Behav Res (2017) 49:1193–1209
DOI 10.3758/s13428-016-0779-0

Modeling stimulus variation in three common implicit attitude tasks

Katie Wolsiefer[1] · **Jacob Westfall**[2] · **Charles M. Judd**[1]

Published online: 12 August 2016

Abstract We explored the consequences of ignoring the sampling variation due to stimuli in the domain of implicit attitudes. A large literature in psycholinguistics has examined the statistical treatment of random stimulus materials, but the recommendations from this literature have not been applied to the social psychological literature on implicit attitudes. This is partly because of inherent complications in applying crossed random-effect models to some of the most common implicit attitude tasks, and partly because no work to date has demonstrated that random stimulus variation is in fact consequential in implicit attitude measurement. We addressed this problem by laying out statistically appropriate and practically feasible crossed random-effect models for three of the most commonly used implicit attitude measures—the Implicit Association Test, affect misattribution procedure, and evaluative priming task—and then applying these models to large datasets (average N = 3,206) that assess participants' implicit attitudes toward race, politics, and self-esteem. We showed that the test statistics from the traditional analyses are substantially (about 60 %) inflated relative to the more-appropriate analyses that incorporate stimulus variation. Because all three tasks used the same stimulus words and faces, we could also meaningfully compare the relative contributions of stimulus variation across the tasks. In an appendix, we give syntax in R, SAS, and SPSS for fitting the recommended crossed random-effects models to data from all three tasks, as well as instructions on how to structure the data file.

Keywords Implicit attitudes · Stimulus sampling · Mixed models

Over the past three decades, researchers have increasingly relied on implicit measures to assess automatic associations (Fazio & Olson, 2003; Gawronski & Payne, 2010; Greenwald & Banaji, 1995). These measures typically require that participants make speeded responses to a set of stimuli over repeated trials. The stimuli themselves are selected to represent categories that are thought to elicit automatic or implicit associations; for example, in studies of implicit race attitudes (e.g., Fazio, Jackson, Dunton, & Williams, 1995), the typical stimuli would be pleasant and unpleasant words and photographs of Black and White faces. The responses to stimuli across repeated trials are typically aggregated for each participant, to yield a set of participant-level scores. These scores are then often used to examine group differences or to predict relevant behaviors.

Although such analysis techniques are easy to interpret because they yield a single "bias" score for each participant, they are likely to be problematic because they ignore systematic variation in the trial-by-trial responses due to the individual stimuli that are used. When systematic variation due to stimuli exists, ignoring stimulus variance is analogous to treating stimuli as a fixed factor rather than a random factor, meaning that results can be generalized to other samples of participants but only if the exact same stimuli are used (Clark, 1973; Judd, Westfall, & Kenny, 2012). What this means is that current analytic practices run the risk of finding statistically "significant" results that fail to replicate when different but comparable sets of stimuli are used in those implicit measures. Given the recent attention to failures to replicate experimental findings in psychology (Brandt et al., 2014; Earp & Trafimow, 2015; Francis, 2012; Galak, LeBoeuf, Nelson, & Simmons,

✉ Katie Wolsiefer
katherine.wolsiefer@colorado.edu

[1] University of Colorado, Boulder, CO, USA

[2] University of Texas, Austin, TX, USA

2012; Open Science Collaboration, 2015; Pashler & Harris, 2012; Pashler & Wagenmakers, 2012), it seems prudent for researchers working with stimulus samples to use statistical models that lead to appropriately conservative conclusions about the generalizability of findings to future participant and stimulus samples.

Our goal in this article is to consider in detail three of the most commonly used implicit attitude measures (Nosek, Hawkins, & Frazier, 2011): the Implicit Association Test (IAT), the affect misattribution procedure (AMP), and the evaluative priming task (EPT). For each of these measures we develop the appropriate analytic strategy that treats both stimuli and participants as random factors, thus modeling responses to these implicit measures as a function of both stimulus and participant variation. We then examine the magnitude of stimulus variance in each of the tasks and discuss, for each, the degree of bias that results when the traditional analyses that ignore stimulus variance are conducted.

Implicit attitude measures: analytic considerations

Abundant psychological literature has used implicit attitude measures to index implicit or automatic evaluations of many different attitude objects and to predict behavior from these. Clinical psychologists have used implicit attitude measures of fear associations to examine the efficacy of phobia interventions (Teachman & Woody, 2003) and implicit attitude measures of death/suicide associations to predict suicide attempts (Nock et al., 2010). In the realm of political psychology, undecided voters' voting behavior has been predicted using implicit policy attitude measures (Arcuri, Castelli, Galdi, Zogmaister, & Amadori, 2008). Measures of implicit attitudes are, perhaps, most prevalent in research on stereotyping and prejudice. Responses on the EPT, one common implicit attitude measure, have been shown to predict trait inferences for members of different racial groups (Olson & Fazio, 2004). Additionally, scores on implicit racial attitude measures are related to nonverbal behavior during interracial interactions (Dovidio, Kawakami, & Gaertner, 2002). In sum, implicit attitude measures have been used in many areas of psychology to measure many different kinds of implicit evaluations and they have demonstrated their validity and utility by predicting important behaviors.

All implicit tasks involve participants' trial-by-trial responses to a large number of presentations of different stimuli. Typically, for analysis purposes, those responses are reduced to a single score for each participant, such as mean latencies (or mean latency differences on different trials), proportions of responses of a given type across trials (or again differences in proportions), or some other aggregate score that collapses across trials. These scores are then used in further analyses, assessing group differences in them as a function of some independent variable or predicting some behavioral outcome from them. With few exceptions, the unit of analysis in these studies is the participant, and confidence intervals of any estimated effects of interest are derived from the variability in scores across participants.

However, the individual trials in implicit tasks are not exact replicates of each other, and therefore potentially important trial-to-trial variability in responses within a participant may exist. In fact, in nearly all implicit tasks, responses are given to particular stimuli that vary across trials. Such stimuli are thought to be instantiations of stimulus *categories* that really are the focus of theoretical interest. For instance, responses might be given to particular words of a positive or negative valence (e.g., horrible, terrible, pleasant, wonderful), particular category exemplars (e.g., specific Black or White faces, or names of Republican and Democratic politicians) or particular pronouns that are self-referencing or other-referencing (e.g., me, I, he, they). Again, what are of interest are responses to the general category thought to be instantiated by the particular stimuli used. In essence, the stimuli used in particular implicit tasks can be thought of as a sample of stimuli that might have been drawn from the categories that are actually of theoretical interest. Put differently, in principle, other stimuli could have been used that would have served the experimenter's purposes just as well as those that did happen to be used. To the extent that there is random variation across these possible stimulus samples, this random variation should be reflected in the standard errors of the parameter estimates of interest in the study.

As we described above, the typical analysis of data from implicit tasks takes into account the participant-to-participant variability in the data but ignores the variability associated with stimuli. In the language of analysis of variance (ANOVA), it treats participants as a random factor but stimuli as a fixed factor. In other words, participants are considered a sample of possible participants and statistical inference asks about generalization to other samples that might be used in future studies. But by ignoring variation associated with individual stimuli, they are treated as fixed rather than as random, meaning that the traditional analysis only permits generalization to future studies that use the exact same stimuli. If in fact one would like to generalize beyond the particular stimuli used in any one study to the category of stimuli as a whole that might have been used, then it can be shown that the traditional analyses, ignoring stimulus variance, leads to inappropriately small standard errors and, hence, inflated Type I error rates (Clark, 1973; Judd et al., 2012). This means that many effects reported in the literature may not be replicated when other stimulus samples are employed, even if the future stimulus samples are comparable in all theoretically and methodologically relevant ways to the original stimuli.

To permit generalization to studies involving different samples of both participants and stimuli, analyses need to treat both factors as random,[1] estimating appropriate error variances from variability in responses across both random factors. Until recently, such estimation was possible only in highly restricted experimental designs through the use of quasi-*F* ratios (Clark, 1973; Winer, 1971). More recently, a much more efficient and general approach has been outlined using what are called linear mixed models (Baayen, Davidson, & Bates, 2008; Judd et al., 2012). Such models have been relatively widely used in education and intervention research, in the context of hierarchically nested or multilevel research designs (e.g., Raudenbush & Bryk, 2001; Snijders & Bosker, 2011) in which the multiple random factors are nested under one another. For instance, in education, students within classrooms might be measured and both the students and the classrooms are appropriately considered to be random factors. In most implicit tasks, the two random factors of participant and stimuli are not nested, but rather are crossed, in that each participant gives responses to the same full set of stimuli across trials. Westfall, Kenny, and Judd (2014; see also Judd, Westfall, & Kenny, in press) provide details on model specification, appropriate standard errors, and power estimation for a range of designs involving crossed random effects of participants and stimuli.

Our specific goals

In this article, we consider three commonly used implicit evaluation measures in detail: the IAT (Greenwald, McGhee, & Schwartz, 1998), the AMP (Payne, Cheng, Govorun, & Stewart, 2005), and the EPT (Fazio et al., 1995) We used data from a large national dataset involving the responses of many participants to these three tasks in three different domains: implicit racial attitudes, implicit political attitudes, and implicit self-evaluations. In considering each of these three implicit tasks, our goals were threefold. First, we wanted to lay out the mixed-model specification that would permit an analysis of data from each of the tasks to appropriately treat both participants and stimuli as crossed random factors. An Appendix contains syntax for estimating these models in R, SAS, and SPSS, as well as instructions on how to structure the data file. Second, on the basis of the model estimation for the data at hand, we examined the extent to which variation due to the stimuli contributed to the responses in each task, thus permitting us to examine the extent to which the traditional analyses, which collapse across stimuli, result in bias (i.e., standard errors that are too small, and test statistics that as a result are too large). Finally, we examined individual participant scores from each task, using both the traditional analytic approach and the newer mixed-model estimation, to examine the extents to which the two approaches converge or diverge in indexing participant implicit evaluations in the three attitude domains.

[1] In saying that both participants and stimuli ought to be treated as random factors, we are not suggesting that in typical studies they are actually sampled randomly from defined populations. Rather, most typically the sample of stimuli used in any given study is a convenience sample, much like the sample of participants in nearly all studies in psychology is a convenience sample. The point is that by treating both participants and stimuli as random factors, one becomes able to generalize the results to other samples of both participants and stimuli, sampled in similar "convenient" ways in future studies.

Dataset and measures

Dataset

The data that we used come from Bar-Anan and Nosek (2014). They were collected on the Project Implicit website (implicit.harvard.edu) between November 6, 2007, and May 30, 2008, as part of a larger study. Participation in studies on this website is open to the public. The participants were visitors to the website who volunteered to be randomly assigned to complete one or more tasks from a set of implicit attitude measures, explicit attitude measures, and various other measures. Participants who did not complete an IAT, AMP, or EPT were excluded from our analysis. The initial dataset contained data from 24,015 participants. Of those, 11,878 participants had completed at least one of the three implicit measures of interest (i.e., the IAT, AMP, and EPT). The final sample sizes for each individual task are listed below.

Although this is a large and in many ways appealing dataset for our purposes, it is not without its limitations. Participants could complete multiple tasks, and these may have contaminated each other. Also, some parameters of the various tasks did not exactly mirror the procedures that the original creators of the task recommended, and some of the stimulus samples were unusual. We discuss these limitations in the Discussion section of this article. Nevertheless, for our purposes, this dataset seemed to us to be the best available, for reasons we also detail below.

Measures

Implicit association test

In the IAT, participants are asked to simultaneously sort two classes of stimuli (e.g., good/bad words and Black/White faces) sharing the same response keys. In the race IAT, for example, participants were asked to press one key each time they saw a Black face or a bad word and to press a different key each time they saw a White face or a good word. After completing a short (20 trials) and a long (40 trials) block in this configuration, participants completed two tasks with the

response mappings reversed (i.e., with Black faces and good words sharing a response key and White faces and bad words sharing a response key). Blocks in which the response mappings were consistent with cultural attitudes (Race: Black/bad and White/good; Politics:[2] Democrats/good and Republicans/bad; Self-Esteem: Self/good and Other/bad) will be referred to as "congruent," and blocks with the opposite response mappings will be referred to as "incongruent" (see Table 1). In this task, implicit attitudes were indexed by the difference in the speeds with which participants responded to trials in the incongruent versus the congruent blocks.

Data cleaning The IAT data were cleaned according to suggestions provided by Greenwald, Nosek, and Banaji (2003). First, trials in which participants made errors were identified and an error rate was calculated for each participant. Only four participants had error rates over 50 % across the three IATs. Given this small number, no participants were excluded. Next, we identified participants who had a high proportion of trials with extremely fast response times. Trials on which participants responded faster than 300 ms were identified, and a by-participant "fast rate" was calculated. In total, 194 participants (21 in the race IAT, 101 in the politics IAT, and 62 in the self-esteem IAT) had fast rates on more than 10 % of trials; these participants were excluded from the analysis. Those who did not complete all blocks (for race 93, for politics 140, and for self-esteem 143) were also excluded from the analysis. Finally, participants were excluded if the stimuli appeared to be inconsistent with the type of task they were recorded as having completed (one participant on the race IAT). This resulted in the following numbers of participants for each task—race IAT, 3,240; politics IAT, 3,084; self-esteem IAT, 3,080.

In addition to the excluded participants, trials were excluded if the response latency was extremely fast (<400 ms) or extremely slow (>10,000 ms). Response latencies were not adjusted for error versus nonerror trials. This is based on the fact that the response latencies included in the datasets were the times elapsed from presentation of the target until a *correct* response.

Scoring The standard individual-difference score yielded from an analysis of IAT data is an IAT D-score (Greenwald et al., 2003). IAT D-scores are calculated by first averaging the response latencies[3] for a given participant by block. The difference in the average response times for the congruent block is subtracted from the average response latency for the incongruent block for the short and long IAT blocks separately. Separate D-scores for the short and long blocks are then calculated by dividing each of the obtained differences by a participant's overall standard deviations across both incongruent and congruent blocks of the same length. Finally, these scores are averaged together to form the participant's overall score. Higher positive scores indicate higher levels of implicit positivity toward the normatively more positive attitude object category.

[2] Although it is clear in the case of the race and self-esteem categories which response corresponds to the attitude generally held by the culture, our designation of the Democrat/good, Republican/bad block as representing the "culturally held attitude" is arbitrary.

[3] The response latency in this task is defined as the time from stimulus onset to a participant's correct response on a given trial.

Affect misattribution procedure

In the AMP, participants are briefly exposed to a prime (e.g., a Black face) and then shown a neutral stimulus. Participants are asked to rate the neutral stimulus (e.g., a Chinese ideograph) as either pleasant or unpleasant relative to other Chinese ideographs. Participants were specifically instructed to ignore the primes: "evaluate each Chinese drawing and not the image that appears before it. The images are sometimes distracting." After a block with three practice trials, participants completed two 36-trial blocks. Each block contained 12 trials from each of the categories of interest (i.e., Black and White, Democrat and Republican, and Self and Other) and 12 trials with a control prime. On each trial, the prime was presented for 75 ms, and the Chinese ideograph was presented for 100 ms (followed by a pattern mask that remained on screen until participants had responded). Typically, implicit attitudes are characterized as the relative proportions of Chinese ideographs that are rated as pleasant following the presentation of one type of prime (e.g., a Black face) versus the proportion that are rated as pleasant following the other type of prime (e.g., a White face).

Data cleaning Participants who completed more than the specified 72 trials were excluded from the analysis. Participants were also excluded if the stimuli appeared to be inconsistent with the type of task they were recorded as having completed. Participants who responded in the same way on all trials (either pleasant or unpleasant) were also excluded. In addition to exclusions on a by-participant basis, trials involving the neutral stimulus were deleted for all participants. This resulted in three datasets with the following numbers of participants: race AMP, 3,276; politics AMP, 3,363; self-esteem AMP, 3,293.

Scoring Typical scoring of the AMP first involves calculating, for each participant, the proportion of trials on which a "pleasant" rating is made (relative to an "unpleasant" rating) for each prime category (e.g., Black vs. White faces). The means for the two prime categories are then compared using a within-subjects ANOVA with prime type as a predictor. We used both the raw and logit-transformed proportions in this analysis.

Table 1 Contrast codes by block type or trial type for each task, as well as the expected directions of these effects

Effect	Block Type (IAT) or Trial Type (AMP and EPT)						Expected Direction of Effect
	Good Word		Bad Word		No Word		
	White Dem. Self	Black Rep. Other	White Dem. Self	Black Rep. Other	White Dem. Self	Black Rep. Other	
Word Type							
Positive vs. Negative	1	1	–1	–1	NA	NA	NA
IAT Effects (block type)							
Congruent	–1	1	1	–1	NA	NA	+
AMP Effects (prime type)							
Black vs. White	1	–1	1	–1	1	–1	+
Dem. vs. Rep.	1	–1	1	–1	1	–1	NA
Self vs. Other	1	–1	1	–1	1	–1	+
EPT Effects (Prime Type × Target Type)							
Black/White × Good/Bad	–1	1	1	–1	NA	NA	+
Dem./Rep. × Good/Bad	–1	1	1	–1	NA	NA	NA
Self/Other × Good/Bad	–1	1	1	–1	NA	NA	+

Note that for the IAT, the effect of interest (congruent vs. incongruent) varies at the block level, whereas for the AMP and EPT, the effects of interest vary at the trial level. IAT = Implicit Association Test, AMP = affect misattribution procedure, EPT = evaluative-priming task, Dem. = Democrat, Rep. = Republican, Con. = congruent IAT block, Incon. = incongruent IAT block, NA = not applicable

Evaluative priming task

As in the AMP, participants who completed the EPT were first shown a prime (the attitude object) followed by a stimulus to which they were asked to respond. In the case of the EPT, the stimuli to which participants responded were positively and negatively valenced words that they were to categorize as either "good words" or "bad words." Participants were instructed to categorize the words as quickly as possible while making as few mistakes as possible. A 1,500-ms timeout window was given for each trial. Participants completed one block without primes (28 trials), followed by three blocks that included primes. Each of these blocks contained 60 trials (15 trials for each prime category/target combination). Each prime trial consisted of a 200-ms presentation of the prime, followed by a blank screen for 50 ms. Finally, the valenced word appeared until either a response was made or the 1,500-ms timeout was reached. A final block tested participants' memory of the stimuli and will not be analyzed or discussed further. Implicit attitudes are typically characterized as the relative difference in the speeds of sorting good versus bad words for one attitude object category compared to another.

Data cleaning Participants were excluded if they completed more than 60 trials in any of the three critical blocks. Participants were also excluded if the stimuli appeared to be inconsistent with the type of task they were recorded as having completed. In addition to these by-participant exclusions, trials were eliminated if the response was an error (e.g., categorizing a "good" word as "bad"), if the response latency was under 300 ms, or if the trial ended in a timeout (over 1,500 ms). These exclusions resulted in three datasets with the following numbers of participants: race EPT, 3,172; politics EPT, 3,060; self-esteem EPT, 3,286.

Scoring Implicit attitude scores are calculated from the EPT by first log-transforming the response latencies on all correct trials.[4] Next, four mean log latencies are calculated for each participant: one for each possible trial type (e.g., Black face/good words, Black face/bad word, White face/good word, White face/bad word). These means are then submitted to a 2 × 2 within-subjects ANOVA. The expectation is that, in a participant sample that has a relatively positive implicit attitude toward Whites relative to Blacks (on average), responses will be faster on trials with good words and White faces than on trials with bad words and White faces, and the reverse pattern will emerge for trials involving Black faces. Thus, a prime type by word valence interaction is the expected effect.

[4] We used a log transformation because this is standard practice in studies involving the EPT. We also fit a series of models using an inverse transformation (1/reaction time) to see whether this led to more normal distributions of the model residuals and random effects than did the log transformation, but we ultimately found that it made little difference for these data: The distribution of level-1 model residuals was perhaps closer to normal under the inverse transformation, but the distributions of random effects were not appreciably different, nor were the fixed- and random-effect results.

Models and analyses

The three goals of our article—deriving mixed models for three implicit attitude measures, estimating the consequences of ignoring stimulus variance in such tasks, and comparing traditional individual-difference estimates of implicit attitudes with estimates derived from mixed models—were accomplished using two sets of analyses. To address our first two goals, we estimated mixed-effects models that allowed effects of interest to vary by both participants and stimuli for each implicit measure. To achieve the third goal, we examined the degree to which allowing effects to vary on the basis of stimuli yielded participant-level "bias" scores that differed from the bias scores derived from the standard analyses.

One of the advantages of the dataset used is that the attitude object stimuli were consistent across the three types of implicit measures, thus permitting comparisons across the three measures based on the same actual stimuli. Accordingly, across measures, we used consistent contrast codes to differentiate the different categories of attitude objects (i.e., Blacks vs. Whites, Democrats vs. Republicans, and self vs. other). As is shown in Table 1, the normatively more positively viewed attitude objects (i.e., Whites, Democrats, self) were coded as +1, whereas the normatively negatively viewed attitude object stimuli (i.e., Blacks, Republicans, other) were coded as –1. Additionally, both the EPT and IAT included both positively and negatively valenced words. For models examining each of these tasks, +1 was used to denote positive words, and –1 was used to denote negative words. In the following sections we outline, for each task, the models used to estimate implicit associations, using both mixed-effects models that treat stimuli as random effects and more conventional methods for examining implicit associations.

Model estimation

Implicit association test

The mixed-model analysis of the IAT modeled the latency on each individual trial divided by the standard deviation across all trials in that block (i.e., across all long- or short-block trials, depending on the block type), following the recommendations of Greenwald, Nosek, and Banaji (2003). Each trial occurred in either a congruent or incongruent block and the specific target on each trial was either a category object (i.e., category exemplar) or a valenced word (i.e., positive or negative). The latencies were modeled as a function of whether the trial was in a congruent or incongruent block (–1 if congruent, +1 if not), and three contrast codes that coded the category of the specific stimulus used. The first of these coded whether the stimulus on the trial was a category object or a valenced word; the second coded the category of the attitude object, given that the trial presented a category of attitude object, using the codes defined earlier; and the third coded the word valence, given that the trial presented a valenced word, again using the codes defined in Table 1. Additionally, three interaction terms were included in the model, involving products of congruency block code with the three codes that coded the specific stimulus type used on each trial, in order to test whether stimulus type moderated the congruency effect. Accordingly, including the intercept, a total of eight fixed effects were included in these models.

The best way to estimate the random effects for the IAT model was not immediately obvious. Participants responded to both valenced words and category exemplars, but these stimuli appeared in separate, alternating trials. Thus, random stimulus effects could reasonably be separated out into two different random factors. Additionally, since responses to both valenced word stimuli and category exemplar stimuli were crossed with block type (i.e., congruent/incongruent), it would be possible to estimate separate random congruency slopes due to word stimulus and to exemplar stimulus. Ultimately, we chose to combine these two potential random factors into one overall Stimulus random factor that consisted of both valenced words and category exemplars. Combining these two types of stimuli into a single random factor allowed for a simpler, easier-to-estimate model requiring fewer parameter estimates. However, before combining these random factors into one, we did examine the degrees to which the variance components for valenced words and category exemplars were different. Likelihood ratio tests comparing the models we report here to models that would allow the two stimulus types to have different random variances provided little statistical evidence that the stimulus types had any meaningful difference in their random variances [the likelihood ratios for these comparisons in the race, politics, and self models were, respectively, $\chi^2(1) = 0.99$, $p = .319$; $\chi^2(1) = 3.34$, $p = .068$; and $\chi^2(1) = 3.82$, $p = .051$].

The so-called *maximal* model for this task would involve estimating many parameters (Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015). Since participant was crossed with word type, face type, and congruency, all of the fixed effects in the IAT model could be allowed to vary randomly across participants. Since stimuli were responded to in both congruent and incongruent blocks, the effect of congruency could vary randomly across the stimuli. And because each participant responded to each stimulus multiple times in the congruent block and multiple times in the incongruent block, it was possible estimate random intercepts and congruency slopes for the participant-by-stimulus interaction. However, fitting such a model would be extremely computationally challenging, given the very large number of parameters to be estimated. In fact, in our case the maximal model failed to converge, even with the relatively large datasets we were using.

We instead recommend a more simplified model, the parameters of which are shown in Table 2. In this model, we

Table 2 Implicit Association Test (IAT) mixed-model results

	Race		Politics		Self	
IAT fixed effects	Coefficient	*t*	Coefficient	*t*	Coefficient	*t*
Intercept	2.45	124.07***	2.51	102.7***	2.57	104.44***
Congruent	0.14	31.09***	0.13	23.16***	0.23	36.42***
Word vs. Face	–0.12	–8.29***	0.08	3.67***	~0	–0.13
Word type	~0	–0.05	0.04	0.84	–0.01	–0.3
Face type	0.04	1.8†	0.02	0.64	0.08	2.59**
Congruent × Word vs. Face	–0.04	–10.67***	–0.02	–6.76***	–0.04	–7.29***
Congruent × Word Type	–0.02	–4.12***	~0	–0.54	–0.01	–1.26
Congruent × Face Type	0.01	1.65†	~0	–0.83	–0.01	–1.71†
IAT random effects	*SD*	% of variance	*SD*	% of variance	*SD*	% of variance
Participant						
Intercept	0.75	33.1 %	0.71	31.14 %	0.71	30.79 %
Congruent	0.17	1.7 %	0.25	3.94 %	0.14	1.28 %
Correlation	(–.09)	–	(–.03)	–	(–.05)	–
Stimulus						
Intercept	0.07	0.3 %	0.08	0.4 %	0.1	0.6 %
Congruent	0.01	0.01 %	0.01	0.01 %	0.02	0.04 %
Correlation	(–.28)	–	(–.43)	–	(0)	–
Residual	1.05	64.89 %	1.02	64.52 %	1.04	67.3 %

Note that this model also estimates all applicable covariances between the random effects, although these parameter estimates are not shown in the table. The percentages of variance for the random effects (i.e., the variance-partitioning coefficients, or VPCs) were computed by dividing each variance component by the total random variance, computed as the sum of the random-effect variances (Westfall, 2015; Westfall et al., 2014). *t* = *t* statistic associated with each regression coefficient, *SD* = standard deviation. † $p < .1$, ** $p < .01$, *** $p < .001$

estimated random intercepts, random congruency slopes, and their covariances for both participants and stimuli. This model allowed for the estimation of the random effects that are of particular interest in most cases, as well as being more likely to converge when using smaller datasets. The statistical rationale behind simplifying the model in this particular way was that this model contains all and only the random effects that contribute substantially to the estimation of the standard error for the fixed congruency slope (Barr et al., 2013; Westfall, 2015), which is typically the test of interest in the IAT.[5]

The standard method for analyzing IAT scores, the D-score (Greenwald et al., 2003), is essentially a within-participant comparison of incongruent and congruent blocks. Thus, the average D-score within a sample gives an estimate of the extent to which responses are faster in one type of block than in the other, across participants and ignoring stimuli. The analogous linear mixed-model effect is the fixed effect of block type, since it will examine, on average across participants and stimuli, the difference in response times for congruent and incongruent blocks. The important distinction is that the mixed-effects model estimate will also control for random stimulus effects and account for their variance. That is, the model partials out any variance due to the stimuli. If stimulus variance really does contribute to an inflated test statistic in the case of the traditional D-score analysis, then the test statistic for the single-sample *t* test of the average D-score, relative to 0, should be higher than that of the corresponding effect in the mixed-effects model. Additionally, the estimate of the degree to which the effect of block varies from stimulus to stimulus should be larger to the extent that the stimulus variance is more heavily influencing the outcomes with regard to the traditional analysis.

Affect misattribution procedure

The traditional analysis for the AMP involves computing for each participant the simple proportion for each type of prime (e.g., Black vs. White faces) for which the subsequently presented target Chinese character is judged as pleasant (rather than unpleasant). Then the means of these are compared, as a function of prime type, using a within-participant ANOVA.

Since mean proportions are calculated for each category of prime, the traditional analysis collapses across stimulus types. To examine the extent to which stimulus variance may

[5] To verify that none of the variance components omitted from the model would in fact substantially alter the test statistics for the fixed congruency effects, we also estimated models that included random slopes for all of the simple effects in the model (but not the interactions, since the resulting maximal models usually failed to converge). The test statistics for the congruency effects in these expanded models were negligibly different from those in the simpler models; they differed by less than 1 %.

influence AMP effects, mixed-effects models that modeled the likelihood of categorizing an ideograph as positive on each trial were used (see the Appendix for example data structures). The by-trial analysis allowed us to treat both participants and stimuli as random factors. Since the outcomes for these models were dichotomous, the analysis was conducted using a generalized linear mixed model with a logistic link function.

This mixed model contains two random stimulus factors, since on each trial both a prime stimulus (e.g., a White face or a Black face) and a target stimulus (i.e., a specific Chinese character about which a judgment is given) were presented. In total, 12 different primes (e.g., six White faces and six Black faces) and 72 different Chinese characters were used as target stimuli. The mixed model that was estimated (again with a logistic link function) allowed random intercepts for the prime stimuli and random intercepts and prime type slopes for the target stimuli.

It should be noted that, once again, we did not estimate the maximal model with all possible random effects. For example, since participants completed trials with different pairings of the target and prime stimuli (where the target and prime stimuli were both repeated within participants), it would be possible to estimate several types of random interaction effects. However, in the interest of parsimony (and model convergence), we estimated only the random effects that would contribute substantially to the standard error of the fixed effect of interest.

Evaluative priming task

The traditional analysis of data from the EPT involves a log transformation of the latencies from correct trials. For each participant, four mean log latencies are computed: one for each cell of the crossed target (positive vs. negative) and prime (e.g., Black vs. White faces) factors. Implicit associations are conceptualized as the prime by target interaction resulting from the within-subjects ANOVA of these log latencies is the effect of interest. The expectation is that congruent means (e.g., White primes with good target and Black prime with bad target) will be smaller (i.e., shorter latencies) than the incongruent combination means.

Comparing means for the four types of trials in a within-participant ANOVA collapses across stimulus type and removes the possibility of examining whether the implicit associations measured depend on the stimulus. The estimation of a mixed-effects model addresses this issue. For the mixed-model analyses, we modeled the log-transformed latencies at the level of the individual trial, with contrast codes for prime (i.e., attitude object category), target valence (i.e., word valence), and their interaction (see Table 1). Modeling latencies at the level of the trial rather than the participant allowed for the preservation of information regarding different response patterns to different types of stimuli across participants. This allowed for the treatment of both prime stimulus and target stimulus as random factors. Accordingly, we estimated random intercepts and target type slopes for the prime stimuli, and random intercepts and prime type slopes for the target stimuli.

As with the IAT and AMP, the maximal model could have contained random participant-by-stimulus interaction effects, because each participant saw each stimulus multiple times with different types of target stimuli. However, these more complex models were avoided, since they would be much more computationally challenging to fit, and because these random participant-by-stimulus effects would not contribute to the standard error of the fixed effect of interest to any meaningful extent (Westfall, 2015).

Results

Comparing traditional analyses to mixed-model analyses

The parameter estimates for all nine mixed models are given in Tables 2, 3, and 4. For each model, we extracted the relevant test statistic (t or z statistic) that tested whether there was a nonzero implicit preference, on average—for example, in the IAT tasks, this would be the test statistic associated with the simple effect of congruency—and compared these to the corresponding test statistics based on the traditional analyses, which ignore random stimulus variability. This comparison is illustrated in Fig. 1, which shows that in all nine cases the traditional test statistics are inflated to some extent. A linear regression predicting the traditional test statistics from the corresponding mixed-model test statistics yielded a regression line that essentially passes through the origin, with an estimated slope of 1.60, indicating that the traditional test statistics are inflated by an average of about 60 % in the large datasets that we examined (95 % CI = [1.11, 2.09]).

It is interesting that the inflation of the test statistics is so substantial, despite the fact that the stimulus variance components themselves are all quite small in comparison to the other variance components. The explanation for this is that the degree of inflation depends not only on the size of the stimulus variance components that are being ignored in the traditional analyses, but also on the sample sizes. For example, the deleterious effect of stimulus variance on the traditional analyses is attenuated when there are more stimuli. In the present case, although the stimulus variances are small, the stimulus sample sizes are also small: 10 to 28 in most tasks, with the notable exception that the AMP uses a sample of 72 target stimuli (Chinese ideographs). Because of these small sample sizes, the test statistics can be substantially inflated even when the degree of stimulus variability is relatively small in comparison to the other variance components.

Table 3 Affect misattribution procedure (AMP) mixed-model results

	Race		Politics		Self	
AMP fixed effects	Coefficient	*z*	Coefficient	*z*	Coefficient	*z*
Intercept	0.73	15.64***	0.34	5.92***	0.62	14.22***
Prime type	–0.16	–6.07***	0.24	4.93***	0.08	5.63***
AMP random effects	*SD*	% of variance	*SD*	% of variance	*SD*	% of variance
Participant						
Intercept	1.03	22.7 %	0.9	17.32 %	0.9	18.59 %
Prime type	0.48	5 %	0.71	10.84 %	0.34	2.71 %
Correlation	(–.27)	–	(–.04)	–	(–.01)	–
Target						
Intercept	0.29	1.82 %	0.27	1.52 %	0.33	2.51 %
Prime type	0.01	0 %	0.01	0 %	0.01	0 %
Correlation	(–.23)	–	(1)	–	(1)	–
Prime						
Intercept	0.09	0.15 %	0.14	0.44 %	0.03	0.03 %
Residual	1.81	70.32 %	1.81	69.88 %	1.81	76.15 %

Note that this model also estimates all applicable covariances between the random effects, although these parameter estimates are not shown in the table. This is a mixed-effects logistic regression model, so all of the parameter estimates are on the logit scale. The percentage of variance for the random effects (i.e., the variance-partitioning coefficient, or VPC) depends on the variance of the latent residual term, which is not identifiable, so we have fixed its standard deviation to $\pi^2/\sqrt{3}$ (see Goldstein, Browne, & Rasbash, 2002, p. 228). The percentage of variance was then computed by dividing each variance component by the sum of the random-effect variances, including the residual variance. *t* = *t* statistic associated with each regression coefficient, *SD* = standard deviation. *** $p < .001$

Comparing variance components across tasks

As we mentioned earlier, each attitude category was represented by the same words and photographs across all three tasks. This permitted meaningful comparisons of the amounts of stable stimulus variance across tasks. For example, since participants responded to the same stimuli when taking the IAT and the EPT, larger stimulus variance components in one of these tasks

Table 4 Evaluative priming task (EPT) mixed-model results

	Race		Politics		Self	
EPT fixed effects	Coefficient	*t*	Coefficient	*t*	Coefficient	*t*
Intercept	6.455	934.72***	6.454	919.89***	6.488	1,024.29***
Prime type	0.002	2.62**	~0	–0.14	–0.002	–1.73†
Target type	0.019	3.48***	0.014	2.54*	0.014	2.8**
Prime × Target	0.002	0.99	0.007	3.21**	0.009	13.99***
EPT random effects	*SD*	% of variance	*SD*	% of variance	*SD*	% of variance
Participant						
Intercept	0.245	49.50 %	0.238	47.42 %	0.224	45.62 %
Prime × Target	0.011	0.1 %	0.018	0.27 %	0.01	0.1 %
Correlation	(–.15)	–	(–.02)	–	(.02)	–
Target						
Intercept	0.028	0.65 %	0.028	0.64 %	0.026	0.6 %
Prime type	0	0 %	0	0 %	0.002	0 %
Correlation	(–1)	–	(1)	–	(.33)	–
Prime						
Intercept	0.003	0.01 %	0.006	0.03 %	0.004	0.01 %
Target type	0.006	0.03 %	0.006	0.03 %	0.001	0 %
Correlation	(.81)	–	(.89)	–	(.69)	–
Residual	0.245	49.71 %	0.248	51.61 %	0.243	53.67 %

Note that this model also estimates all applicable covariances between the random effects, although these parameter estimates are not shown in the table. The percentages of variance for the random effects (i.e., the variance-partitioning coefficients, or VPCs) were computed by dividing each variance component by the total random variance, computed as the sum of the random-effect variances (Westfall, 2015; Westfall et al., 2014). *t* = *t* statistic associated with each regression coefficient, *SD* = standard deviation. † $p < .1$, * $p < .05$, ** $p < .01$, *** $p < .001$

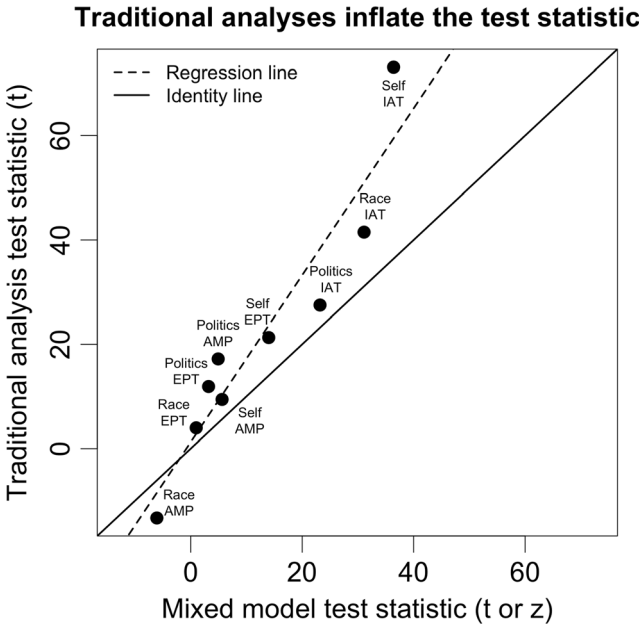

Fig. 1 Comparison of test statistics (*t* or *z*, depending on the task) of participants' average implicit preferences, taking into account random stimulus variability, with the corresponding test statistics based on traditional analyses that ignore stimulus variability

would suggest that the responses in that task were relatively more affected by idiosyncratic stimulus characteristics than those in the other task. The participant variance components were also of interest, since the participant variance is essentially the "signal" that these implicit attitude tasks are designed to measure. For example, the race IAT aims to measure individual differences in people's implicit preferences toward White or Black people, and the variance of those individual differences is directly estimated by the variance of the "congruent" slopes across participants in the mixed model. Thus, comparing the sizes of the participant variance components across tasks allowed us to compare the relative sensitivities of the different tasks in picking up on the relevant psychometric signal.

In Fig. 2, we plot the participant and stimulus variance components for each task and category. Only the variance components that are statistically relevant to the implicit preference being measured (e.g., in the case of the IAT, those that appear in the standard error of the fixed congruency effect) are shown in the figure; more detailed information is given in the figure legend. To facilitate comparisons, the variance components are computed as variance-partitioning coefficients (VPCs; Goldstein, Browne, & Rasbash, 2002; Westfall, 2015; Westfall et al., 2014)—that is, as proportions of the total random variation due to each variance component.

The pattern revealed in Fig. 2 is that the IAT and EPT both seem to involve relatively little stimulus variance, whereas the AMP involves relatively more stimulus variance. As for the participant variance, the AMP consistently shows the most stable participant variance, followed by the IAT, and with the EPT a fairly distant third. In fact, the degree of stable participant variance in the EPT is low enough to be a cause for concern about its psychometric reliability.[6] A comparison of the IAT and AMP does not show either one to be clearly superior in these data; the IAT seems to involve relatively less noise due to the stimuli, but also less signal due to participants, whereas the AMP involves relatively more noise due to the stimuli, but also more signal due to participants.

Results: comparing individual-difference estimates

Next, we wished to compare individual-difference estimates for the traditional and mixed-effects model analyses, to determine whether the estimates differed in any systematic way. However, whereas the unit of analysis using traditional scoring is the individual-difference estimate itself (i.e., D-scores for the IAT, differences in proportions for the AMP, and mean differences between the congruent and incongruent trial types in the EPT), the mixed models do not immediately offer individual-difference estimates for each participant.

It turns out that it is possible to extract individual-difference estimates, known as best linear unbiased predictors (BLUPs), from the mixed-effects models (Kliegl, Wei, Dambacher, Yan, & Zhou, 2011). In this case, each BLUP represents the slope of interest for a given participant, centered around the overall fixed effect and "shrunk" toward the fixed effect in proportion to the precision with which the participant's random slope is estimated. In the case of the IAT, each BLUP is the estimate of the congruency slope for a given participant, centered around the overall fixed congruency slope; in the case of the AMP, each BLUP represents a participant's slope for prime type; and in the case of the EPT, each BLUP is an estimate of a participant's prime type by target type interaction.

Figure 3 shows scatterplots of the traditional individual-difference scores and BLUPs for each type of implicit measure analyzed. First, the relationship between traditional scores and BLUPs is quite strong across all measures, with correlations between the two individual-difference estimates ranging from .87 to .99. Overall, there is strong agreement between the

[6] As we detail in the Discussion section, the limitations of the dataset and the procedures used for the measures may in part be responsible for some of these results.

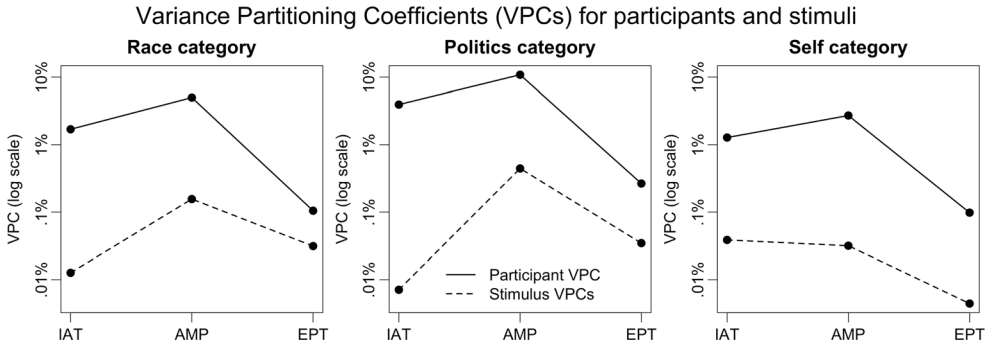

Fig. 2 Variance-partitioning coefficients (VPCs), computed as the proportions of the total random variation due to each variance component—that is, by dividing each variance component by the sum of the random-effect variances (Westfall, 2015; Westfall et al., 2014). Only the variance components that are statistically relevant to the implicit preference being measured are shown in the figure. For the Implicit Association Test (IAT), the participant variance is the variance of the participant congruency slopes, and the stimulus variance is the variance of the stimulus congruency slopes. For the affect misattribution procedure (AMP), the participant variance is the variance of the participant prime type slopes, and the stimulus variance is the sum of the variances of the target prime type slopes and the prime intercepts. For the evaluative-priming task (EPT), the participant variance is the variance of the participant Prime Type × Target Type slopes, and the stimulus variance is the sum of the variances of the target prime type slopes and the prime target type slopes

traditional implicit scores and implicit scores that account for random effects due to the stimulus.

Despite this overall strong agreement, differences in the correlation values as a function of the type of implicit task are striking. The correlations are uniformly high and the scatterplots as expected for the IAT and AMP tasks, with somewhat lower (but still substantial) correlations for the EPT. Thus, although the traditional scores and BLUPS are nearly identical for the IAT and AMP, there is a somewhat smaller degree of correspondence for the EPT.

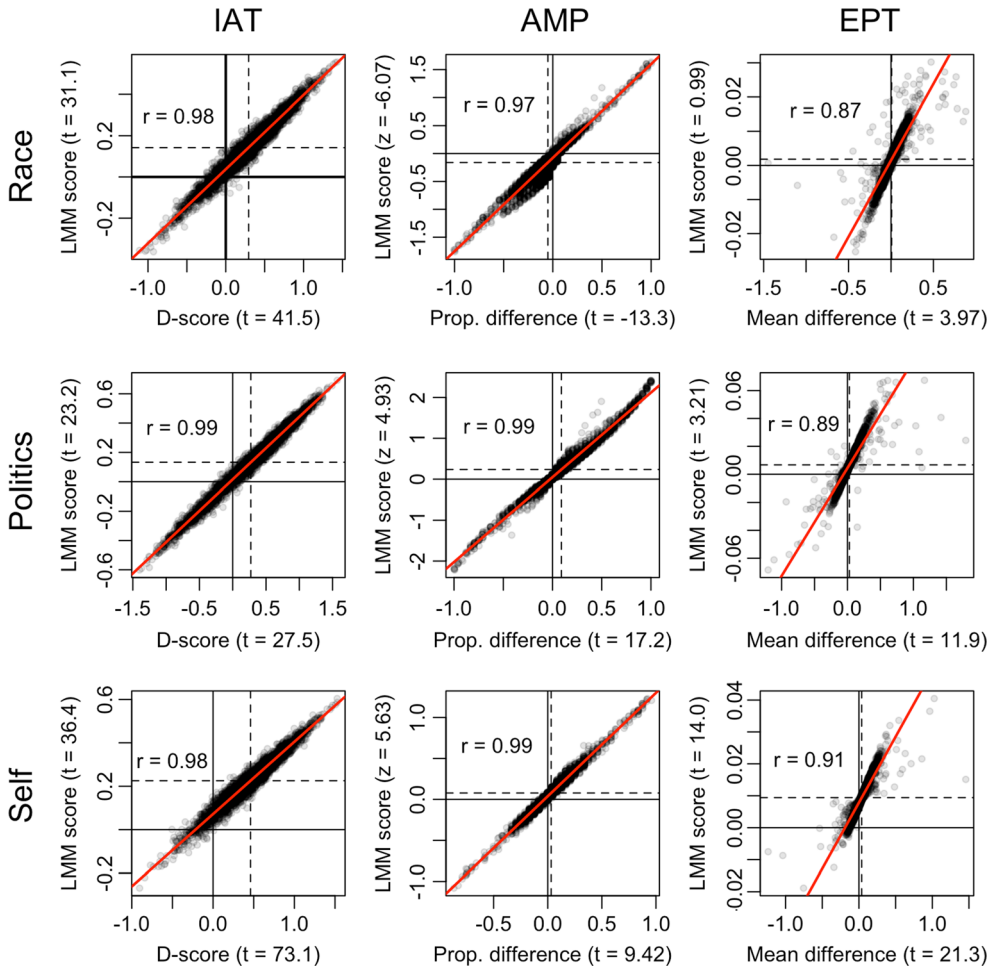

Fig. 3 Comparison of individual-difference scores derived from linear mixed models (LMMs) to individual-difference scores computed in the traditional fashion, for each task and attitude domain. There is generally strong but not perfect agreement between the LMM scores and the traditional scores. Prop. difference = difference in proportions

In sum, it is possible to obtain individual-difference estimates of implicit associations even when accounting for stimulus variance. In fact, the BLUPs look quite similar to those from traditional methods of computing implicit association scores. With that said, there are some differences between BLUPs and traditional scores, as evidenced by the fact that their correlations were less than unity. In particular, the BLUPs obtained from the EPT tend to be less closely related to the estimates obtained from the more traditional forms of estimation.

Discussion

We were able to successfully construct models that accounted for stimulus variance by using linear and generalized linear mixed-effects models that treat stimulus as a random factor. Comparisons of these models to models that did not account for stimulus variance demonstrated that failing to account for stimulus variance inflates the magnitude of the test statistic for all three implicit measures, the IAT, AMP, and EPT. Such inflation was expected because the significance tests for the new models focus on generalization across future studies with different stimulus samples as well as with different participant samples. For both the AMP and EPT, the linear and generalized linear mixed-effects models used the same fixed effects as the traditional analysis. Although models that do not account for stimulus variance have an advantage with regard to simplicity of estimation and interpretation, the inflation of the test statistics for these models relative to those that do account for stimulus variance highlight the importance of accounting for variance due to stimuli.

It is important to note that the magnitude of these differences varies consistently across the three different implicit tasks, with the least inflation in general for the IAT and the largest for the EPT. Across the three stimulus domains on average, the *t*-statistic for the EPT traditional analysis was 308 % the size of the mixed-model *t* statistic. The inflation of the *t* statistic was smaller for the AMP, with the traditional *t*-statistic being 245 % the size of the mixed-model *t* statistic, on average. Finally, the IAT demonstrated the smallest inflation of the *t*-statistic from the linear mixed models to the traditional model, with the traditional *t* statistic being, on average, 171 % of the mixed-model *t* statistic. Clearly the inclusion of stimulus random factors in the EPT and AMP analyses makes a much bigger difference that it does in the case of the IAT. This conclusion seems likely due to the fact that in the IAT on each trial the stimulus and word category labels are continuously present, but in the case of the EPT and AMP, participants are never prompted to think about the categorization of the prime stimuli.

Perhaps surprisingly, the typical results for the implicit tasks involving race were not robust in these data, in spite of the large participant sample sizes used and in spite of the frequency with which these implicit tasks have been used in the literature to assess implicit racial attitudes. That is, the expected implicit preference for White over Black targets does not emerge for all tasks and for all models. In fact, for the AMP the typical congruence effect is reduced and for the EPT it is quite small in absolute magnitude. In the models with stimulus variance components these test statistics are further reduced, leading even to statistical nonsignificance in the case of the EPT. Only in the IAT does this implicit racial bias effect seem robust. This is true in spite of the fact that implicit participant scores for race were reasonably highly correlated across implicit tasks. One wonders whether the choice of stimuli (i.e., professional basketball players) for this task may have influenced these results particularly on those tasks in which category labels were not presented alongside the category facial photographs.

It should be noted that the purpose of this article is not to determine which implicit attitude measure is the "best," but rather to demonstrate how stimulus variance may be modeled in these tasks and to compare the relative contributions of stimulus variance to the test statistics estimated using standard analysis techniques. We refrain from making generalized comparisons of measure quality for several reasons. First, the role of stimulus variance in estimating implicit attitudes should not be the only metric considered when deciding which implicit attitude measure to use. Second, some limitations with regard to the dataset used for this article discourage such comparisons. The status of the Project Implicit website as a well-known attitude measurement website may have impacted participants' responses as could the possibility that the same participants completed the same tasks under multiple study sessions. Further, some concessions were made with regard to the parameters of the implicit measures in order to reduce participant fatigue. For example, the inter-trial interval used for the EPT was considerably shorter than what is typically used for this task. Such factors may severely limit our ability to draw strong conclusions about other psychometric properties that would be important for making strong comparisons across tasks. Finally, in terms of stimulus variance, the "best" measure may depend on one's purpose. If the goal is to measure broad-level category-based attitudes (as is often the case in studies of race or political preference), then a measure with relatively

little stimulus variance may be optimal. However, if one is interested in measuring attitudes toward individuals, a measure that is sensitive to the stimuli used may be more helpful.

In sum, we were able to successfully construct models that account for stimulus variance in three common implicit measures. Comparisons of traditional analysis to analyses that account for stimulus variance overwhelmingly provided evidence that ignoring stimulus variance artificially inflates the value of the test statistic of interest in the IAT, AMP, and EPT. Overall, point estimates of implicit associations that account for stimulus variance appeared quite similar to traditional estimates, although this was somewhat less true for IAT scores. If the goal of using implicit measures is to assess implicit associations toward attitude objects at a broad, categorical level, it is important to account for the fact that we, as researchers, sample from a population of stimuli just as we sample from a population of students. If we desire to draw inferences past the set of stimuli to which participants respond within an implicit measure, it is important to treat stimuli as random factors in the design. As illustrated above, failing to account for random stimulus variance when that is the intent results in an inflation of the test statistic and the increased potential for a Type I error. Although only one effect presented above was reduced to non-significance in this study, it should be noted that the sample sizes are much larger than those of typical studies that use implicit measures. It is the hope of the authors that greater consideration of stimulus variance will be given in studies involving implicit measures or, more broadly, studies that involve repeated responses to multiple stimuli.

Appendix: estimating the mixed models in R, SAS, and SPSS

Data preparation

We assume that the data are loaded in the "long" format (i.e., one row per trial response) and are stored in a data set named dat. RTD is the trial response time divided by the standard deviation across the short- or long-block IAT trials (Greenwald, Nosek, & Banaji, 2003), logRT is the natural logarithm of the trial response time, and the response itself (a column of 0 or 1 values indicating which response was made) is named Y. The columns of identifying labels for the participants, stimuli, primes, and targets are named, respectively, participant, stimulus, prime, and target. The predictors are defined as in Tables 1, 2, 3, and 4 of the main text; importantly, they are assumed to be entered as numeric variables consisting of contrast code values, *not* as a factor (in R), class (in SAS), or string (in SPSS) variable. Examples of how the dataset should be structured for each analysis are shown in Tables 5, 6, and 7.

Syntax

R

Mixed models can be fit in R using the lme4 and lmerTest packages (Bates, Maechler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2015), the latter of which tests fixed effects using the Satterthwaite approximate degrees of freedom. Before running any of the models below, the user must first install and load these packages using the following commands:

```
install.packages(c("lme4","lmerTest"))
library("lme4")
library("lmerTest")
```

IAT model

```
install.packages(c("lme4","lmerTest"))
library("lme4")
library("lmerTest")
```

```
model <- glmer(Y ~ primeType + (primeType | participant) +
   (1 | prime) + (primeType | target),
   data=dat, family=binomial(logit), verbose=3)

summary(model)
```

AMP model

```
model <- glmer(Y ~ primeType + (primeType | participant) +
   (1 | prime) + (primeType | target),
   data=dat, family=binomial(logit), verbose=3)

summary(model)
```

EPT model

```
model <- lmer(logRT ~ primeType*targetType +
   (primeType:targetType | participant) + (targetType | prime) +
   (primeType | target), data=dat, verbose=3,
   control=lmerControl(optCtrl = list(maxfun = 10000)))

summary(model)
```

SAS

IAT model

```
proc mixed covtest data=dat;
class participant stimulus;
model RTD = congruent|wordVsFace + congruent|wordType +
   congruent|stimType / ddfm=satterth solution;
random intercept congruent / sub=participant type=un;
random intercept congruent / sub=stimulus type=un;
run;
```

AMP model

```
proc glimmix asycov data=dat;
class Y participant prime target;
model Y(event="1") = primeType / dist=binary solution;
random intercept primeType / sub=participant type=un;
random intercept / sub=prime;
random intercept primeType / sub=target type=un;
run;
```

EPT model

```
proc mixed covtest data=dat;
class participant prime target;
model logRT = primeType|targetType / ddfm=satterth solution;
random intercept primeType*targetType / sub=participant
   type=un;
random intercept targetType / sub=prime type=un;
random intercept primeType / sub=target type=un;
run;
```

SPSS

IAT model

```
mixed RTD with congruent wordVsFace congruent*wordVsFace
/fixed = congruent wordVsFace congruent*wordVsFace
/print solution testcov
/random = intercept congruent | subject(participant) covtype(un)
/random = intercept congruent | subject(stimulus) covtype(un).
```

AMP model

```
To the authors' knowledge, there is not a way to estimate a
logistic mixed effects model with crossed random factors in SPSS.
```

EPT model

```
mixed logRT with primeType targetType primeType*targetType
/fixed = primeType targetType primeType*targetType
/print solution testcov
/random = intercept primeType*targetType | subject(participant)
   covtype(un)
/random = intercept targetType | subject(prime) covtype(un)
/random = intercept primeType | subject(target) covtype(un).
```

Table 5 Example of how the dataset should be structured for fitting the IAT model

Participant	Stimulus	Congruent	Wordvsface	Wordtype	Facetype	RTD
p1	Wonderful	–1	–1	–1	0	2.38
p1	Poison	–1	–1	1	0	2.46
p1	mtmmbm2.jpg	–1	1	0	–1	2.01
p1	mtmmwm1.jpg	–1	1	0	1	1.83
p1	Love	1	–1	–1	0	1.71
p1	Grief	1	–1	1	0	1.58
p1	mtmmbf1.jpg	1	1	0	–1	1.17
p1	mtmmwm2.jpg	1	1	0	1	1.47
p2	Paradise	–1	–1	–1	0	1.57
p2	Poison	–1	–1	1	0	1.21
p2	mtmmbm3.jpg	–1	1	0	–1	1.37
p2	mtmmwf2.jpg	–1	1	0	1	1.96
p2	Cheer	1	–1	–1	0	1.35
p2	Bomb	1	–1	1	0	1.16
p2	mtmmbf1.jpg	1	1	0	–1	1.50
p2	mtmmwf3.jpg	1	1	0	1	1.30

Table 6 Example of how the dataset should be structured for fitting the AMP model

Participant	Prime	Target	Primetype	Y
p1	epbf3	ampt14	–1	1
p1	epbm1	ampt33	–1	0
p1	epbf3	ampt19	–1	0
p1	epbm2	ampt16	–1	1
p1	epwf1	ampt39	1	0
p1	epwf3	ampt10	1	1
p1	epwm3	ampt25	1	0
p1	epwf2	ampt35	1	0
p2	epbm2	ampt56	–1	1
p2	epbm1	ampt26	–1	1
p2	epbm3	ampt22	–1	1
p2	epbf2	ampt5	–1	0
p2	epwf3	ampt7	1	1
p2	epwm3	ampt2	1	1
p2	epwm2	ampt21	1	1
p2	epwf3	ampt9	1	1

Table 7 Example of how the dataset should be structured for fitting the EPT model

Participant	Prime	Target	Primetype	Targettype	LogRT
p1	mtmmbm3	Annoying	–1	–1	6.50
p1	mtmmbm2	Stink	–1	–1	6.85
p1	mtmmbf1	Beautifu	–1	1	6.60
p1	mtmmbm1	Likeable	–1	1	6.36
p1	mtmmwf1	Offensiv	1	–1	7.11
p1	mtmmwm2	Abuse	1	–1	6.47
p1	mtmmwm1	Cheer	1	1	6.50
p1	mtmmwf3	Love	1	1	6.68
p2	mtmmbm2	Ugly	–1	–1	6.13
p2	mtmmbm1	Abuse	–1	–1	6.63
p2	mtmmbm3	Glee	–1	1	6.29
p2	mtmmbf3	Enjoy	–1	1	6.49
p2	mtmmwm3	Offensiv	1	–1	7.05
p2	mtmmwf1	Stink	1	–1	6.76
p2	mtmmwm1	Wonderfu	1	1	6.05
p2	mtmmwm3	Splendid	1	1	6.81

References

Arcuri, L., Castelli, L., Galdi, S., Zogmaister, C., & Amadori, A. (2008). Predicting the vote: Implicit attitudes as predictors of the future behavior of decided and undecided voters. *Political Psychology, 29,* 369–387. doi:10.1111/j.1467-9221.2008.00635.x

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. doi:10.1016/j.jml.2007.12.005

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods, 46,* 668–688. doi:10.3758/s13428-013-0410-6

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. Unpublished manuscript. Preprint at http://arxiv.org/abs/1506.04967

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. doi:10.18637/jss.v067.i01

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. doi:10.1016/j.jesp.2013.10.005

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359. doi:10.1016/S0022-5371(73)80014-3

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Interpersonal Relations and Group Processes, 82,* 62–68.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Quantitative Psychology and Measurement, 6,* 621. doi:10.3389/fpsyg.2015.00621

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. doi:10.1037/0022-3514.69.6.1013

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54,* 297–327. doi:10.1146/annurev.psych.54.101601.145225

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19,* 975–991. doi:10.3758/s13423-012-0322-y

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology, 103,* 933–948. doi:10.1037/a0029709

Gawronski, B., & Payne, B. K. (Eds.). (2010). *Handbook of implicit social cognition: Measurement, theory, and applications* (1st ed.). New York, NY: Guilford Press.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics, 1,* 223–231. doi:10.1207/S15328031US0104_02

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27. doi:10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (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. doi:10.1037/0022-3514.85.2.197

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54–69. doi:10.1037/a0028347

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology.*

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology, 1,* 238. doi:10.3389/fpsyg.2010.00238

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models. R package version 2.0-29. Retrieved from https://CRAN.R-project.org/package=lmerTest

Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. *Psychological Science, 21,* 511–517. doi:10.1177/0956797610364762

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences, 15,* 152–159. doi:10.1016/j.tics.2011.01.005

Olson, M. A., & Fazio, R. H. (2004). Trait inferences as a function of automatically activated racial attitudes and motivation to control prejudiced reactions. *Basic and Applied Social Psychology, 26,* 1–11. doi:10.1206/s15324834basp2601_1

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* 943. doi:10.1126/science.aac4716

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536. doi:10.1177/1745691612463401

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528–530. doi:10.1177/1745691612465253

Payne, B. K., Cheng, C. 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. doi:10.1037/0022-3514.89.3.277

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Snijders, T. A. B., & Bosker, R. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA: Sage.

Teachman, B. A., & Woody, S. R. (2003). Automatic processing in spider phobia: Implicit fear associations over the course of treatment. *Journal of Abnormal Psychology, 112,* 100–109. doi:10.1037/0021-843X.112.1.100

Westfall, J. (2015). *PANGEA: Power ANalysis for GEneral Anova designs*. Unpublished manuscript. Available at http://jakewestfall.org/publications/pangea.pdf

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal designs in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143,* 2020–2045. doi:10.1037/xge0000014

Winer, B. J. (1971). *Statistical principles in experimental design* (Vol. 2). New York, NY: McGraw-Hill.