Personalizing the Implicit Association Test increases explicit evaluation of target concepts

Brian A. Nosek
University of Virginia

Jeffrey J. Hansen
Project Implicit

Abstract

In an effort to remove a presumed confound of extrapersonal associations, Olson and Fazio (2004) introduced procedural modifications to attitude versions of the Implicit Association Test (IAT). We hypothesized that the procedural changes increased the likelihood that participants would explicitly evaluate the target concepts (e.g., rating Black and White faces as liked or disliked). Results of a mega-study covering 58 topics and six additional studies (Total N=15,667) suggest that: (a) after personalizing, participants are more likely to explicitly evaluate target concepts instead of categorizing them according to the performance rules, (b) this effect appears to account for the personalized IAT's enhanced correlations with self-report, (c) personalizing does not alter the relationship between the IAT and cultural knowledge, and (d) personalized and original procedures each capture unique attitude variation. These results provide an alternative interpretation of the impact of personalizing the IAT. Additional innovation may determine whether personalizing implicit cognition is viable.

Abstract = 150 words

Personalizing the Implicit Association Test increases explicit evaluation of target concepts

Modern conceptions of the mind make a distinction between deliberate, intentional, or explicit thoughts and feelings, and automatic, unintentional or implicit thoughts and feelings (Greenwald & Banaji, 1995). The theoretical distinction is advanced with a proliferation of measurement methods that assess social constructs – attitudes, stereotypes, and identity – without requiring an act of introspection or self-knowledge. The Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2007) is a popular method, in part, because it is adaptable for many research applications, relatively reliable as a measure of associative strength, elicits strong effects, reveals evaluations that are distinct, but related to self-report (Nosek & Smyth, 2007), and shows predictive validity of judgment and behavior across a variety of topics (Greenwald, Poehlman, Uhlmann, & Banaji, in press).

A topic of significant theoretical and practical interest is the extent to which the IAT and other implicit measures reflect something about the person or the culture they inhabit, or whether such a distinction is even meaningful. Some authors argue that the potential influence of cultural knowledge or, more generally, extrapersonal associations, contaminates the measurement of implicit attitudes (Arkes & Tetlock, 2004; Karpinski & Hilton, 2001; Olson & Fazio, 2004), whereas others suggest that such influence could be understood as a distinguishing feature of implicit and explicit attitudes (Banaji, Nosek, & Greenwald, 2004; Nosek & Hansen, 2008).[1]

Practical interest has led to procedural innovations meant to influence the extent to which personal or extrapersonal associations influence IAT performance. Olson and Fazio (2004) introduced a personalized IAT to reduce the presumed influence of extrapersonal associations – such as cultural knowledge about race on performance of a racial attitude IAT. They found, for example, that the personalized procedure elicited stronger correlations between the IAT and self-report for a couple of topics than did the original procedure.

In this article, we report an investigation of the impact of the personalizing procedural changes on the IAT's construct validity. We found that, compared to the original procedure, personalizing make participants less likely to categorize items into their superordinate categories as required by the task performance rules (e.g., identify a Black face as belonging to the category "Black"), and more likely to explicitly evaluate the target concepts (i.e., rate whether a Black face is liked or disliked). This effect, rather than reducing the IAT's relationship with cultural knowledge, appears to account for the enhanced relationship between the IAT and self-reported attitudes due to personalizing.

Following task instructions is a condition of measurement

A presumption for psychological measurement is that participants can and do follow the performance instructions. For example, in the Stroop task, proper task performance requires that participants attempt to name the ink color of words with either matching ("BLUE" printed in blue ink) or mismatching ("GREEN" printed in blue ink) meaning. If participants ignore the rule and instead just name the words, then measurement and interpretation are compromised. Experimenters might believe that participants are performing the task one way, but because of unclear instructions, expediency, or deliberate malfeasance, participants perform the task another way.

Researchers who have administered the IAT recognize that a common misunderstanding by participants is the rule for categorizing stimulus items. The IAT requires categorization of items representing four different categories (e.g., Black faces, White faces, good words, bad words) as members of those categories using two response keys. In one sorting condition, *Black faces* and *good words* are categorized with one response and *White faces* and *bad words* with the other response. In the other sorting condition *White faces* and *good words* are categorized with one response, and *Black faces* and *bad words* with the other.

Sometimes participants do not follow the task instructions for a few trials or more changing it from a categorization task with four categories (Black, White, good, bad), to an evaluation task with two categories (good, bad). Participants who make this error explicitly evaluate Black and White faces as good or bad, rather than categorizing them as Black or White.

Because this confusion is common, procedures have emerged to improve adherence to task instructions (Nosek et al., 2007). As a categorization task there is supposed to be a right answer for each item (i.e., it belongs to one and only one category). Providing error feedback and requiring a correct response before moving to the next trial quickly educates (or reminds) participants if they use the wrong rules. Other procedural standards also assist with communicating the performance rules, such as: [1] comprehensive instructions at the beginning and at each stage of the task, [2] presentation of the categories and stimuli in advance to clarify the correct categorization, [3] presentation of target labels (e.g., Black people/White people) and items in one color (white) and attribute labels (e.g., good/bad) and items in another color (green), [4] mixing the stimulus modality, such as attributes presented as words and target concepts as pictures, [5] fixed alternation of target and attribute trials so that the pattern of target or attribute categorization is predictable, and [6] avoidance of items that are easily coded as belonging to more than one of the categories. These procedures continue to evolve to maximize comprehension, accurate performance, and strength of inference (see the procedure: https://implicit.harvard.edu/).

Personalizing the IAT makes two procedural changes that may alter task performance

In an effort to remove the presumed influence of extrapersonal associations, Olson and Fazio (2004) introduced two procedural innovations to the original IAT design. First, for IATs that measure evaluative associations, the attribute labels were changed from “Pleasant” and “Unpleasant” to “I like” and “I don’t like.” Second, in the original IAT design, when a participant makes an error, they receive error feedback (a red “X”) and the trial does not end until the correct response is made. In the personalized design, error feedback is removed and no correction is required.[2]

We believe that these changes make it more likely that participants will misunderstand and not follow the IAT’s performance rules. Error feedback is one of the only ways to correct inaccurate task performance once the response block has begun. Without requiring correct responses, participants who *evaluate* rather than *categorize* will not know that they are performing the task incorrectly when those judgments require opposing behavioral responses. Further, by design, the labels “I like” and “I dislike” are intended to emphasize personal evaluations, as Olson and Fazio (2004) state: “changing the labels from ‘[Un]pleasant’ to ‘I [don’t] like’ may be enough to direct participants to construe the items presented in terms of their own attitudes and to reduce the influence of extrapersonal associations used to solve the IAT’s mapping problem” (p. 658). Our hypothesis is that the changes unintentionally increase the influence of *explicit evaluations* to solve the IAT’s mapping problem. In other words, these changes may subtly encourage participants to evaluate the target concepts too.

Overview

Together, the data summarized in this report support five claims: (1) the pattern of errant responses in the personalized IAT suggest an increased likelihood of explicitly evaluating target concepts rather than follow the task’s categorization instructions; (2) both label changes and error removal contribute to this effect; (3) the increased correlation between the IAT and explicit attitudes by personalizing appears to be due to this effect; (4) personalizing does not alter the relationship between the IAT and cultural knowledge, and (5) the personalized and original procedures each capture unique attitude variation – i.e., personalizing assesses a different component of attitudes rather than just removing a contaminating influence.

This article summarizes seven studies investigating the effects of personalizing the

IAT. Portions of the data from the six of the seven studies in this article were reported by Nosek and Hansen (2008). There we reported evidence that the original IAT procedure has reliable relations of varying magnitude with self-report, and little to no independent relation with cultural knowledge across 99 topics (158 samples, *N* = 107,709). In that report we did not examine the data collected with the personalized procedure, or compare the original and personalized procedures, the focus of the current article. With significant space constraints for this special issue, we present a full methods section for the main study – a mega-study of 58 topics and a sample of more than 12,000. The other six studies used very similar procedures and are summarized briefly in text. More detail is available in Nosek and Hansen (2008) and in supplementary materials available at http://briannosek.com/.

Method

Participants

A total of 12,972 tasks were completed by 7,401 volunteers at the Project Implicit research site (https://implicit.harvard.edu/; Nosek, 2005; Nosek & Hansen, 2008) between October 13, 2003 and September 17, 2004. Participants were randomly assigned to one of the 58 topics. Once assigned to a topic, that user was not assigned to the same topic again on future sessions. In effect, the study consists of 58 data collections on different topics with a common procedure. Of the 7,401 participants, 5,023 (68%) completed just one study session. Using only the first study completion for each participant for analysis does not alter the substantive interpretations reported here.

Of the 7,401 participants who reported demographic information (98% response rate) the following was observed: 68% female, 32% male; 1% American Indian, 5% Asian, 7% Black, 5% Hispanic, 73% White, 1% Biracial (Black-White), 4% Multiracial, and 4% Other; 20% Conservative, 31% Neutral or Moderate, and 49% Liberal; and, the average participant was born in 1974 (i.e., ~30 years old; *SD* = 11.6 years). Following data cleaning (dropping tasks with missing data or when >10% of response latencies were shorter than 300ms; see Greenwald, Nosek, & Banaji, 2003), a total of 12,152 usable completed study sessions across the 58 topics remained (*n*~210 per topic).

Materials

Implicit Association Test. The original and personalized IATs were the same except for two procedural differences (Olson & Fazio, 2004). One difference involves changing the evaluative category labels from ones thought to emphasize normative judgments in the original IAT (Good/Bad, Pleasant/Unpleasant) to ones that emphasize idiosyncratic judgments in the personalized IAT (I Like/I Dislike). The other change eliminated error feedback and requiring correction of the error before moving to the next response trial. All other procedural details were identical and are described below using attitudes toward George Bush relative to John Kerry as illustrative target concepts.

Participants completed seven blocks of response trials for either the original or personalized IAT. First, participants sorted evaluative words for 20 trials into categories (Pleasant/Unpleasant for the original IAT; I like/I dislike for the personalized IAT) using two response keys on a keyboard. Second, using the same response keys participants sorted faces and words associated with Bush and Kerry for 20 trials into categories (Bush/Kerry). Third, participants sorted items for all four categories (Bush, Kerry, Pleasant [I like], Unpleasant [I dislike]) for 20 trials using the two response keys. One key was used to categorize Kerry and Pleasant [I like] items; the other key was used to categorize Bush and Unpleasant [I dislike] items. Fourth, the same key mapping was repeated for 40 more trials. Fifth, like the 2nd block, participants sorted Bush and Kerry items again for 40 trials except that the response mapping was reversed (i.e., if Kerry items were categorized with the left key before, they were now categorized with the right key). Sixth, again participants sorted items from all four categories for 20 trials except that the response mappings for the category exemplars (Bush/Kerry) were opposite of the 3rd and 4th blocks. So, in this example, Kerry and Unpleasant [I dislike] were sorted with one key and Bush and Pleasant [I like] were sorted with the other. And, seventh, participants repeated the sorting conditions in the 6th block for 40 more trials.

In blocks with four categories, trials alternated between presenting target (Kerry, Bush) and attribute (Pleasant [I like], Unpleasant [I dislike]) items. Also, reminder labels appeared at the top of the screen for all blocks reminding participants of the current categorization rules. Further, to emphasize the distinction between the category and attribute dimensions, “Kerry/Bush” labels and items appeared in white, and “I like/I dislike” labels and items appeared in green on a black background. For the original IAT only, categorization errors were identified with a red ‘X’ below the stimulus item and participants had to correct the response before continuing to the next trial. An interstimulus delay of 150 milliseconds separated each trial. Finally, the order of the category mapping conditions (Kerry with Unpleasant [I dislike] before or after Kerry with Pleasant [I like]) was counterbalanced between-subjects.

IAT analysis followed recommendations of Greenwald et al. (2003). Two D algorithms are featured in this report: the standard *D* in which latencies for error trials are retained ‘as is’ and the *D600* in which latencies for error trials are replaced with the mean response latency for that block plus a 600 millisecond penalty.

Self-report measures. Explicit attitudes were indicated on 9-point thermometer scales, in response to “How warmly do you feel toward X?” and the difference between two ratings indicated a preference for one concept compared to the other (e.g., Kerry-Bush). Cultural knowledge measures paralleled the attitude measure in form, “How warmly does the average person feel toward X?”, and scale. Additional measures for the other studies described in the results assessed attitudes and cultural knowledge in a variety of ways including historical treatment, most people’s feelings, cultural warmth ratings, and societal portrayals. A summary report appears in Table 1 of Nosek and Hansen (2008).[3]

Design and Procedure

Once randomly assigned to a study, participants completed self-report measures and either the original or personalized IAT in a randomized order. Self-report measures were presented on a single webpage, in a randomized order for each participant.

Results and Discussion

Procedures that personalize the IAT inadvertently encourage participants to explicitly evaluate target concepts instead of categorizing them

Our hypothesis leads to a specific prediction about the pattern of errors that should emerge because of the personalizing procedural modifications. Explicitly evaluating all stimuli instead of categorizing might have little effect on error rates when the explicit evaluation of target categories matches the key assignment (e.g., for Republicans, “John Kerry” with “I dislike” on one key and “George Bush” with “I like” with the other) because it simplifies the task from a 4-category judgment task (John Kerry, George Bush, I like, I dislike) to a 2-category task (I like, I dislike). On the other hand, task recoding to explicit evaluation should increase error rates when the explicit evaluation of target categories mismatches the key assignment (e.g., for Republicans, “John Kerry” with “I like” with one key and “George Bush” with “I dislike” with the other). So, evidence for this task recoding will be observed if there is a magnitude increase in the absolute difference in error rates between the two critical conditions for the IAT for the target concepts, especially in line with explicit preferences.

Personalizing increases differential error rates. A study session included either a personalized or original format IAT for one of 58 topics. The data were analyzed with a multilevel model with the condition and other variables predicting absolute error differentials and attitude topic as the grouping variable (see also Nosek, 2005; Nosek & Hansen, 2008).

This approach separates variability across individuals from variability across attitude objects avoiding inferential problems due to the ecological fallacy. All inferential tests reported in text were conducted with multi-level models using attitude topic as the grouping variable. Aggregated means, percentages, or correlations are reported in text to facilitate comprehension of the results.

Across all topics, the absolute difference in error rates between the two IAT conditions was higher for the personalized than original IAT (t = 18.5, p<.0001, d=.34). A second model tested and found a significant random effect of topic (z = 2.96, p=.002) indicating that the impact of personalizing on differential error rates varied across topics. Figure 1 presents the estimates for the absolute differential error rates by topic in white for the original task and in black for the personalized task. The variation across topics suggests that there are unidentified factors that make recoding particularly likely. Identifying those factors is beyond the present scope, but an intriguing line for future investigation. The key for the present purposes is the main effect and that none of the 58 topics had estimated differential error rates higher for the original than the personalized task.

While recoding could occur with whichever stimulus items were presented, our hypothesis anticipated that the change in differential error rates between original and personalized procedures would be more evident in the target concept trials than in the evaluation trials. That is, participants who were evaluating rather than categorizing concepts would be more likely to miscategorize items representing John Kerry and George Bush, for example, than items representing “I like” and “I dislike.” As expected, the absolute differential error rate increased with personalizing for both target and evaluative stimuli, but the change was larger for target concepts with a 65% increase (original = 6.5%; personalized = 10.7%) than for evaluative items with a 27% increase (original = 7.9%; personalized = 10.0%; test of interaction of IAT version and target-evaluative items F[1, 12036] = 36.17, p<.0001, d=.11).[4]

In summary, as hypothesized, the personalizing modifications significantly increased the differential error rate between response blocks, especially for the target concepts. In the personalized task, participants appear more likely to explicitly evaluate target concepts instead of following the categorization performance instructions for the IAT. That increased likelihood might mean that explicit evaluation of target concepts occurs for just a few trials or more, and not necessarily for every participant. On its own, this result is consistent with our hypothesis, but does not unambiguously identify task recoding as the reason. The next sections provide additional evidence for our hypothesis.

Task recoding becomes more pronounced for people with strong preferences. The task recoding hypothesis suggests that people with explicit preferences for one concept over the other will be more likely to show increased error rates when the categorization rules oppose their likely evaluation. We calculated the absolute difference between explicit preference ratings such that the lowest value (0) indicated no explicit preference between the two target concepts, and the highest value (8) indicated the strongest reportable preference between the target concepts. The differential error rates between blocks were coded such that higher values indicated more errors in the response block that was incompatible with their explicit preference (e.g., more errors for Kerry supporters when Kerry was paired with negativity and Bush with positivity). This way, evidence for task recoding would be observed if there was a positive slope between explicit preferences and differential error rates.

Both original and personalized tasks show a significant positive slope such that participants with stronger explicit preferences show patterns of errors in the IAT consistent with those preferences. However, this effect was larger for the personalized (slope = .0152) compared to the original procedure (slope = .0058; slope comparison t=9.81, p<.0001, d=.19). Error rates in the personalized procedure are more sensitive to explicit preferences consistent with our task recoding hypothesis.

The increased differential error rate was not caused by increased response competition. An alternative explanation for these data could preserve the viability of the personalizing procedural changes, or even enhance it. The IAT effect relies on response competition – the conflicting tendency to respond with the wrong key press when the response pairings are incompatible with one's associations (Greenwald & Nosek, 2001). Indeed, the mental costs associated with overriding the tendency to evaluate the target concepts when evaluation and categorization require different responses is thought to be a central component to the IAT effect itself (Mierke & Klauer, 2003). It is possible that the greater error rate differential for the personalized task is a function of the increased response competition rather than evidence for task recoding. That is, an error in the IAT could reflect either a misunderstanding of the rules, or a manifestation of the phenomenon of interest – e.g., strong associations between *Bush* and *bad* for Democrats lead them to respond more slowly *and* make more errors when Bush and bad have opposing response assignments. Indeed, response latency and error rate differences are positively correlated in the original IAT showing that response competition both slows down responding and increases errors when the key assignments are "incompatible" (Greenwald et al., 1998; Greenwald & Nosek, 2001; Klauer & Mierke, 2005; Mierke & Klauer, 2003).

Our explanation, on the other hand, anticipates that the increase in differential error rates, if anything, will not affect or could even decrease response competition because some participants have an increased likelihood of explicitly evaluating the concepts instead of overriding the incompatible response tendency. If personalizing *increases* response competition, then when error differences between conditions are larger the response latency differences should be larger too (Greenwald & Nosek, 2001; Klauer & Mierke, 2005; Mierke & Klauer, 2003). Our alternative explanation predicts that RT differences may not change or even could decrease despite error differences being larger. And, more importantly, if our hypothesized secondary process - explicit evaluation - influences error differences, then the positive correlation between RT and error differences, a function of the common influence of response competition, would decline.

Variables were coded such that positive values indicated a preference for the concept implicitly preferred on average. Mean RT differences revealed that, across the 58 topics, effects were smaller for the personalized (M = 123, SD = 310) than the original (M = 154, SD = 314) procedure (t=3.96, p=.0002, d=.08).[5] Also as expected, in the original procedure, the correlation between the RT and error differences was r = .47 presumably reflecting the common influence of response competition (Greenwald & Nosek, 2001; Mierke & Klauer, 2003). However, with the personalizing changes, that correlation was weaker, r = .20 (difference: t=24.63, p<.0001, d=.45), suggesting that personalizing disrupted the common influence of response competition.[6] These data are not consistent with the alternative account that the enlarged error differentials in the personalized IAT are a function of *increased* response competition. Instead, the error rates appear to be affected by a secondary process – explicit evaluation.

Label changes and removing error feedback both increase recoding to an explicit evaluation task. We conducted three additional studies on different topics (Study 2: Kerry/Bush [n = 1124], Study 3: Black/White [n = 735], Study 4: Apple/Candy bar [n = 1197]) to test whether one or both personalizing procedural changes contributed to the task recoding effect. In addition to the original and personalized IATs, these studies added a 'hybrid' IAT (between subjects) that removed error feedback but did not change the category labels from *pleasant/unpleasant* to *I like/I dislike*. If both changes encourage task recoding, then the personalized task might show the most recoding, the original procedure the least, and the hybrid task somewhere in between.[7] Because the studies shared a common procedure, we analyzed them in aggregate for simplicity.

Replicating the main study, task recoding was especially prevalent in the target concept (original = 6.6%, hybrid = 10.6%, personalized = 13.2%) versus evaluation (original = 7.9%, hybrid = 8.1%, personalized = 7.3%) trials. In fact, the task difference was exclusive to concept trials. The interaction between IAT version (original, hybrid, personalized) and trial type (concept, evaluation) was significant (*F*[2,2876] = 32.13, *p*<.0001, *d*=.21), as was the primary main effect of IAT version (*F*[2,2876] = 22.67, *p*<.0001, *d*=.18).[8] Follow-up tests on just the target concept trials showed that the hybrid version elicited significantly more recoding than the original version (*F*[1,1914] = 38.49, *p*<.0001, *d*=.28), the personalized version elicited significantly more recoding than the hybrid version (*F*[1,1925] = 9.67, *p*=.002, *d*=.14), and neither increase was moderated by attitude topic (*d*'s=.05, *p*'s>.28). In summary, both removing error feedback and changing the evaluative labels from *pleasant/unpleasant* to *I like/I dislike* exacerbate task recoding.

Task recoding may, itself, account for changes in correlations between IAT scores and self-reported preferences.

The evidence in the previous section implies that the personalized IAT is more strongly correlated with self-report in some cases because more participants are explicitly evaluating the targets rather than categorizing them. In conditions that error patterns contribute directly to the IAT score, this should lead to an elevation of correlation between the IAT and self-report – but because of the changed performance behavior, not because the automatic responses are more in line with explicit evaluation as expected if the changes are removing a confounding influence.

This is evident in examination of the pattern of relationships with self-reported attitudes. Returning to the main study, across the 58 topics, the relationship between the average RT difference and self-report was *r* = .32 for the original procedure, and was not different for the personalized procedure, *r* = .34 (*t*=.92, *p*=.36, *d*=.02). Instead, the difference in IAT correlation with self-reports manifested in the pattern of errors reflecting the fact that the personalized task encouraged explicit evaluation. The average error difference correlated *r* = .20 with self-repot for the original task, and *r* = .28 for the personalized task (*t*=6.65, *p*<.0001, *d*=.12).

Two of the best performing D algorithms for calculating IAT effects (Greenwald et al., 2003) have different strategies for dealing with errors, leading to dissimilar expectations about how the personalizing changes would affect each. One *D* does not adjust the latency of errant responses – it just uses the latency from the beginning to the end of the trial. An alternative, *D600*, replaces each error with the mean of correct responses in that block plus a 600 millisecond penalty. As a consequence, if participants are explicitly evaluating the target concepts, then the greater number of errors in the "incompatible" versus "compatible" block will create an explicit influence via the penalty assessment. As such, the personalized procedures might increase the IAT's correlation with self-report using the D600 calculation, but perhaps not with the unadjusted *D*. As expected, across the 58 topics, the IAT correlation with self-report was uninfluenced by the personalizing changes with *D* (original, *r* = .38; personalized, *r* = .37; *t*=.39, *p*=.70, *d*=.007). However, the IAT correlation with self-report was reliably stronger with the personalizing changes with *D600* (original, *r* = .37; personalized, *r* = .41; *t*=.4.04, p<.0001, *d*=.07).

Four path models appearing in Figure 2 provide summary confirmation of the findings described here. For the original (left) and personalized (right) procedures, we tested the extent to which RT and error differences contributed to the *D* (top) and *D600* (bottom) effects and whether RT and error differences uniquely contributed to predicting explicit attitudes after adding the *D* calculation to the model. Considering the models in the top row (*D*) first, the RT and error difference effects were much more weakly related in the personalized procedure as discussed earlier. Further, for the personalized task, error differences did not uniquely contribute to the *D*. Most importantly, error differences had only a small effect predicting explicit attitudes for the original task (β=.17), but a large unique effect for the personalized task (β=.51). Further, when the *D* was included in the model, it partially accounted for the error differences unique prediction of explicit attitudes for the original IAT, but not at all for the personalized IAT.

There are two notable differences when *D* is replaced with *D600*. First, error differences predicted the *D600* effect better for the original procedure and went from zero to very strong prediction for the personalized procedure. This occurs by design as the *D600*, with the penalty assessment, formally integrates error differentials into the scoring. However, the size of the increase for the personalized IAT, especially compared to the original IAT, is surprising. Second, the *D600* now partially accounts for the large explicit-error difference relationship for the personalized IAT.

In summary, using the *D* calculation, for which the personalizing changes did not affect the correlation between IAT and self-report, the error differences did not contribute to *D*, but *were* substantially related to explicit attitudes. This occurred in sharp contrast to the original IAT that had a relatively weak relationship between error differentials and explicit attitudes. Using *D600*, on the other hand, revealed that the error differences had a stronger impact on the score and that partially accounted for the strong error-explicit attitude relationship.[9]

The basis of the stronger correlation between the IAT and self-report appears to be due to explicit evaluation influencing IAT performance to a greater degree after making the personalizing changes. This explanation is an alternative to Olson and Fazio’s (2004) proposal that the change in correlation magnitude was due to removal of the influence of extrapersonal associations, such as cultural knowledge. Next, we tested whether there were any changes in the relationship between the IAT and cultural knowledge as a function of the personalizing procedural changes.

Personalizing the IAT does not alter its (non) relationship with cultural knowledge

The key assumption driving the introduction of the personalizing procedural modifications was that they would reduce the influence of extrapersonal associations – such as cultural knowledge – that are distinct from one’s attitudes (Olson & Fazio, 2004). Putting aside the conceptual ambiguities of distinguishing person from culture and personal from extrapersonal (see Gawronski et al., in press; Nosek & Hansen, 2008), we tested whether the personalizing changes did influence the relationship between the IAT and cultural knowledge. Our previous report (Nosek & Hansen 2008) showed that the original IAT procedure had little to no relationship with cultural knowledge across 99 topics after accounting for common variation in explicit attitudes. This suggests that the personalizing changes would not alter the relationship between cultural knowledge and the IAT because there is no relationship to change in the first place.

We conducted a series of multilevel analyses predicting the IAT *D600* score with the attitude topic as a grouping variable. We used *D600* to give the personalized task as much benefit as possible, as this was the algorithm that increased the IAT-explicit correlation in the previous section. All effects were reliable (p<.05) unless noted otherwise. In the first model, consistent with Nosek and Hansen (2008; and Nosek, 2005), explicit attitudes predicted IAT scores (d=.32) and a significant random effects factor showed that the strength of IAT-explicit correspondence was stronger for some topics than others (z=3.93). Also, cultural knowledge was not uniquely predictive of IAT scores (d=.02) and that non-relationship did not vary across topics (z=.12). In a second model, a fixed effect of IAT version interacted with explicit attitudes (d=.08) confirming that the personalized task correlated more strongly than the original task with self-report (when using the *D600* algorithm). However, the non-relationship between the IAT and cultural knowledge did not vary significantly by version of the IAT (*d*=.04) suggesting that the personalizing changes had little or no impact on the relationship with cultural knowledge. Further, addition of a random effects factors of IAT version, cultural knowledge, and their interaction elicited no significant effects. Most important of those, the non-significant interaction (*z*=.03) indicated that the personalizing changes did not affect the relationship between the IAT and cultural knowledge for some topics and not for others. In short, cultural knowledge had no independent relationship with the IAT after accounting for explicit attitudes, and the personalizing changes had no effect on that relationship.[10]

Personalized and original IAT designs each contain unique attitude variance

The evidence summarized thus far supports our hypothesis that the personalizing procedures encourage task recoding that alters measurement, but not by removing the influence of extrapersonal associations, at least in the form of cultural knowledge. It is possible that the personalizing procedures do remove contaminating variance – just not what was assessed with our cultural knowledge measures. If the procedural modifications only remove contaminating variance from the IAT, then any attitudinal variation in the original IAT should be redundant with that measured by the personalized IAT. If, on the other hand, the procedural modifications alter attitude measurement, as suggested by the evidence that participants are more likely to explicitly evaluate target concepts, then the original IAT may retain unique attitudinal variation. Following Olson and Fazio's (2004) use of explicit attitudes as a criterion variable to index attitudinal variation in the IAT, we tested whether original and personalized IATs predicted explicit attitudes after removing the common variance between them.[11]

The studies described so far administered the original and personalized IATs between subjects so they are not useful for testing this hypothesis. We conducted three additional studies (Study 5: Bush/Kerry [*n*=82], Study 6: Black/White [*n*=142]; Study 7: Peanuts/Shellfish [*n*=235]) and administered both the original and personalized IATs to all participants to evaluate their unique contributions to predicting explicit attitudes.

To test whether the two IAT versions each contained unique attitudinal variation, in structural equation models using three parcels for each IAT as indicators and a latent factor of multiple explicit attitude measures, we regressed explicit attitudes on the original and personalized IATs simultaneously. That way, all common variance between the IATs was removed. If the personalized IAT is just less contaminated by non-attitudinal associations then the original IAT, then only the personalized version should be a significant predictor in the simultaneous models. As we predicted, in Study 5 measuring political attitudes both the personalized (β = .55, p < .001) and original (β = .31, p = .004) IATs had significant regression coefficients predicting explicit attitudes showing that each measure contained unique attitudinal variation. Study 6 measured racial attitudes, a topic that does not usually elicit strong correlations between implicit measures and self-report. In the structural model, the personalized IAT contained significant unique attitudinal variation (β = .27, p = .02), whereas the original IAT retained a non-significant amount of unique attitudinal variation in common with self-report (β = .20, p = .08). This test may have been underpowered considering that with racial attitudes there is less available shared variance for multiple measures to explain. Study 7 investigated attitudes toward foods. A structural model in which explicit attitude was regressed on original and personalized IATs showed that both the modified (β = .32, p < .001) and original (β = .36, p < .001) IATs contained significant unique attitudinal variation. In summary, these final three studies suggest that personalizing alters attitude assessment with the IAT. The previous analyses suggest that the alteration is a consequence of occasional intrusions of explicitly evaluating the target concepts. Those intrusions might occur for just a few trials or for an entire block, and for a few participants or many.

Conclusion

Summary of findings

Olson and Fazio (2004) observed stronger correlations between self-reported attitudes and the personalized IAT compared to the original IAT. They interpreted this as evidence that personalizing removed extra-personal contaminating variance in the original IAT, thus bolstering its relation with self-reported attitudes. We agree that such a difference in correlations is a necessary condition for showing the reduction of contaminating variance, but it is not sufficient to reveal the identity of the contaminating variance, nor does it require a conclusion that removal of contaminating variance is the operative cause. We hypothesized that the personalizing changes increase the likelihood that participants will explicitly evaluate all stimulus items instead of categorizing them.

In one mega-study with 58 topics and six additional studies we found that: (1) the pattern of errant responses in the personalized IAT suggest an increased likelihood of explicitly evaluating the target concepts instead of categorizing them; (2) both removing error feedback and changing category labels to *I like/I dislike* contributed to the effect; (3) the pattern of errors accounted for the changes in IAT-explicit correlations; (4) personalizing did not alter the relationship between the IAT and cultural knowledge, and (5) the personalized and original procedures each captured unique attitude variation rather than the personalizing task just removing a contaminating influence. Together, these data provide an alternative account of the implications of the personalizing procedural changes to the IAT.

What these data do not say

These data provide a substantial amount of evidence that the personalizing procedural changes have unintended performance effects, and may not remove cultural knowledge influences in IAT measurement. However, for the most part, the data do not have implications for the validity of the theoretical claims that spurred the procedural innovations. For one, while the personalizing procedural innovations may not 'personalize' the IAT as originally conceived, other changes might be effective. The notion that procedural changes could shift the type of associative information that influences IAT performance – personal or otherwise – is attractive and would provide measurement flexibility. Second, these data do not say that task recoding does not occur for the original IAT. Indeed, we know that it does with participant reports of misunderstanding the instructions. This is the origin of the numerous procedural enhancements mentioned in the introduction, and cause for continuing vigilance to improve instructions and maximize adherence to task instructions. Third, these data do not have implications for the conceptualization of personal-extrapersonal distinctions beyond that discussed in Nosek and Hansen (2008) and by Gawronski and colleagues (in press). Cultural knowledge, a presumed example of extrapersonal associations (Karpinski & Hilton, 2001; Olson & Fazio, 2004), does not relate to the IAT independently of explicit attitudes. This introduces conceptual and empirical challenges to (a) define extrapersonal associations (Gawronski et al., in press), (b) show that they have influence on the IAT, (c) explain why cultural knowledge does not (at least as measured by Nosek & Hansen, 2008), and (d) clarify whether it is a contaminating influence or a meaningful component of the attitude construct (Banaji, et al., 2004; Nosek & Hansen, 2008). In summary, these data provide a reinterpretation of the psychometric impact of the personalizing changes to the IAT, but they do not provide critique of the goal or viability of personalizing in the first place.

References

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.

Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2004). No place for nostalgia in science: A response to Arkes & Tetlock. *Psychological Inquiry, 15*, 279-310.

Gawronski, B., Peters, K. R., & LeBel, E. P. (in press). What makes mental associations personal or extra-personal? Conceptual issues in the methodological debate about implicit attitude measures. *Social and Personality Psychology Compass*.

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. (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. (2001). Health of the Implicit Association Test at Age 3. *Zeitschrift für Experimentelle Psychologie, 48*, 85-93.

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.

Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., & Banaji, M.R. (in press). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81*, 774 – 788.

Klauer, K.C., & Mierke, J. (2005). Task-Set Inertia, Attitude Accessibility, and Compatibility-Order Effects: New Evidence for a Task-Set Switching Account of the Implicit Association Test Effect. *Personality and Social Psychology Bulletin, 31*, 208-217.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology, 85*, 1180-1192.

Nosek, B. A. (2005). Moderators of the relationships between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134*, 565-584.

Nosek, B. A. (2007). Implicit-explicit relations. *Current Directions in Psychological Science, 16*, 65-69.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting intergroup implicit attitudes and beliefs from a demonstration website. *Group Dynamics, 6*, 101 – 115.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 365-292). Psychology Press.

Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion*.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test: Implicit and explicit attitudes are related but distinct constructs. *Experimental Psychology, 54*, 14-29.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & 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. (2004). Reducing the influence of extra-personal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86*, 653-667.

Author Note

This research was supported by a grant from the National Institute of Mental Health (R01 MH68447). Thanks to Yoav Bar-Anan and Kate Ranganath for helpful comments. Correspondence concerning this article should be addressed to Brian Nosek, Department of Psychology, University of Virginia, 102 Gilmer Hall, P.O. Box 400400, Charlottesville, VA 22904-4400. Electronic mail may be sent to nosek@virginia.edu. Supplements and related information can be found at http://briannosek.com.

Footnotes

[1] Gawronski, LeBel, and Peters (in press) point out that the construct definition of "extrapersonal associations" is ambiguous and many of the alternative interpretations have conceptual or explanatory limitations. This article focuses on clarifying the impact of personalizing procedural changes on the IAT's construct validity, regardless of the validity of extrapersonal associations. For the latter, see Gawronski et al, and Nosek and Hansen (2008).

[2] Olson and Fazio (2004) also reported a procedural change to the stimulus items representing the evaluative categories, but later found that it was non-essential. We used clearly pleasant and unpleasant items (e.g., horrible, wonderful; see Nosek & Hansen, 2008).

[3] Nosek and Hansen (2008) also found that the cultural knowledge measures had substantial interindividual variability and predictive validity of cultural variables – such as cultural attitude knowledge about race predicting perceived likelihood of job discrimination.

[4] While it could not explain the larger differential error rates for target than evaluative concepts, one might wonder if the procedural changes shifted participants' speed-accuracy tradeoffs. It appears not. Overall, participants were slightly *slower* with the personalized (M = 1015, SD = 304) than the original procedure (M = 987, SD = 282), while also having slightly *more* errors on average (personalized = 11.9%; original = 9.2%).

[5] Calculated as absolute values the RT differences were very similar (Original M=249, SD=245; Personalized M=245, SD=226). Also, after dropping the error trials completely, the RT differences were similar (Original M=122, SD=252; Personalized M=119, SD=290), and after calculating those as absolute values the original task effects were slightly weaker (Original M=200, SD=197; Personalized M=230, SD=213).

[6] This difference persisted even when only correct response trials were included in the RT calculation (original r=.36, personalized r=.20; t=13.77, p<.0001, d=.25).

[7] The reverse hybrid of changing the labels to "I like/I dislike" but retaining error feedback could be jarring to participants because the task would tell them that their personal likes and dislikes were wrong if they went against normative standards. Note, however, that this is also a feature of the original design and part of Olson and Fazio's key concern that normative information may be a confounding influence.

[8] There were some other significant effects in the model that are irrelevant to the present goals such as a main effect revealing a greater differential error rate on target concept than evaluation trials, and a significant interaction of trial type with attitude topic suggesting that some topics elicited a greater difference between targets and evaluation trials than others.

[9] A reviewer noted that, despite being the only algorithm that showed the stronger IAT-explicit correlation, *D600* is a less defensible analysis strategy because it introduces a penalty for incorrect responses while the participant gets no accuracy feedback. This is particularly true for the evaluative items if it is sensible to disagree with the normative response. Our use of relatively unambiguous evaluative items (e.g., wonderful, horrible) partially addresses this issue, but the reviewer's point emphasizes the personalized task's dependence on error patterns without procedurally emphasizing accurate responding.

[10] The effects were very similar with *D* except that personalizing did not increase the

relationship between the IAT and explicit attitudes. These analyses were replicated with Studies 2, 3, and 4 mentioned earlier using a wide variety of cultural knowledge measures using comparative model fitting to maximize opportunity to see an effect on cultural knowledge if it existed. We also replicated them in Studies 6 and 7 (see next section) in a within-subjects design. These analyses are available in the supplements.

[11] Explicit attitudes are an appropriate criterion because, by definition, explicit attitudes are personal. If they did contain extrapersonal associations, then Olson and Fazio’s (2004) observation that the personalized IAT relates more strongly with explicit attitudes could mean that it is *more* influenced by extrapersonal associations rather than less.

Figure 1. Estimated magnitudes of absolute differential error rates by topic.

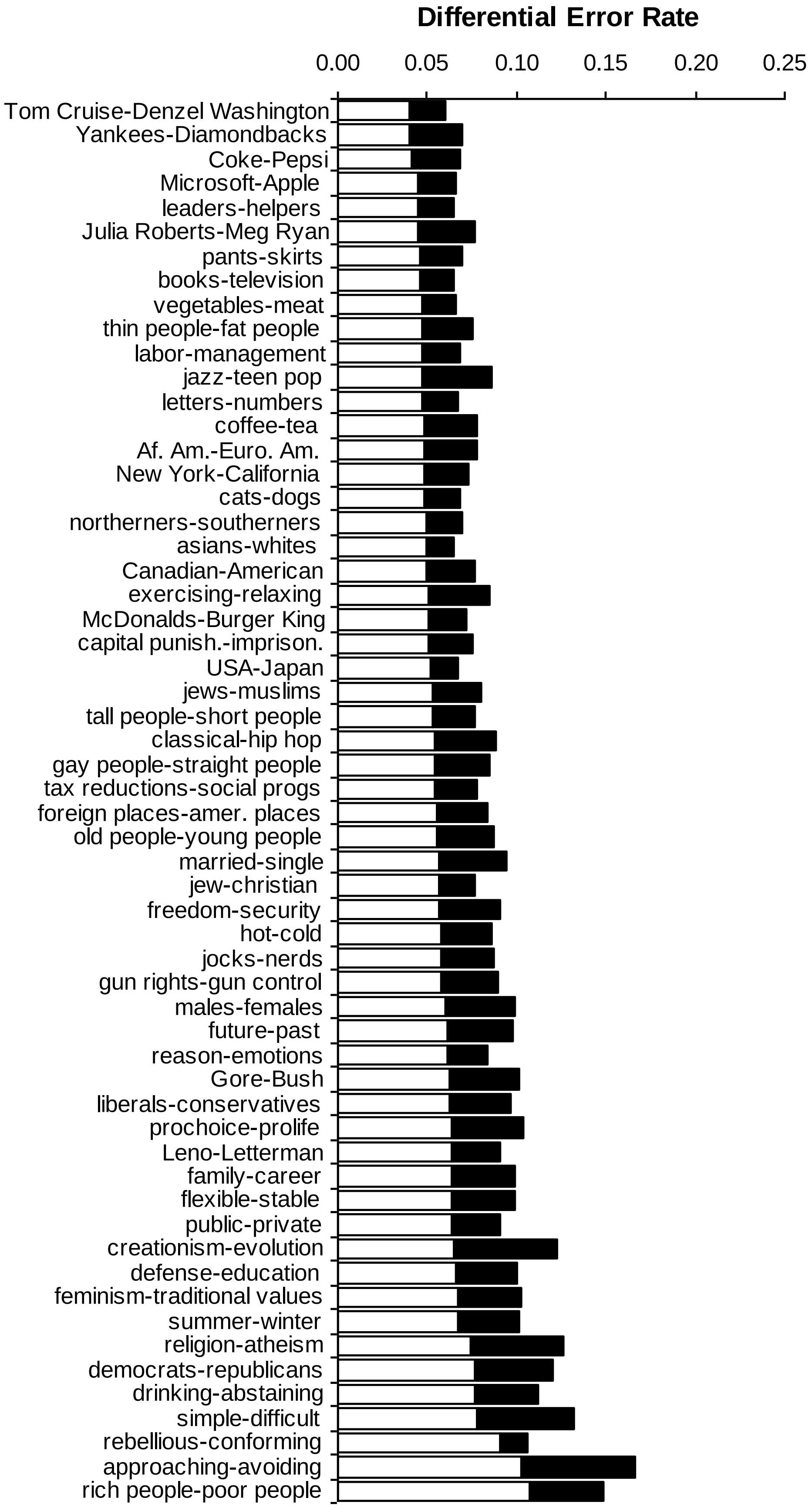
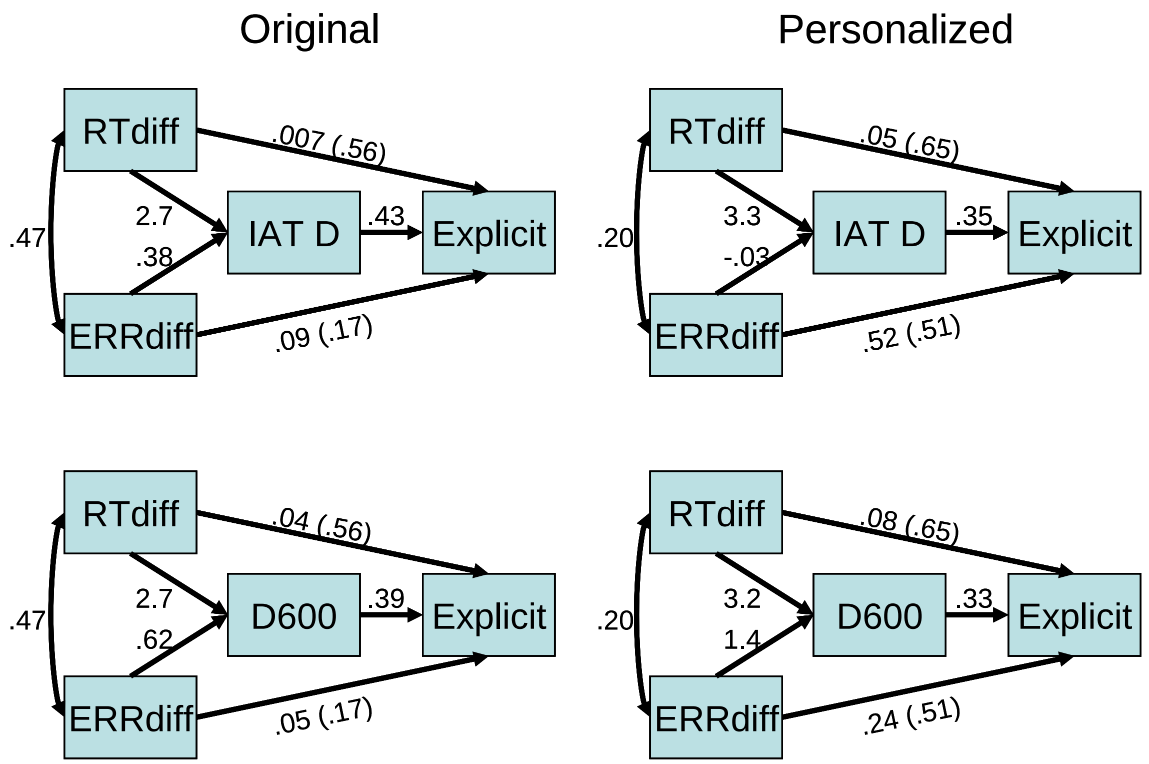


Figure 2. Four regression models testing unique effects of RT and error differences for the original and personalized IAT in predicting explicit preferences



Note: The top two models use the IAT D score, the bottom two models use the IAT D600 score. Values are standardized beta weights. Values in parentheses indicate relationship without the IAT score included in the model.