \sigma_T^2 / \sigma_X^2,$$

where ρ_{xx} is the reliability of the measure X, which reflects the ratio of true score variance to total observed variance. Thus, to produce a reliability for variable X of .90, one would set the ratio of the true score variance to error variance at 9 to 1.

To generate observed dependent variable scores X with a given amount of error in our simulation, we first drew a sample of N true scores from one of our prespecified populations and a sample of N error scores from another (independent) prespecified population, such that the standard deviations of the two populations were configured to the desired proportion. The true and error scores were then summed to yield observed scores with the requisite amount of error variance for the reliability level under investigation. For example, to achieve a .90 level of reliability for a dependent variable, we drew a sample of N true scores from a normal distribution with a mean of 0 and a standard deviation of 1 and then drew a sample of N error scores from a separate distribution with a mean of 0 and a standard deviation of roughly .33. As reliability is the ratio of true score variance to observed variance, we have $\rho_{xx} = (1.0^2)/(1.0^2 + .33^2) = .90$. These true score and error score vectors were then summed and treated as observed control group scores on the dependent variable in subsequent analyses.[3] The treatment group's observed scores were generated in exactly the same way, except that the true scores were sampled from populations with nonzero means (i.e., $\mu = .2$, $\mu = .5$, or $\mu = .8$). Table 3 summarizes, for an effect size of .5, our simulated population parameters, varying in levels of reliability and sample sizes for both the treatment and control groups.

Procedure

For each of the 150 unique conditions in our $10 \times 5 \times 3$ factorial design, 5,000 iterations were executed. For each iteration, a t test was used to compute the statistical significance of the mean difference between the treatment and control groups on the simulated dependent variable for the conditions being evaluated. The proportion of statistically significant effects out of 5,000 was then tabulated, yielding the

Table 3. Summary of Population Parameters Evaluated for Both the Treatment and Control Groups at 10 Levels of Reliability and a Population Effect Size of .5

Reliability levels									
.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
Treatment groups (X = T + E)									
T = N(.5, 1)	T = N(.5, 1)	T = N(.5, 1)	T = N(.5, 1)	*T* = N(.5, 1)	T = N(.5, 1)	T = N(.5, 1)	T = N(.5, 1)	T = N(.5, 1)	T = N(.5, 1)
E = N(0, 3)	E = N(0, 2)	E = N(0, 1.528)	E = N(0, 1.225)	E = N(0, 1)	E = N(0, .816)	E = N(0, .655)	E = N(0, .5)	E = N(0, .33)	E = N(0, 0)
Control groups (X = T + E)									
T = N(0, 1)	T = N(0, 1)	T = N(0, 1)	T = N(0, 1)	T = N(0, 1)	T = N(0, 1)	T = N(0, 1)	T = N(0, 1)	T = N(0, 1)	T = N(0, 1)
E = N(0, 3)	E = N(0, 2)	E = N(0, 1.528)	E = N(0, 1.225)	E = N(0, 1)	E = N(0, .816)	E = N(0, .655)	E = N(0, .5)	E = N(0, .33)	E = N(0, 0)

Note: X = observed scores; T = true scores; E = random error scores. Random observations were drawn from normal distributions with a mean and standard deviation N(μ, s), as specified in each condition, at five different sample sizes. Observed scores were computed as the sum of true and random error scores.

observed replicability for that condition. Although, conceptually, replication is the probability of replicating an effect that has already been found in one study, and is thus a conditional probability, in the context of simulation studies (e.g., Cumming, 2008), each sampling is independent from each other, and so the order in which statistically significant samples are observed is irrelevant. Hence, in a simulation study context, replication is isomorphic with statistical power.

Results

Type I Error and Statistical Power Checks

Preliminary analyses were executed to verify the Type I error rate and statistical power of the Monte Carlo procedure vis-à-vis the independent groups *t* test. The Type I error rate was verified by setting the population means of both the treatment and control groups equal to 0 on the dependent variable (each with a standard deviation of 1) and testing mean differences at $\alpha = .05$ over many random samples. Across the 50 relevant conditions (5 levels of sample size × 10 levels of reliability), we found an average of 247.8 out of 5,000 (4.96%) statistically significant *t* tests (i.e., $p < .05$), confirming the nominal alpha level of 5%.

Statistical power was verified by executing 5,000 iterations as above, but at various levels of effect size. Sample size was varied, but no random measurement error was added to the dependent variable. Our findings were compared with power estimates as computed using the G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007). Results showed that observed power closely mapped onto predicted power estimates. For example, predicted statistical power for an effect size of .5 is .19 for $N = 10$ and is .70 for $N = 50$. Our simulation results conformed closely to these estimates, with 921 of 5,000 (power = .184 at $N = 10$) and 3,493 of 5,000 (power = .699 at $N = 50$) samples showing statistically significant mean differences.

Replication Results

Our main simulation results are presented graphically in Figures 1, 2, and 3, for population effect sizes of .2, .5, and .8, respectively. As depicted in Figure 1, for such a small population effect size of only .2, our simulation results just barely revealed a trend suggestive of a linear relation between reliability and replication. This was most clearly seen for the largest sample size (i.e., $N = 50$), such that a reliability of .9 yielded a replicability rate of about .17. That value was approximately 3 times greater than the replicability of about .06 at a reliability of .1 (the latter replicability approaching the Type I error rate) and was close to the calculated power for that condition of .168 (calculated power estimates are shown by the rightmost point of each curve in Figure 1). It is important to note, however, that these less-than-clear results are not surprising given the very low levels of baseline statistical power. In other words, for small treatment effects, floor effects in statistical power tend to obscure any relation between reliability and replicability (the maximum replicability for any condition of Figure 1 was only about .18).

The relation between reliability and replicability was much clearer for simulation results for medium population effect sizes, as depicted in Figure 2. That figure shows an apparent linear increase in replicability as reliability levels increased, for all sample sizes. Also observable in Figure 2 is the finding that differences in replicability due to sample size were much more pronounced at higher levels of reliability. Interestingly, the reliability of the dependent variable does not seem to have much of an effect on replicability at small sample sizes (e.g., $N = 10$), where statistical power is low. With relatively large sample sizes (e.g., $N = 50$), however, the deleterious effect of unreliability on probability of replication is substantial.

It might be instructive to compare some of the specific results illustrated in Figure 2 at this point. For example, consider a dependent variable with a reliability of .80 (typical of well-constructed explicit measures) versus a dependent variable with a reliability of .30 (typical of some implicit measures).

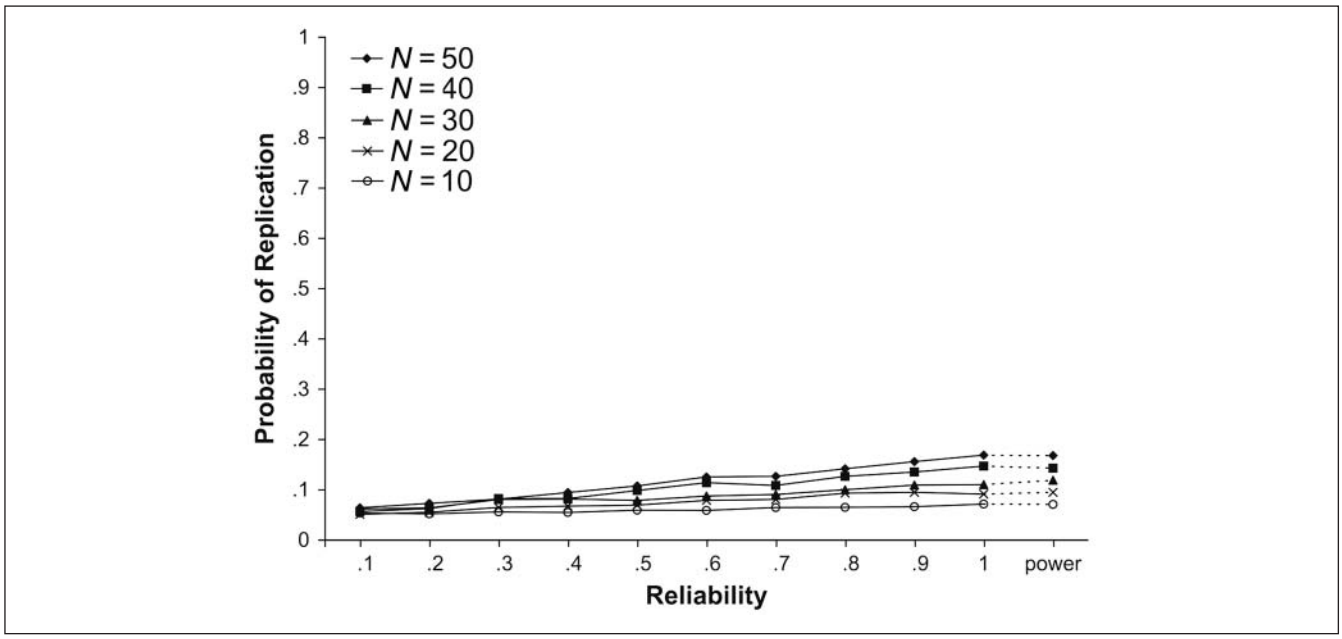

Figure 1. Observed probability of replicating a two-group mean difference effect as a function of sample size (*N*) and dependent variable reliability; population effect size equals .2
Note: Shown at the extreme right of the graph are the theoretical statistical power values.

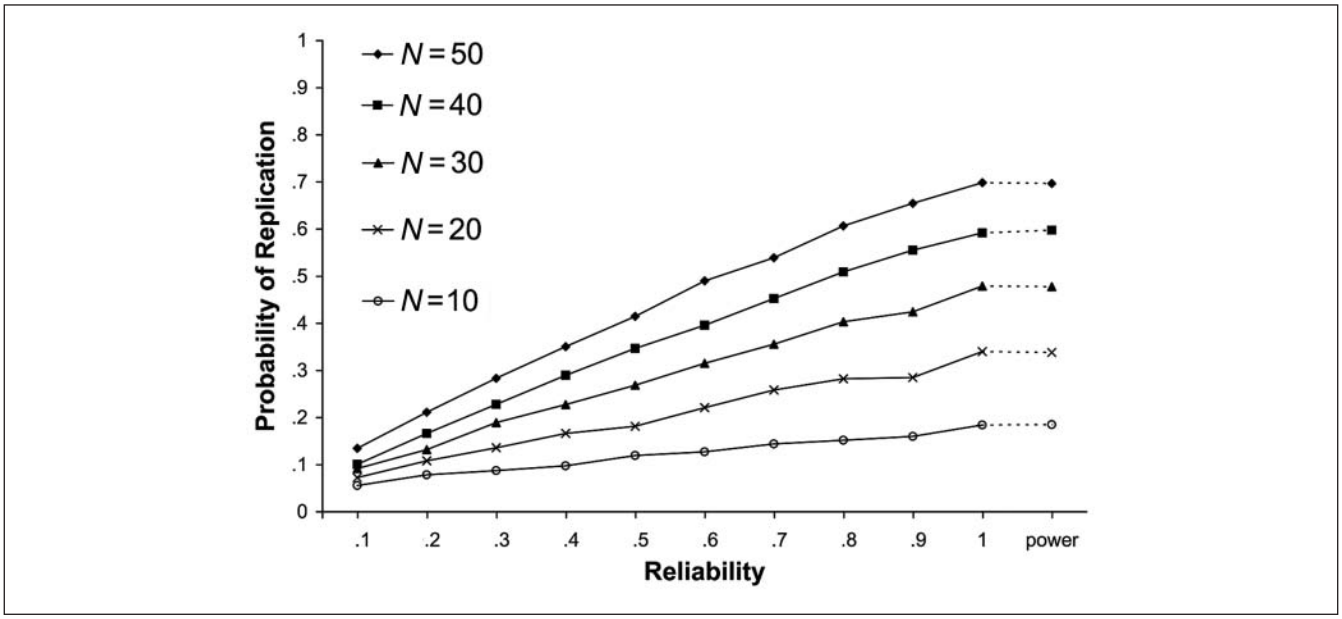

Figure 2. Observed probability of replicating a two-group mean difference effect as a function of sample size (*N*) and dependent variable reliability; population effect size equals .5
Note: Shown at the extreme right of the graph are the theoretical statistical power values.

In the former case, and at $N = 40$, the estimated probability of replicating a statistically significant mean difference found between two independent groups is approximately .50, whereas in the latter case it is only about .25. This is a notable difference, one that might in part explain why some effects involving implicit measures may be more difficult to replicate than comparable findings involving explicit measures.

The patterns of simulation results for a large population effect size are shown in Figure 3, and they generally paralleled our findings for a medium population effect size shown

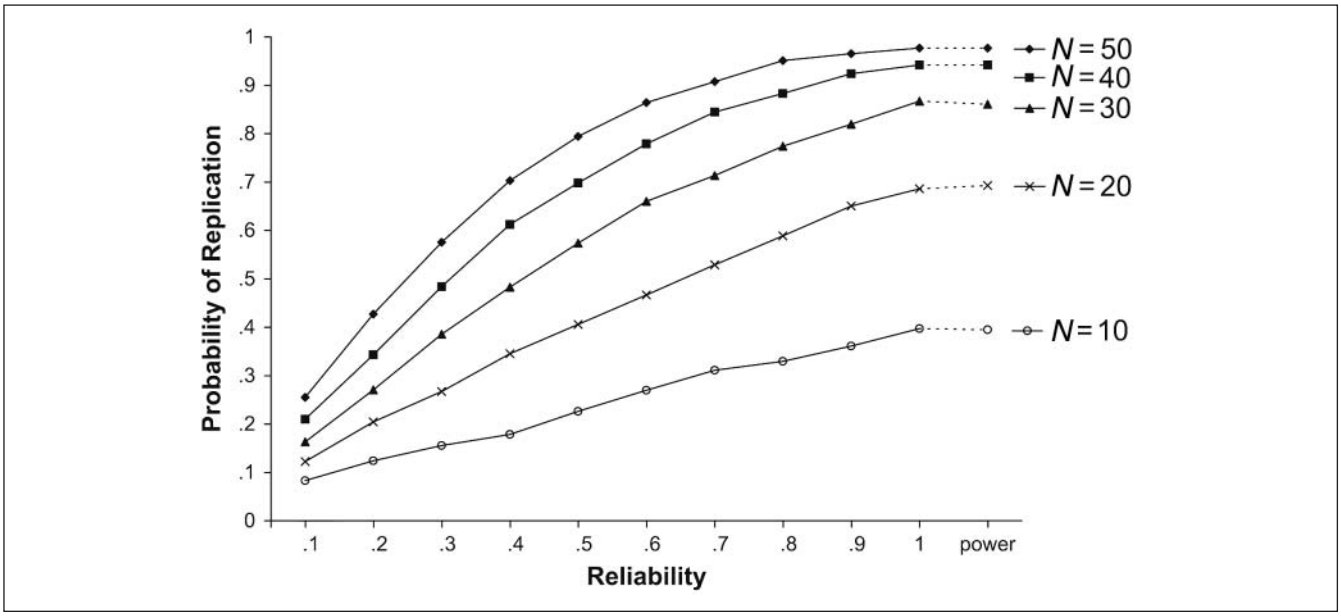

Figure 3. Observed probability of replicating a two-group mean difference effect as a function of sample size (*N*) and dependent variable reliability; population effect size equals .8
Note: Shown at the extreme right of the graph are the theoretical statistical power values.

in Figure 2. As depicted in both figures, replicability rates generally increased monotonically as levels of reliability increased, for all sample sizes. Two slight differences that appeared for large versus medium effect sizes, however, are worth mentioning. First, the differences across sample sizes (for a given reliability level) in replicability rates appear to be more pronounced for the larger effects. Second, for larger effect sizes, the relation between reliability and replicability is curved (nonlinear) for sample sizes greater than 20 or so, as replication approaches the ceiling of 1.0. This suggests diminishing returns on replicability with improvements in reliability beyond about .70, especially with sample sizes around 40 or greater. Overall, however, examining replicability rates for particular experimental conditions with large effect sizes (Figure 3) revealed the same general conclusion reached about medium effect sizes (Figure 2). For example, Figure 3 shows that using a measure with a reliability of .80 as compared to .30 (for a sample size of 30) essentially doubles the probability of replication from about .40 to .80.

Discussion

Using a Monte Carlo simulation, the present research investigated the impact of random measurement error on the replicability of experimental findings in the context of increasingly popular implicit measures of mental attributes. Although implicit measures have contributed to important insights for various psychological phenomena in almost every area of psychology, they have exhibited considerable variation in reliability, with many measures having low levels of reliability. The intended goal of the current investigation was to focus on an overlooked cost of such low reliability, that is, lower probability of replication when such measures are used as dependent variables in experimental research.

Results from our Monte Carlo simulation revealed that the probability of replicating an experimental effect systematically decreased as the random measurement error contaminating the scores increased. This pattern was especially pronounced for "medium" and "large" population effect sizes and for moderate to large sample sizes (i.e., *N* equal to or greater than 30 per condition). These simulation results resonate particularly well with Nunnally's (1982) general ideas, as reflected in our opening quote; that is, empirical results that are influenced by random measurement errors will not be exactly repeatable. Our results also echo the poor replicability of experimental effects that we and other colleagues have sometimes experienced regarding studies involving implicit measures.

Implications of Our Simulation Study

The current simulation results have at least three important implications for research involving implicit measures (or research using any other measures with questionable reliability, implicit or otherwise). The first, and most important, is the metascientific implication that unreliability of implicit measures can decrease more than just statistical power. As demonstrated empirically in our simulation, the unreliability of implicit measures can have far-reaching effects on replicability, in some cases dramatically reducing the probability of repeating a real experimental effect. For instance, for a large effect (and using a typical sample size of 30 or so), the probability of replicating a between-group experimental effect decreases from

approximately .80 to .40 when using a measure with a reliability of .30 compared to .80. This vividly illustrates the potent repercussions of implicit measures' typically lower reliability, with respect to the development of a cumulative science.

The second implication of our simulation findings is an exhortation that one should carefully take into account the reliability of implicit measures when evaluating research findings involving such measures. That is, researchers should calibrate their confidence in their experimental results as a function of the amount of random measurement error contaminating the scores of the dependent variable. In this sense, less confidence should be placed in experimental results based on implicit measure scores contaminated with a high degree of random measurement error. This, of course, assumes that the reliability estimates in the sample, or within each experimental condition, are an accurate reflection of the amount of measurement error inherent in the scores (i.e., one must rule out other factors that can reduce the accuracy of reliability estimates, such as restriction of range, outliers, calculation misspecifications, etc.). Hence, evaluating (and reporting) the reliability of scores produced by an implicit measure should be viewed as a mandatory requirement when gauging the robustness of a finding, above and beyond the evaluation of sample size, *p* values, and confidence intervals (also see Kashy, Donnellan, Ackerman, & Russell, 2009).

The third important implication of our study is that, to increase the probability of replicating an effect involving an implicit measure, researchers should attempt to increase the reliability of implicit measures having known psychometric shortcomings. Alternatively, researchers might choose to use only those implicit measures that have typically demonstrated acceptable levels of reliability, such as the IAT (Greenwald et al., 1998) or the AMP (Payne et al., 2005). However, given the critical importance of using a multimethod approach to corroborate the accuracy of theoretical claims involving constructs assessed using implicit measures (Gawronski, Deutsch, LeBel, & Peters, 2008), we argue that more research should be focused on improving implicit measures with unacceptable levels of reliability (see, e.g., Gawronski, Cunningham, LeBel, & Deutsch, 2010, with respect to the affective priming task) rather than simply abandoning those measures altogether.[4]

Replicability Concerns Specific to Implicit Measures

The results of our simulation study apply equally well to implicit measures, explicit measures, physiological measures, and any other type of measure. Our focus in this article, however, is on implicit measures in psychology because of their well-known issues with reliability. The question that then arises is, why is reliability particularly problematic for implicit measures as opposed to explicit measures? There are three main reasons why reliability may be lower in implicit measures to begin with, compared to explicit measures, and hence why replicability of experimental results may be more of a problem for the former. First, although the initial promise of implicit measures was that they could provide context-independent and stable indices of mental representations (e.g., Bargh, 1999; Fazio, Jackson, Dunton, & Williams, 1995), more recent empirical evidence and theorizing suggest that implicit measures may actually be *more* context dependent than are explicit measures (Ferguson & Bargh, 2007). According to Ferguson and Bargh's (2007) perspective, implicit measures might tap into mental representations that are highly sensitive to momentary personal or contextual factors, such as a person's recently activated memories, current goals, or present mood or even the race of the experimenter (e.g., Barden, Maddux, Petty, & Brewer, 2004). Hence, to the extent that these factors vary across measurement occasions, they can contribute to replication difficulties. Although some of these factors could also affect constructs assessed with explicit measures, the metacognitive and introspective processes involved in explicit measures may encourage stability in the assessment of the construct that would not emerge otherwise (Ferguson & Bargh, 2007).

Another reason why reliability may be lower in implicit measures as compared to explicit measures concerns procedural factors. For instance, the fact that many implicit measures are based on reaction times may contribute to unreliability because reaction times can vary considerably from one testing situation to the next. This variation can be a function of physiological, hormonal, emotional, or other changes in a respondent. Such factors are less likely to have an effect on the responses to a typical self-report questionnaire. Finally, the scoring of implicit measures may also contribute to lower levels of reliability. The scoring algorithms for many implicit measures, in contrast to explicit measures, often involve computing difference scores. Such aggregate scores suffer in reliability in direct proportion to the correlation between the individual components scores (Cronbach, 1958; Edwards, 2002).[5]

What can be said about the few implicit measures that have relatively high levels of reliability (e.g., IAT, AMP)? This demonstrates that the construction of such instruments is not impossible. An important open question here, however, is to what extent systematic construct-unrelated variance is driving the reliability estimates of these measures (either internal consistency or test–retest). Hence, even though these implicit measures might ostensibly have acceptable levels of reliability, it could still be the case that their experimental effects are difficult to replicate because of changing systematic variance contaminating the test scores across experiments (such as the momentary personal or contextual factors referred to above). Our analysis here implies that more attention needs to be focused on understanding the systematic artifacts possibly afflicting implicit measures (and explicit measures) with respect to their reliability estimates.

Advantages of Low Reliability?

We have heard some researchers express contrary views about the implications of implicit measures' low reliability for experimental results. First, there is the position that ignoring random measurement error is conservative because it makes it harder to find statistically significant effects. Although it is true that random measurement error will, all else being equal, reduce the size of most statistics, whether it is "conservative" to ignore random measurement error depends on the nature of the research question (Thye, 2000). For instance, if in a particular study making a Type II error (concluding no effect is present when one in fact exists) is more costly than making a Type I error (concluding an effect is present when none actually exists), then it would hardly be conservative to ignore random measurement error (e.g., investigating the harmful effects of being a victim of prejudice). In addition, ignoring random measurement error can lead to erroneous research conclusions in particular experimental conditions (Schmidt & Hunter, 1996; Thye, 2000). For example, if random measurement error is differentially associated with the treatment conditions of an experiment, a mean between-condition difference can be found when no true effect exists (i.e., a Type I error; Thye, 2000).

Another comment we have heard concerns the belief that experimental effects detected with unreliable measures may be *more* robust than those detected with reliable measures. Based on our simulation results, it is true that, for low levels of reliability (e.g., $\rho_{xx} = .20$), large experimental effects were more likely to replicate than were small and medium effects (e.g., for $N = 30$, replicability was about .25 for large effects, compared to about .15 and .05 for medium and small effects, respectively). However, it is important to note that a replicability rate of about .25 is nonetheless quite low in an absolute sense. In addition, it is essential to remember that the foundations of the scientific enterprise rest on the sound measurement of constructs. Hence, it is antithetical to any empirical science to use measurement instruments known to be psychometrically questionable. Indeed, some have argued that it is imperative to increase our understanding of the various sources of measurement error (whether systematic or random) contaminating the scores of our measures rather than to simply control measurement error post hoc (Deshon, 1998).

Finally, other researchers, based on their own experiences, have claimed that they have been able to easily replicate effects using certain implicit measures, despite their low reliabilities. But this cannot generally be the case because random measurement error unquestionably reduces replicability. In cases where their claim has been true, we would argue that the reliability estimates for the implicit measures in question may be inaccurate. Certainly, more measurement-oriented research is needed to tackle the difficult question of how best to estimate reliability, for reaction time tasks in particular and for implicit measures more broadly. New psychometric developments concerning methods of estimating reliability could be helpful here (Cronbach, 2004; Duhachek & Iacobucci, 2004).

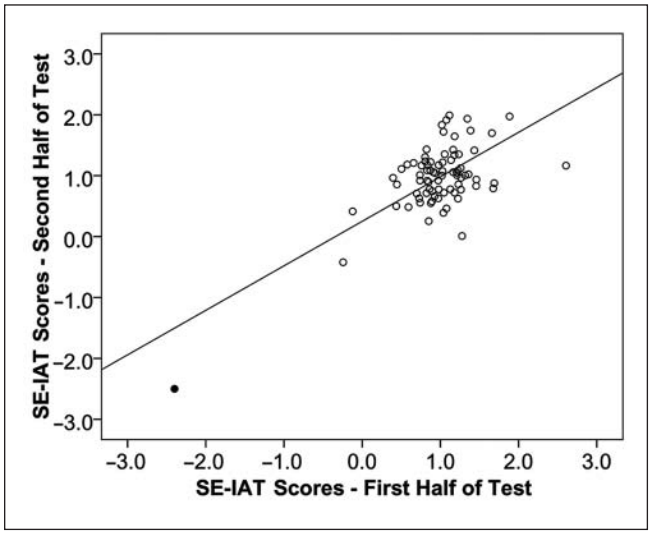

Figure 4. Scatterplot and regression line of 80 self-esteem Implicit Association Test (SE-IAT) scores on the second half of the test plotted against scores of the first half of the test
Note: The internal consistency reliability estimate across all subjects is $\alpha = .81$ ($r = .68$), but only $\alpha = .57$ ($r = .40$) if one outlier (closed circle) is removed.

Recommendations for Reliability Estimation

For researchers to adjust their confidence in experimental results involving implicit measures, the reliability of these measures needs to be routinely and accurately assessed and reported in studies where they are used as dependent variables. This raises issues concerning how reliability should be estimated for implicit measures in the context of experimental studies. First, because Cronbach's alpha is based on the mean interitem correlation among items in a measure, careful attention must be paid to the potential influence of outliers that can affect that correlation (see Liu & Zumbo, 2007, for a discussion of this issue). This is particularly relevant for the common practice of estimating reliability for implicit measures by splitting the test into two halves, computing an implicit measure score for both halves, and then computing a Cronbach's alpha based on those two scores (e.g., as is typically done to estimate reliability for IAT scores).

In the context of split-half reliability, the presence of an outlier or two can severely distort reliability estimates by potentially inflating (or deflating) the correlation between the two test halves (see LeBel & Gawronski, 2009). Figure 4 demonstrates an example of this problem. Depicted is a scatterplot of 80 self-esteem IAT scores for the first and second half of the test (based on an unpublished data set). Now, the internal consistency of the data shown by the 79 open circles of Figure 4 is $\alpha = .57$. With the addition of only one outlier to this data set, however, shown by the closed circle at the lower left of the

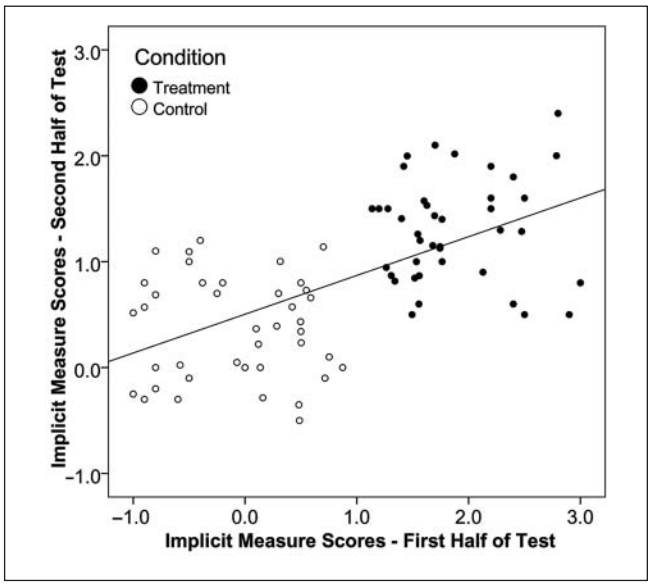

Figure 5. Scatterplot and regression line of hypothetical scores on the second half of an implicit measure plotted against scores on the first half of the measure
Note: Open circles represent control group scores and closed circles represent treatment group scores.

figure, the reliability estimate jumps to $\alpha = .81$. Such an outlier could arise if, for example, a person received a particularly low score on the self-esteem IAT due to any number of construct-irrelevant factors (e.g., fatigue, faking strategies, construct-unrelated response tendencies, etc.). As this example vividly demonstrates, it is important to probe interitem scatterplots for anomalies that can cause spurious reliability estimates.

We also note in the present context that reliability estimates should be reported separately for individual experimental conditions rather than for all conditions combined. These estimates, especially in an experimental context, can be artifactually inflated due to an increase in true score variance (relative to total observed score variance) caused by the experimental manipulation (Thye, 2000). Consider the hypothetical dataset shown in Figure 5. These data illustrate first-half and second-half test scores on an implicit dependent variable measure for a control group (open circles) and a treatment group (closed circles). The two groups clearly represent two different populations, having their own distributions of scores. As is evident in the figure, even though within each experimental condition there is virtually no association between test halves ($r = .06$ for the control group, and $r = -.04$ for the treatment group), a strong association emerges when collapsing across conditions (i.e., $r = .61$). Hence, if internal consistency is computed on the entire sample, the reliability estimate ($\alpha = .70$) would be artificially inflated due to group mean differences and completely erroneous. Along similar lines, another important reason to report reliability estimates separately for each experimental condition is to confirm that the construct in question was measured equally well across groups that are being compared. If one cannot be confident that the psychometric integrity of the measure remained the same across experimental groups (i.e., if reliability is drastically different across the conditions), then differences in observed scores across groups cannot be meaningfully interpreted (DeShon, 2004).

As a concrete example of our reliability estimation recommendations for implicit measures, consider a situation where a researcher uses such a measure as a dependent variable in an experimental study. To assess the reliability of the dependent measure, the researcher could begin by computing implicit measure scores for two separate subsets of trials of the implicit measure for each subject. For example, an implicit measure score could be calculated for a subject for all odd-numbered trials and another implicit measure score calculated separately for all even-numbered trials. A split-half reliability estimate is then calculated for the sample of subjects based on those two sets of scores (i.e., the correlation between the odd and even sets of scores, adjusted by the Spearman–Brown formula; see Lord & Novick, 1968, p. 112; Paunonen, 1984, p. 385). And the cautions expressed earlier pertain. That is, the researcher should carefully probe for the influence of outliers on the reliability estimates, and the reliability estimation should be done separately for each condition of the experimental study. The reliability estimates should then be reported, so that other researchers can gauge the replicability of the results due to random measurement error contamination.

Conclusion

Using a Monte Carlo simulation, we examined the impact of random measurement error in a dependent variable (i.e., unreliability) on replicating a between-group mean difference effect. We placed this research in the context of increasingly popular implicit measures, which show considerable variation in reliability, with many measures having low levels of reliability. Although our conclusions apply equally well to any measure with questionable reliability (implicit or otherwise), our simulation results support the following three main conclusions:

1. The probability of replicating an experimental effect decreases as random measurement error (i.e., low reliability) contaminates the dependent variable.
2. To inspire confidence in experimental results involving implicit measures, researchers need to improve those implicit measures having unacceptable levels of reliability or then utilize implicit measures known to have acceptable psychometric properties.
3. The reliability of implicit (or explicit) measures should be routinely (and accurately) assessed and reported in research articles, and in the case of experimental studies, reliability estimates should be reported separately for each experimental condition.

Acknowledgments

We would like to thank Bertram Gawronski for providing valuable feedback on an earlier version of this article and for his constructive discussions with the first author, which inspired many of the ideas in this article. We would also like to thank Ben Bowles for his valuable comments on a previous version of this article and Kurt Peters for fruitful conceptual discussions.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial Disclosure/Funding

The authors disclosed receipt of the following financial support for the research and/or authorship of this article: The present research was supported by the Social Sciences and Humanities Research Council of Canada Doctoral Fellowship 767-2007-1425 to the first author and Research Grant 410-2010-2586 to the second author.

Notes

1. Note that classical test theory can be seen as a special case of the broader measurement theory framework of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), which allows for the modeling of multiple sources of error variance within the error component (e.g., error due to raters, time, settings). Hence, the conclusions arising from the current investigation are not strictly limited to adherents of the classical test theory perspective.
2. The focus of the current investigation is strictly on the impact of *random* measurement error (i.e., nonsystematic error variance) on the replicability of observed scores. It is also possible, of course, to partition observed scores into systematic error (construct-irrelevant) variance and true score (construct-relevant) variance. However, such an investigation would concern the measure's validity and, therefore, is beyond the scope of this article. Furthermore, the conclusions we reach based on our simulation study, regarding random error and replicability, hold regardless of the extent to which systematic error variance impinges on a measure's observed scores.
3. The population standard deviations of the error scores, σ_E, for desired levels of reliability, ρ_{xx}, were calculated using the following general formula: $\sigma_E = \sqrt{\frac{(1-\rho_{xx})}{\rho_{xx}}}$.
4. Another point made salient by our findings is that replicability could be enhanced in social psychological research by using larger sample sizes in general (as is illustrated clearly in Figures 1 to 3). We know of no statistician, however, who would recommend using larger sample sizes as a means of compensating for unsound psychometric instruments.
5. We thank an anonymous reviewer for this suggestion.

References

Algom, D., Chajut, E., & Shlomo, L. (2004). A rational look at the emotional Stroop phenomenon: A generic slowdown, not a Stroop effect. *Journal of Experimental Psychology: General*, *133*, 323-338.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or would Jesse Jackson 'fail' the Implicit Association Test? *Psychological Inquiry*, *15*, 257-278.

Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes. *Journal of Personality and Social Psychology*, *87*, 5-22.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361-382). New York, NY: Guilford.

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, *62*, 254-263.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*, 27-41.

Blanton, H., & Jaccard, J. (2008). Unconscious racism: A concept in pursuit of a measure. *Annual Review of Sociology*, *34*, 277-297.

Blanton, H., Jaccard, J., Gonzales, P., & Christie, C. (2006). Decoding the Implicit Association Test: Perspectives on criterion prediction. *Journal of Experimental Social Psychology*, *42*, 192-212.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychology Inquiry*, *13*, 103-124.

Borkenau, P., & Mauer, N. (2007). Well-being and the accessibility of pleasant and unpleasant concepts. *European Journal of Personality*, *21*, 169-189.

Borkenau, P., Paelecke, M., & Yu, R. (2009). Personality and lexical decision times for evaluative words. *European Journal of Personality*, *24*, 123-136.

Bosson, J. K., Lakey, C. E., Campbell, W. K., Zeigler-Hill, V., Jordan, C. H., & Kernis, M. H. (2008). Untangling the links between narcissism and self-esteem: A theoretical and empirical review. *Personality and Social Psychology Compass*, *2/3*, 1415-1439.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, *79*, 631-643.

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology*, *40*, 227-259.

Campbell, W. K., Bosson, J. K., Goheen, T. W., Lakey, C. E., & Kernis, M. H. (2007). Do narcissists dislike themselves "deep down inside"? *Psychological Science*, *18*, 227-229.

Charles, E. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, *10*, 206-226.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *12*, 1-16.

Cronbach, L. J. (1958). Proposals leading to analytic treatment of social perception scores. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior* (pp. 353-379). Stanford, CA: Stanford University Press.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3, 286-300.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measurement: Consistency, stability, and convergent validity. *Psychological Science*, *12*, 163-170.

De Houwer, J. (2003). The Extrinsic Affective Simon Task. *Experimental Psychology*, *50*, 77-85.

De Houwer, J., & De Bruycker, E. (2007). The identification-EAST as a valid measure of implicit attitudes toward alcohol-related stimuli. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*, 133-143.

De Raedt, R., Schacht, R., Franck, E., & De Houwer, J. (2006). Self-esteem and depression revisited: Implicit positive self-esteem in depressed patients? *Behaviour Research and Therapy*, *44*, 1017-1028.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods*, *3*, 412-423.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*, 137-149.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*, 62-28.

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*, 792-808.

Edwards, J. R. (2002). Ten difference score myths. *Organizational Research Methods*, *4*, 265-287.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*, 1013-1027.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297-327.

Ferguson, M. J., & Bargh, J. A. (2007). Beyond the attitude object: Implicit attitudes spring from object-centered contexts. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes: Procedures and controversies* (pp. 216-246). New York, NY: Guilford.

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. *Science*, *321*, 1100-1102.

Gawronski, B. (2002). What does the Implicit Association Test measure? A test of the convergent and discriminant validity of prejudice related IATs. *Experimental Psychology*, *49*, 171-180.

Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorization influence spontaneous evaluations of multiply categorizable objects? *Cognition and Emotion*, *24*, 1008-1025.

Gawronski, B., Deutsch, R., & Banse, R. (in press). Response interference tasks as indirect measures of automatic associations. In K. C. Klauer, C. Stahl, & A. Voss (Eds.), *Cognitive methods in social psychology*. New York, NY: Guilford.

Gawronski, B., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment*, *24*, 218-225.

Gawronski, B., LeBel, E. P., & Peters, K. P. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science*, *2*, 181-193.

Gray, N. S., McCulloch, M. J., Smith, J., Morris, M., & Snowden, R. J. (2003). Violence viewed by psychopathic murderers: Adapting a revealing test may expose those psychopaths who are most likely to kill. *Nature*, *423*, 497-498.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4-27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197-216.

Hopkins, K. D., & Hopkins, B. R. (1979). The effect of the reliability of the dependent variable on power. *Journal of Special Education*, *13*, 463-466.

Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, *117*, 348-357.

Johnson, R. E., & Steinman, L. (2009). Use of implicit measures for organizational research: An empirical example. *Canadian Journal of Behavioural Science*, *41*, 202-212.

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality and Social Psychology*, *85*, 969-978.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*, 1131-1142.

Kawakami, K., & Dovidio, J. F. (2001). Implicit stereotyping: How reliable is it? *Personality and Social Psychology Bulletin*, *27*, 212-225.

Kopriva, R. J., & Shaw, D. G. (1991). Power estimates: The effect of dependent variable reliability on the power of one-factor ANOVAs. *Educational and Psychological Measurement*, *51*, 585-595.

LeBel, E. P., & Gawronski, B. (2009). How to find what's in a name: Scrutinizing the optimality of five scoring algorithms for the name-letter task. *European Journal of Personality*, *23*, 85-106.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, education, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181-1209.

Liu, Y., & Zumbo, B. D. (2007). The impact of outliers on Cronbach's coefficient alpha estimate of reliability. *Educational and Psychological Measurement*, *67*, 620-634.

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

Meier, B., & Perrig, W. J. (2000). Low reliability of perceptual priming: Consequences for the interpretation of functional dissociations between explicit and implicit memory. *Quarterly Journal of Experimental Psychology*, *53*, 211-233.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition*, *19*, 625-666.

Nunnally, J. C. (1982). Reliability of measurement. In H. E. Mitzel (Ed.), *Encyclopedia of educational research* (pp. 1589-1601). New York, NY: Free Press.

Nuttin, M. J., Jr. (1985). Narcissism beyond gestalt and awareness: The name letter effect. *European Journal of Social Psychology*, *64*, 723-739.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science*, 14, 636-639.

Papies, E. K., Stroebe, W., & Aarts, H. (2009). Who likes it more? Restrained eaters' implicit attitudes towards food. *Appetite*, *53*, 279-287.

Paunonen, S. V. (1984). The reliability of aggregated measurements: Lessons to be learned from psychometric theory. *Journal of Research in Personality*, *18*, 383-394.

Payne, B. K., Cheng, S. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*, 277-293.

Perez, L. A., Peynircio, Z. F., & Blaxton, T. A. (1998). Developmental differences in implicit and explicit memory performance. *Journal of Experimental Child Psychology*, *70*, 167-185.

Robie, C., & Komar, S. G. (2007). Simulation, computer approach. In S. G. Rogelberg (Ed.), *Encyclopedia of industrial and organizational psychology* (pp. 723-724). Thousand Oaks, CA: Sage.

Rogers, W. T., & Hopkins, K. D. (1988). Power estimates in the presence of a covariate and measurement error. *Educational and Psychological Measurement*, *48*, 647-56.

Schimmack, U., & Diener, E. (2003). Predictive validity of explicit and implicit self-esteem for subjective well-being. *Journal of Research in Personality*, *37*, 100-106.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 27 research scenarios. *Psychological Methods*, *1*, 199-223.

Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality*, *19*, 595-605.

Schott, B. H., Henson, R. N., Richardson-Klavehn, A., Becker, C., Thoma, V., Heinze, H.-J., & Düzel, E. (2005). Redefining implicit and explicit memory: The functional neuroanatomy of priming, remembering, and control of retrieval. *Proceedings of National Academy Science*, *102*, 1257-1262.

Spruyt, A., Hermans, D., De Houwer, J., Vandekerckhove, J., & Eelen, P. (2007). On the predictive validity of indirect attitude measures: Prediction of consumer choice behavior on the basis of affective priming in the picture-picture naming task. *Journal of Experimental Social Psychology*, *43*, 599-610.

Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika*, *23*, 9-17.

Teige, S., Schnabel, K., Banse, R., & Asendorpf, J. B. (2004). Assessment of multiple implicit self-concept dimensions using the Extrinsic Affective Simon Task. *European Journal of Personality*, *18*, 495-520.

Thye, S. R. (2000). Reliability in experimental sociology. *Social Forces*, *78*, 1277-1309.

Uhlmann, E. L., Pizarro, D. A., & Bloom, P. (2008). Varieties of unconscious social cognition. *Journal for the Theory of Social Behaviour*, *38*, 293-322.

Williams, R. H., & Zimmerman, D. W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology*, *116*, 359-369.

Wittenbrink, B. (2007). Measuring attitudes through priming. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes: Procedures and controversies* (pp. 17-58). New York, NY: Guilford.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*, 262-274.

Wittenbrink, B., & Schwarz, N. (Eds.). (2007). *Implicit measures of attitudes.* New York, NY: Guilford.

Woltz, D. J., & Shute, V. J. (1993). Individual difference in repetition priming and its relationship to declarative knowledge acquisition. *Intelligence*, *17*, 333-359.

Zeigler-Hill, V. (2006). Discrepancies between implicit and explicit self-esteem: Implications for narcissism and self-esteem instability. *Journal of Personality*, *74*, 119-143.