Supplement to Greenwald & Lai (2020):
A Meta-Analysis of the Internal Consistency & Test-Retest Reliability of Implicit Measures

Last updated: September 26, 2019
Questions about the meta-analytic method or results? Contact Calvin Lai at calvinlai@wustl.edu.

Table of Contents

Summary

We conducted a meta-analysis of the internal consistency (IC) and test-retest reliability (TRR) of implicit measures using robust variance estimation (Hedges, Tipton, & Johnson, 2010). The ICs were acceptable to good for most measures, with an overall meta-analytic Cronbach's alpha of .77, 95% CI [.76, .79]. The TRRs were satisfactory for most measures, with an overall meta-analytic effect correlation of .44, 95% CI [.40, .49]. Implicit measures differed considerably in the IC and TRR estimates. See Tables S1 and S2 for a summary of the results.

For all data, analysis scripts, a codebook, instructions for reproducing the meta-analysis, and any updates to this document, see https://osf.io/cv83q/.

Method

Inclusion criteria. We set the following inclusion criteria:
(1) *The study must include an implicit measure listed in Table 1 of Greenwald & Lai (2020)*. We excluded variants of implicit measures that deviated greatly from typical practice (e.g., a study that used auditory stimuli on the IAT; Roessel, Schoel, & Stahlberg, 2018).

(2) *The article must report relevant reliability statistics.* For internal consistency, we recorded Cronbach's alpha, split-half correlations, and Spearman-brown-adjusted split-half correlations. For test-retest reliability, we recorded unadjusted correlations between multiple administrations of the same implicit measure.

(3) *If there is an experimental procedure before the implicit measure, the experimental procedure is not expected to change the measure's internal reliability or test-retest reliability.* As one example, two articles that found instructions to pay explicit attention to category

membership reduced the internal consistency of the evaluative priming task were excluded (Gawronski, Cunningham, LeBel, & Deutsch, 2010; Olson & Fazio, 2003).

(4) The *study is reported in English*. We excluded studies that were not written in English.

Article retrieval. We searched for relevant articles using the Scopus database with the following search terms: (Article Title, Abstract, Keywords: names of implicit measures or acronyms, e.g., "Affect Misattribution Procedure", "ST-IAT") AND (All fields: "internal consistency" or "split-half reliability" or "odd-even reliability" or "test retest" or "Cronbach") AND Document Type: Article. For information on the search terms we used for particular measures see "List of Measure Search Terms" below. This resulted in 185 records that potentially matched our inclusion criteria. We supplemented this search with a review of the primary publication named in Table 1 (7 records containing eligible estimates), articles mentioned in three major psychometric reviews not covered in the original search (39 records containing eligible estimates; Golijani-Moghaddam, Hart, & Dawson, 2013; Payne & Lundberg, 2014; Rae & Olson, 2018) and articles obtained from direct requests the authors of the primary publications (21 records containing eligible estimates). This process yielded a final sample of 166 articles that reported results from 331 independent samples with 562 IC effect sizes and 169 TRR effect sizes.

Effect size selection. In many articles, authors provided several possible reliability estimates to record. In those cases, we followed these decision rules:

- If an overall estimate of an implicit measure was given, we recoded that estimate.
- If estimates were only given for multiple trial types within an implicit measure, we took the average reliability of those trial types.
- If reliability was given for multiple scoring methods, we recorded reliability for the most commonly used scoring method.
- If only a range of reliability estimates was given, we recorded the midpoint of that range.
- If only an average of the reliability estimates was given, we recorded that average.

Effect size computation.

Internal consistency. All analyses were conducted on alpha coefficients, either as Cronbach's alpha or Spearman-Brown adjusted correlations (also known as standardized Cronbach's alpha). For analysis, estimates that were reported as unadjusted split-half correlations were recalculated into Spearman-Brown adjusted correlations using the formula: $(2r)/(1 + r)$. For analysis, alpha coefficients were transformed into transformed alphas with an approximately normal distribution using Bonett's (2002) transformation: $-\ln(1 - \alpha)$. After analysis, the meta-analytic mean and confidence limits were transformed back into raw alpha coefficients.

Test-retest reliability. For analysis, raw correlation coefficients were transformed using Fisher's *r*-to-*z* transformation to a Z-score with a normal distribution. After analysis, the meta-analytic mean and confidence limits were transformed back into raw correlation coefficients.

Robust variance estimation meta-analysis. All meta-analytic models were fitted using a robust variance estimation method (Hedges et al., 2010) implemented in the *robumeta* package in R (Fisher & Tipton, 2015). This method incorporates the dependence between multiple effect sizes drawn from the same sample and has been useful in prior meta-analyses of implicit measures (Kurdi et al., 2019; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). For dependent correlations, a single parameter estimate is assumed. As recommended, we examined whether the estimated models changed when the assumed dependent correlations is varied (r = .00, r = .50, r =.80, and r = 1.00). The estimates were robust to assumptions about dependent correlations, so we report analyses with the default value provided by the package (r = .80).

Unanalyzed variables. We also recorded several variables that were not used in the present analyses but could be useful in future analyses of this dataset. See the codebook and dataset at https://osf.io/cv83q/ for more information.

Results

Descriptive statistics. Tables S1 and S2 provides overviews of the internal consistency and test-retest reliability estimates respectively.

Internal consistency. The ICs were acceptable to good for most measures, with an overall meta-analytic Cronbach's alpha of .77, 95% CI [.76, .79]. However, there was considerable variation between measures. The three most internally consistent measures were the AMP (α = .81), IAT (α = .80), and Brief IAT (α = .79). The three least internally consistent measures were the EAST (α = .38), Evaluative Priming Task (α = .53), and Approach-Avoidance Task (α = .62).

Test-retest reliability. The TRRs were satisfactory for most measures, with an overall meta-analytic effect correlation of .44, 95% CI [.40, .49]. As with IC, there was considerable variation between measures. The three most reliable measures were the Name-Letter Effect (r = .56), AMP (r = .52), and IAT (r = .50). The four least reliable measures were the Approach-Avoidance Task (r = .10), EAST (r = .24), SC-IAT (r = .25), and Evaluative Priming Task (r = .26).

Special cases. A description of ad-hoc decisions are described in "Notes" variable in the online dataset. Here, we review two large-scale datasets presented unique challenges for meta-analysis. The first dataset was from Nosek and colleagues (2007), which reported IC estimates from a large-scale study that included 2,575,535 participants across 17 IATs. Due to the size of this dataset, it is possible that the results could be skewed by the results of this one study. As a robustness check, we computed internal consistency with and without this dataset. When IC is estimated without this dataset, the results were almost identical: α = .81, 95% CI [.79, .82]. The second dataset was Bar-Anan & Nosek (2014), which was a large-scale study that employed a planned-missingness design. In this design, each participant was randomly assigned to complete a portion of a set of implicit measures. This led to estimates that were mostly (but not completely) independent from other estimates within the sample. For analysis, we treated these results as if they were fully independent samples.

Table S1. Summary table of internal consistency estimates. # of articles = Number of articles contributing to the meta-analytic estimate, # of samples = Number of independent samples included in the model, # of effects = Number of effect sizes included in the model, α = raw alpha coefficient, CI = confidence interval, τ^2 = residual heterogeneity.

Families of Measures	**# of articles**	**# of samples**	**# of effects**	**α (95% CI)**	**τ^2**	**I^2**
Priming variations						
Evaluative Priming (EPT)	12	14	21	.53 (.35, .66)	0.21	89.87
Semantic Priming Task	N/A	N/A	N/A	N/A	N/A	N/A
Lexical Decision Priming (LDT)	N/A	N/A	N/A	N/A	N/A	N/A
Affect Misattribution Procedure (AMP)	36	68	73	.81 (.77, .83)	0.53	96.39
Implicit Association Test variations						
Implicit Association Test (IAT)	89	167	257	.80 (.78, .81)	0.03	99.30
Go/No-Go (GNAT)	8	17	18	.66 (.57, .74)	0.18	87.80
Single category (SC-IAT)	13	19	33	.76 (.71, .80)	0.13	79.28
Implicit Relational Assessment Procedure (IRAP)	13	18	23	.60 (.47, .69)	0.26	78.02
Single target (ST-IAT)	6	10	16	.78 (.64, .86)	0.16	93.68
Brief IAT (BIAT)	9	14	61	.79 (.74, .83)	0.16	91.13
Recoding Free (IAT-RF)	2	4	4	.69 (.32, .86)	0.00	0.00
Other (Miscellaneous)						
Name-Letter Effect (NLE)	5	6	8	.66 (.54, .75)	0.04	56.28
Linguistic Intergroup Bias (LIB)	N/A	N/A	N/A	N/A	N/A	N/A
Stimulus Response Compatibility Task (SRCT)	2	2	5	.81 (.00, 1.00)	0.84	96.00
Extrinsic Affective Simon Task (EAST)	13	18	24	.38 (.2, .52)	0.20	78.22
Stereotypic Explanatory Bias (SEB)	N/A	N/A	N/A	N/A	N/A	N/A
Approach-Avoidance Task (AAT)	10	12	19	.62 (.44, .74)	0.18	87.06
MouseTracker	N/A	N/A	N/A	N/A	N/A	N/A

Table S2. Summary table of test-retest reliability estimates. # of articles = Number of articles contributing to the meta-analytic estimate, # of samples = Number of independent samples included in the model, # of effects = Number of effect sizes included in the model, *r* = correlation coefficient, CI = confidence interval, τ^2 = residual heterogeneity.

Families of Measures	**# of articles**	**# of samples**	**# of effects**	***r* (95% CI)**	**τ^2**	**I^2**
Priming variations						
Evaluative Priming (EPT)	7	9	16	.26 (.06, .44)	0.06	84.83
Semantic Priming Task	N/A	N/A	N/A	N/A	N/A	N/A
Lexical Decision Priming (LDT)	N/A	N/A	N/A	N/A	N/A	N/A
Affect Misattribution Procedure (AMP)	3	5	7	.52 (.26, .70)	0.05	85.09
Implicit Association Test variations						
Implicit Association Test (IAT)	26	43	58	.50 (.45, .55)	0.08	89.22
Go/No-Go (GNAT)	3	5	5	.48 (.23, .67)	0.04	84.39
Single category (SC-IAT)	3	4	7	.25 (.15, .33)	0.00	0.00
Implicit Relational Assessment Procedure (IRAP)	2	3	5	.43 (.17, .63)	0.00	0.00
Single target (ST-IAT)	2	4	8	.43 (.26, .58)	0.01	49.14
Brief IAT (BIAT)	6	11	32	.43 (.28, .56)	0.07	87.60
Recoding Free (IAT-RF)	N/A	N/A	N/A	N/A	N/A	N/A
Other (Miscellaneous)						
Name-Letter Effect (NLE)	4	4	24	.56 (.29, .74)	0.04	82.73
Linguistic Intergroup Bias (LIB)	N/A	N/A	N/A	N/A	N/A	N/A
Stimulus Response Compatibility Task (SRCT)	N/A	N/A	N/A	N/A	N/A	N/A
Extrinsic Affective Simon Task (EAST)	3	3	3	.24 (-.16, .57)	0.01	58.09
Stereotypic Explanatory Bias (SEB)	N/A	N/A	N/A	N/A	N/A	N/A
Approach-Avoidance Task (AAT)	4	4	4	.10 (-.15, .34)	0.01	64.54
MouseTracker	N/A	N/A	N/A	N/A	N/A	N/A

List of Measure Search Terms

Measure	Measure's Search Terms
Priming variations	
Evaluative Priming (EPT)	(TITLE-ABS-KEY("sequential priming" OR "evaluative priming" OR "evaluative decision task")) OR (TITLE-ABS-KEY("affective priming task") AND ALL("attitude" or "stereotyp*"))
Semantic Priming Task	TITLE-ABS-KEY("stereotype priming" or "stereotype prime")
Lexical Decision Priming (LDT)	TITLE-ABS-KEY(lexical decision) AND ALL("attitude" or "stereotyp*")
Affect Misattribution Procedure (AMP)	TITLE-ABS-KEY("affect misattribution procedure")
Implicit Association Test variations	
Implicit Association Test (IAT)	TITLE-ABS-KEY("Implicit Association Test" or "Implicit Association Task")
Go/No-Go (GNAT)	TITLE-ABS-KEY("Go/No-Go Association Task")
Single category (SC-IAT)	TITLE-ABS-KEY("Single-Category Implicit Association Test" or "SC-IAT" or "SC-IATS")
Implicit Relational Assessment Procedure (IRAP)	TITLE-ABS-KEY("implicit relational assessment procedure")
Single target (ST-IAT)	TITLE-ABS-KEY("Single-Target Implicit Association Test" or "ST-IAT" or "ST-IATS")
Brief IAT (BIAT)	TITLE-ABS-KEY("Brief Implicit Association Test" OR "Brief IAT" or "BIAT" or "BIATS")
Recoding Free (IAT-RF)	TITLE-ABS-KEY("Recoding Free IAT" or "RF-IAT" or "RF-IATS" or "IAT-RF" or "IAT-RFS")
Other methods	
Name-Letter Effect (NLE)	TITLE-ABS-KEY("Name letter effect" or "initial preference task" or "initials preference task")
Linguistic Intergroup Bias (LIB)	TITLE-ABS-KEY("Linguistic Intergroup Bias")
Stimulus Response Compatibility Task (SRCT)	TITLE-ABS-KEY("Stimulus Response Compatibility Task" OR "SRC task") and ALL("approach")
Extrinsic Affective Simon Task (EAST)	TITLE-ABS-KEY("Stereotypic Explanatory Bias")
Stereotypic Explanatory Bias (SEB)	TITLE-ABS-KEY("Extrinsic Affective Simon Task")
Approach-Avoidance Task (AAT)	TITLE-ABS-KEY("approach avoidance task" OR "approach avoid task")
MouseTracker	TITLE-ABS- KEY("mousetracker" OR "Mouse Tracker" OR "Mouse Tracking") AND ALL("attitude" or "stereotyp*")

References

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods, 46,* 668–688.

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

Fisher, Z., & Tipton, E. (2015). *robumeta: An R-package for robust variance estimation in meta-analysis*. Retrieved from https://arxiv.org/abs/1503.02220v1

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

Golijani-Moghaddam, N., Hart, A., & Dawson, D. L. (2013). The Implicit Relational Assessment Procedure: Emerging reliability and validity data. *Journal of Contextual Behavioral Science, 2,* 105–119.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1,* 39–65.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., … Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *The American Psychologist*. https://doi.org/10.1037/amp0000364

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., … Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18,* 36–88.

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

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105,* 171–192.

Payne, K., & Lundberg, K. (2014). The Affect Misattribution Procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass, 8,* 672–686.

Rae, J. R., & Olson, K. R. (2018). Test-retest reliability and predictive validity of the implicit association test in children. *Developmental Psychology, 54,* 308–330.

Roessel, J., Schoel, C., & Stahlberg, D. (2018). What's in an accent? General spontaneous biases against nonnative accents: An investigation with conceptual and auditory IATs. *European Journal of Social Psychology, 48,* 535–550.