Faulty assumptions: A comment on Blanton, Jaccard, Gonzales, and Christie (2006)

Brian A. Nosek

University of Virginia

N. Sriram

University of Virginia

Contact information:

Brian Nosek

Box 400400; 102 Gilmer Hall

Department of Psychology

University of Virginia

Charlottesville, VA 22911

Email: nosek@virginia.edu

Word count = 3,300

Abstract

Blanton, Jaccard, Gonzales, and Christie (BJGC, 2006) assert that the Implicit Association Test (IAT) imposes a model that portrays relative preferences as the additive difference between single attitudes. This assertion is misplaced because relative preferences do not necessarily reduce to component attitudes. BJGC also assume that the IAT conditions represent two indicators of the same construct. This assumption is incorrect, and is the cause of their poor-fitting models. The IAT, like other experimental paradigms, contrasts performance between interdependent conditions, and cannot be reduced to component parts. This is true whether calculating a simple difference between conditions, or using the IAT *D* score. *D* – an individual effect size that is monotonically related to Cohen's *d* – codifies the interdependency between IAT conditions. When their unjustified psychometric assumptions are replaced with plausible assumptions, the models fit their data very well, and basis for their poor-fitting models becomes clear.

Word Count = 145

Blanton, Jaccard, Gonzales, and Christie (2006; hereafter "BJGC") criticize relative measurement and the use of difference scores for comparing experimental conditions. This article makes two points: (1) BJGC are too quick to dismiss the value of relative measurement and mistakenly suggest that relative attitude measures are necessarily reducible to component evaluations; and (2) BJGC misperceive the Implicit Association Test (IAT; Nosek, Greenwald, & Banaji, in press) conditions as analogous to two items of a scale and consequently impose invalid psychometric assumptions by decomposing the performance conditions as separate indicators of the construct. BJGC mistook their poor model fits as showing problems with the IAT rather than as indicating problems with their assumptions.

Relative preference ≠ additive difference

Relative measurement operationalizes key psychological concepts. For example, sensitivity (d') in signal detection is a relative comparison of the ability to separate signal from noise (Green & Swets, 1966; Greenwald, Nosek, & Sriram, 2006). Social judgments often rely on relative choices such as whether to hire a Black versus a White job candidate, or bet on the Yankees versus the Red Sox. These relative measures unambiguously represent a neutral or zero point of indifference such as when d'=0 and when the choice alternatives are equally attractive.

In the summary paragraph opening their General Discussion, BJGC (2006, p. 207) state:

> "Despite the enthusiasm for the [IAT] approach, a closer analysis reveals that the IAT framework requires restrictive causal assumptions and makes measurement assumptions that can be questionable. Because the IAT is focused on relative implicit evaluations at the conceptual level, theorists who use the IAT to predict psychological criteria must assume that the implicit evaluations of two distinct objects combine in an additive fashion to impact the criterion of interest."

Contradicting their statement that theorists *must assume* that two distinct objects combine additively in the IAT, BJGC acknowledge that it may not be true by noting that "One can use a number of different mathematical functions [of two distinct attitudes] to represent a construct that is 'relative'" (p. 208). Also,

both statements are overly restrictive in that they inappropriately imply that relative evaluations are necessarily describable as a function of two component evaluations.

Relative preferences often do not conform to the simple additive model assumed by BJGC. Whether such objects can be placed on a unidimensional, additive scale must be determined empirically. And, there are numerous cases in which relative preferences are not decomposable into component attitudes at all. For example, intransitive relative preferences, violating unidimensionality, have been demonstrated repeatedly (e.g., Bar-Hillel & Margalit, 1988; Birnbaum, 1992; Shafir, Osherson, & Smith, 1990; Tversky, 1969) as have sub- or super-additivity (Anderson, 1977). Also, Hsee, Loewenstein, Blount, and Bazerman (1999) identified circumstances in which most participants preferred object A to object B when considered separately, while B was preferred to A in a relative comparison. In short, relative preferences cannot be assumed to be an additive combination of two distinct evaluations, nor can they be assumed to be decomposable into distinct evaluations.

Therefore, BJGC's assumption that "implicit evaluations of two distinct objects combine in an additive fashion to impact the criterion of interest" (p. 207) is their own theoretical commitment; it is not imposed by the IAT procedure, nor is it a general characteristic of relative preferences.[1]

Similarly, BJGC characterization of the IAT as double-barreled is flawed (p. 204). Double-barreling concerns measurement situations in which agreement with one statement simultaneously requires agreement (or disagreement) with a second, independent statement. Consider, for example, the political preference question – Do you prefer George Bush to John Kerry? This is not double-barreled because the response indicates a relative preference, not independent assessments of liking for Bush and liking of Kerry. This is a proper procedure in the same sense that it is proper to ask relative preference questions in other well-established methods, such as Coombs's unfolding method of evaluative scaling (Coombs, 1950), or in studies of preferences among gambles or decision frames (Kahneman & Tversky, 1984; Tversky & Kahneman, 1981).

The value of relative preference measures

The IAT, in its general form, is unsuited for single association assessments (Nosek et al., 2005). Development of effective measures of single association strengths would be welcome contributions, and work to add measurement flexibility is on-going (e.g., DeHouwer, 2003; Nosek & Banaji, 2001). At the same time, the examples above illustrate the value of relative measurement.

BJGC state that "the relative construct of the IAT invokes a causal model that is restrictive in form and that will be inappropriate in many criterion-prediction contexts." (p. 193). To this statement, we add that interest in *single* evaluations at the conceptual level also imposes a causal model that is restrictive in form. Such models would be unable to detect relative preferences that are qualitatively different than those assessed as independent attitudes toward each component (e.g., Hsee et al., 1999). Selection of assessment – whether single or relative – should conform to the goals of the research question.

The IAT's combined task conditions are interdependent and not analogous to parallel items of a measurement scale

Distinct from the conceptual interpretation of the IAT's relation to relative versus single attitudes, BJGC also examined the IAT measurement model contrasting its two performance conditions. Development of a measurement model for the comparison of conditions is independent of whether the evaluation assessed is relative or not. For example, like the IAT, evaluative priming contrasts two performance conditions; and, unlike the IAT, evaluative priming can be used to measure single attitudes.

BJGC draw an analogy between the conditions of an IAT and items of a scale arguing that "… the compatible and incompatible task can be viewed as 'reverse coded' items that reflect opposing aspects of the *same* relative preference" (BJGC, p. 195, italics in original). BJGC's modeling exercise was based on this assumption after stating that "… the conceptualization of the two opposing IAT judgments as reverse scored indicators of the same construct leads to a straightforward psychometric model for the IAT" (p. 195).

While the foil for BJGC's criticisms was the IAT, these statements fail to recognize that the mean latency difference, with a history of over 100 years in experimental psychology, is taken to reflect the

additional processing demands in one condition, over and above the processes in a contrasted condition (Meyer, Osman, Irwin & Yantis, 1988). Since Donders' choice reaction time studies in the nineteenth century (see Donders, 1969), latency difference scores have been used to contrast performance conditions in many experimental paradigms including the Stroop task (MacLeod, 1991), semantic priming (Neely, 1991), task switching (Meiran, Chorev, & Sapir, 2000), evaluative priming (Fazio, Sanbonmatsu, Powell, & Kardes, 1986), lexical-decisions (Balota & Chumbley, 1984), attentional networks (Fan, McCandliss, Sommer, Raz, & Posner, 2002), and the IAT (Greenwald, McGhee, & Schwartz, 1998). Likewise, non-RT paradigms such as neuroimaging (Friston, Frith, Turner, & Frackowiak, 1995) and Evoked Response Potentials (Hillyard & Annllo-Vento, 1998) rely on similar ideas. In none of these cases is BJGC's analogy of the response conditions as a decomposable two-item scale useful for understanding the tasks, or generating a meaningful psychometric model.

The relevant content in these paradigms cannot be inferred from responses in one condition alone, they crucially depend on contrasting the interdependent conditions. For example, in the Stroop task, speed of color naming when the word meaning matches the ink color is not an indicator of the Stroop effect itself. The effect is revealed by contrasting that condition with one in which the words and ink color are mismatched. Also, Fan et al. (2002) measured three independent attentional networks by making pairwise contrasts of three distinct tasks. To assume that each task is simultaneously an indicator of all three networks is untenable. In other words, the conditions of these paradigms do not separately indicate the intended construct. As a consequence, the comparison cannot be decomposed meaningfully.

Condition contrasts in response latency paradigms are driven by construct definitions (e.g., Stroop interference, priming effect, IAT effect), not by a need to control for individual differences in general processing speed as BJGC suggest (p. 209). The content of interest in the IAT is the interference of the response pairings on the latency of stimulus categorization. For example, the IAT tests whether exemplars of the categories *Black*, *White*, *Good*, and *Bad* are more or less difficult to categorize when *Black* shares a response with *Good* (and *White* with *Bad*) compared to when *Black* shares a response with *Bad* (and *White* with *Good*). Like Stroop and priming effects, the IAT effect is inferred by comparing the

contrasted conditions. The response latencies within each condition are not a direct indicator of the content. This kind of interdependence between response conditions is well known in psychological science - it is the basis of the experimental paradigm.

The implication of this interdependence is that the two conditions are not separable indicators of the relevant content – the featured assumption of BJGC's models. For example, to assess the effectiveness of a math exercise on learning, a math test could be administered before and after the exercise. The conceptual interpretation sought in this study is *change in math performance*. The difference between the pre-test and the post-test would be one method of assessing the change in math performance. BJGC's models impose the interpretation that the pre-test and the post-test are "reverse scored indicators of the same construct." If this were true, then a researcher could assess the construct by measuring performance using just one condition, and the intervention's effectiveness could be revealed with pre-test data alone without conducting the intervention or post-test. Obviously, the relevant content – change in math performance – can only be inferred through comparison of the conditions.

The math tests and IAT blocks *might* have meaning in isolation, but not the meaning of interest. In isolation, the individual math tests might be indicators of math ability (higher performance = greater ability), but they do not independently assess change in math performance. Critically, in these cases, the metric coding is the same *before* calculating the difference. However, for BJGC's combining items, calculating the difference reverses one item to match the coding of the other, so the items match *after* calculating the difference (e.g., reversing negatively-coded Rosenberg self-esteem items so that higher values mean higher self-esteem for all items). For response latency measures, like the IAT, the relevant coding is the same before differencing – faster responses indicate stronger associations. The relevant IAT content is the comparative association strengths between conditions. BJGC's commit the fallacy of affirming the consequent by assuming that the converse of a true statement is also true: "reversed scale items are combined as difference scores" does not allow the conclusion that "difference scores are combinations of reversed scale items." Difference scores are not always decomposable as parallel items

that separately measure the intended construct. BJGC's insistence on interpreting the IAT conditions in isolation violates its interpretability as a measure of comparative performance.

BJGC were close to agreeing that the two IAT conditions are not parallel indicators of the IAT effect when they stated that "our [final] model tests suggested an alternative measurement model that treats the 'compatible' and 'incompatible' IAT judgments as distinct psychological constructs that can have different influences on psychological criteria" (p. 204), but they did not note that this invalidates their fundamental assumption treating the IAT's two combined tasks as analogs of two items of a measurement scale. BJGC thus mistakenly attributed the problem of their poor-fitting models to the IAT procedure, rather than to their assumptions.

Modeling combining items versus contrasting conditions

Figure 1 represents BJGC's model based on the analogy of interpreting the difference score of the IAT as combining items of a scale on the top, and our conception of the IAT effect being revealed exclusively by contrasted conditions on the bottom. Analytically, difference scores could be applied in either model. The appropriateness of doing so depends on the construct definition – whether the effect of interest revealed by each component independently (top), or by interdependent comparison of conditions (bottom). In the first case, the difference is a means of recoding and combining parallel items; in the second case, the difference itself is the meaningful unit of analysis. Multiple papers have recognized the latter as being the proper representation of the IAT difference (Cunningham, Preacher, & Banaji, 2001; Cunningham, Nezlek, & Banaji, 2004; Greenwald & Farnham, 2000; Greenwald, Nosek, & Banaji, 2003; Nosek & Smyth, in press). However, as revealed above, BJGC inappropriately applied the top model to the IAT asserting that it could be decomposed meaningfully.

D resolves the ambiguity of difference score models

D is an individual effect size assessment that reflects the comparative difficulty of performing the two response conditions of the IAT, with a value of 0 indicating that the tasks had equivalent performance (Greenwald et al., 2003). The procedure and logic of the IAT remain unchanged whether

scoring with a simple difference or *D*, but *D* is not decomposable meaning that it can only be applied to contrasted conditions.

Mean latency difference scores are biased by two factors: (i) individual differences in general processing speed and (ii) the positive association between mean latency and variance of latency and (Faust, Balota, & Spieler, 1999; Sriram, Greenwald, & Nosek, 2006). The first factor causes mean latencies to be positively correlated across subjects. The second factor causes mean latency differences to violate scale invariance across experimental conditions and causes interpretive difficulties, even in the analysis of single subject data in which individual differences in general processing speed are absent. The two biasing factors cause positive correlations with general processing speed.

Ironically, selection of the IAT as a foil for criticism of difference scores could have been an opportunity to point out that the individualized *D* effect size calculation (Greenwald et al., 2003), mitigates some of the known problems associated with simple differences by formally instantiating the condition interdependency, and this innovation may generalize to other contrasted performance paradigms (Sriram, et al., 2006). Instead, BJGC dismiss *D* saying "New variations of the IAT and new scoring methods recently have been introduced (Greenwald, Nosek, & Banaji, 2003; Nosek, Greenwald, & Banaji, 2005), but these revisions do not address our critique" (p. 194) and characterize *D* as "ad hoc," "dubious," and "empirically unjustified." Further, BJGC interpret the rationale for *D*'s use of dividing by the standard deviation as an "algebraic strategy … to control for general processing speed" (p. 210). In fact, *D* is an effect size calculation based on an individual's comparative performance in those conditions and is not reducible to the difference between response conditions. It can only be applied to the bottom model of Figure 1.

Greenwald, Nosek, and Banaji (2003) proposed standardization at the level of an individual's trial latencies to contrast the two conditions. For each participant, the latency difference $\bar{y} - \bar{x}$ is divided by σ_{xy}, an inclusive SD based on all of the trials in both of the IAT's combined tasks. The prototypical inference test for comparing conditions is the *t*-test that calculates the difference between the response

conditions and divides by the pooled standard deviation with an adjustment for sample size. Similarly, Cohen's *d*, an estimate of effect magnitude that is applicable to *t*-tests, is the difference between conditions divided by the pooled standard deviation (Cohen, 1988).[2]

To preserve the logic of parallel items that is essential to BJGC's critique, it may be tempting to translate *D* into the difference of two components, $\frac{\bar{y}}{\sigma_{xy}}$ and $\frac{\bar{x}}{\sigma_{xy}}$. However, for *t*, *d* and *D*, this translation is illusory. Because the standard deviation calculation depends on observations from both conditions, changing the values in just one of the performance blocks would affect the values of *both D* "components." In other words, the above "components" are blends of both response conditions. *t*, *d* and *D* can only be interpreted as single units. For example, in an experiment, there is no such thing as a *t*-value for the control condition and another *t*-value for the experimental condition. Representing *D* as separate components would violate assumptions of the independence of indicators.[3] In summary, whereas simple difference scores could be mistaken as combinations of parallel items, as BJGC did, *D* is only applicable to contrasted performance conditions because it instantiates the psychometric interdependency of the conditions, fitting our psychometric conceptualization.

Reanalysis of BJGC's Study 1

We reanalyzed data from BJGC's Study 1 to illustrate the value of our psychometric conceptualization in contrast to BJGC's models.[4] We removed BJGC's conception of the two conditions being parallel indicators, and introduced our conceptualization of the fundamental interdependence of conditions. We retained the other features of their models even though some are less than ideal – such as using a non-relative criterion (math identity) with a relative predictor, an attitude IAT (math compared to arts attitudes).

We first tested our psychometric alternative to model represented by BJGC's Figure 2 in which the IAT latent factor predicts a criterion variable. The critical difference is that, in our models, the response conditions are not parallel indicators of the IAT effect. Instead, the indicators include data parcels from both conditions reflecting their interdependence.

BJGC found a very poor fit with their psychometric assumptions that the IAT response conditions were indicators of the same construct (ε_a = .59 when the equal-but-opposite constraint was imposed and ε_a = .35 when that constraint was relaxed; BJGC msp. 14-15; ε_a = root mean square error of approximation; MacCallum, Browne, & Sugawara, 1996, suggested that $\varepsilon_a < .05$ be used as a benchmark for close fits). Our model, presented in Figure 1, respected the interdependency of conditions such that the 4 IAT indicators each reflected a *D* effect calculated on a parcel of the data. In contrast to BJGC's models, this approach provided an excellent fit to the data ($\chi^2(4)=6.3$, p=.18, ε_a=.00; 90% CI for ε_a=.00 to .04).

With their psychometric assumptions in place, BJGC found that their latent IAT factor was strongly related to general processing speed. Using the same general processing speed indicators as BJGC, we tested our psychometric alternative. To identify a baseline model in which general processing speed had no influence on IAT effects, we first constrained to zero the loadings of the four IAT indicators on the processing speed factor. This model resulted in a very good fit to the data ($\chi^2(75)=84.3$, p=.22, ε_a=.025; 90% CI for ε_a=.00 to .049; Figure 1). To test whether an important portion of the variance in IAT scores is explained by processing speed, we next allowed the loadings of the IAT indicators on the processing speed factor to be freely estimated (Figure 3). The fit of this model was also good ($\chi^2(71)=78.0$, p=.27, ε_a=.022; 90% CI for ε_a=.00 to .048), and was not significantly better than the fit of the more constrained model ($\Delta\chi^2(4)=6.2$, p=.18). That is, a model that respects the interdependence of IAT response conditions is not improved by accounting for general processing speed. Also, previous observations found that the *D* algorithm reduces the extraneous influence of general processing speed on IAT effects (Cai, Sriram, Greenwald, & McFarland, 2004; Greenwald, et al., 2003; Mierke & Klauer, 2003; Nosek & Smyth, in press). In conclusion, BJGC's models provided poor fits, not because of properties of the IAT, but because of their use of unjustified psychometric assumptions that derived from their misidentification of the IAT's task conditions as parallel items of a scale. Replacing those assumptions with ones that fit contrasted performance tasks substantially improved model fit.

Conclusion

In this article, we observed that the core assumptions motivating BJGC's critique are unjustified. Conceptually, relative attitude measures do not necessarily "reflect the difference between how a person implicitly evaluates two distinct attitude objects" (p. 193). Operationally, BJGC's faulty interpretation of the IAT blocks as being parallel, decomposable indicators of the same construct led them to impose invalid assumptions. Models imposing their assumptions fit poorly to real data, confirming the inappropriateness of the assumptions. Reanalysis of BJGC's data with models respecting the interdependence of the response conditions provided a very good fit to their data.

Footnotes

[1] BJGC never examined whether the IAT data conformed to their conceptual assumption. Their conceptual analysis (p. 199-200) focused on explicit measures and was irrelevant for the question of whether the relative assessment in their math-arts attitude IAT reflects the additive difference between the component evaluations. A test of this assumption with BJGC's data and additional information appears in a supplement available for download at http://briannosek.com/.

[2] The parallels between *t, d*, and *D* make obvious that one can calculate *t* and *d* scores on IAT data on an individual basis. Cohen's *d* (or its equivalent, *t*) was compared with *D* in analyses conducted by Greenwald et al. (2003, Footnote 6, p. 201). *D* possesses some valuable measurement properties that make it behave like a dominance measure (Handcock & Morris, 1999). Elaborating slightly, *D* is monotonically related to *d* and can assume values in the interval [-2, 2]. *D* is closely related to the probability with which latencies in one distribution exceeds those in a contrasting distribution (see also Sriram et al., 2006), and like *d*, has units that are independent of the units of measurement in the original data. Assuming equal number of trials in both conditions, the relationship between *D* and *d* can be expressed as:

$$D = \frac{2d}{\sqrt{4 + d^2}}$$

[3] Although the presumed components $\frac{\bar{y}}{\sigma_{xy}}$ and $\frac{\bar{x}}{\sigma_{xy}}$ are not interpretable in isolation, the *D* score can be represented as the difference between the mean of within-subject standardized latencies in the two conditions. A value of zero would correspond to the grand mean of the subject's latencies across conditions and positive and negative values are larger and smaller than the overall subject mean latency. In this componentization, the two mean standardized latencies correlate -1 across subjects. That is, a standardized mean latency contains the same information as the mean difference between standardized latencies. In effect, standardization recasts both mean latencies as relative measures and illustrates the absence of separable components.

[4] The authors thank Hart Blanton for providing the data for this reanalysis.

Authors' note

This research was supported by a grant from the National Institute of Mental Health (MH-R01 MH68447-01). Thanks to Mahzarin Banaji, Tony Greenwald, and Fred Smyth for helpful comments in the preparation of this article. Correspondence should be addressed to Brian Nosek, Department of Psychology, University of Virginia, 102 Gilmer Hall, P.O. Box 400400, Charlottesville, VA 22911. Electronic mail may be sent to nosek@virginia.edu. Related information can be found at http://briannosek.com.

References

Anderson, N.H. (1977). Failure of additivity in bisection of length**,** *Perception & Psychophysics*, 22, 213-222.

Balota, D.A., & Chumbley, J.I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage, *JEP: Human Perception and Performance, 10,* 340-357.

Bar-Hillel, M. & Margalit, A. (1988). How vicious are cycles of intransitive choice? *Theory and Decision*, 24, 119-145.

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

Birnbaum, M. (1992). Issues in Utility Measurement. *Organizational Behavior and Human Decision Processes*, 52, 319-330.

Cai, H., Sriram, N., Greenwald, A. G., & McFarland, S. G. (2004). The Implicit Association Test's D measure can minimize a cognitive skill confound: Comment on McFarland and Crouch (2002). *Social Cognition*, *22(6),* 673-684.

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

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review, 57*, 145-158.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12(2),* 163-170.

Cunningham, W.A., Nezlek, J.B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin, 30(10),* 1332-1346.

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology, 50(2),* 77-85.

Donders, F.C. (1969). On the speed of mental processes. *Acta Psychologica*, 30, 412-431. Translated from the original article written in 1868.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*, 340-347.

Faust, M. E., Balota, D. A., & Spieler, D. H. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125(6)*, 777-799.

Fazio, R.H., Sanbonmatsu, D.M., Powell, M.C., & Kardes, F.R.(1986). On the automatic activation of attitudes, *Journal of Personality and Social Psychology, 50*, 229-238.

Friston, K.J., Frith, C.D., Turner, R., & Frackowiak, R.S. (1995). Characterizing evoked hemodynamics

with fMRI, *Neuroimage, 2*, 157-165.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology, 79(6),* 1022-1038.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74(6)*, 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.

Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential validity of the Implicit Association Test: Comment on Blanton and Jaccard , *American Psychologist,* 61, 56-61.

Handcock, M. S., & Morris. M. (1999). *Relative Distribution Methods in the Social Sciences*. Springer.

Hillyard, S.A., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention, *Proceedings of the National Academy of Sciences***,** 95, 781-787.

Hsee, C. K., Loewenstein, G., Blount, S., & Bazerman, M. (1999). Preference reversals between joint and separate evaluations of options: a theoretical analysis. *Psychological Bulletin, 125*, 576-590.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39(4)*, 341-350.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130–149.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109(2),* 163-203.

Meiran, N., Chorev, Z., & Sapir, A. (2000). Component processes in task switching. *Cognitive Psychology, 41(3)*, 211-253.

Meyer, D. E., Osman, A. M., Irwin, D. E., & Yantis, S. (1988). Modern mental chronometry. *Biological Psychology, 26(1-3),* 3-67.

Mierke, J. & Klauer, K. C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology, 85(6)*, 1180-1192.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds), *Basic processes in reading: Visual word recognition (pp. 264-336)*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19(6)*, 625-666.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (in press). The implicit association test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic Processes in Social Thinking and Behavior*. Psychology Press.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31(2)*, 166-180.

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

Shafir, E. B., Osherson, D. N., & Smith, E. E. (1990). The advantage model: A comparative theory of evaluation and choice under risk. *Organizational Behavior and Human Decision Processes, 55(3)*, 325-378.

Sriram, N., Greenwald, A.G., & Nosek, B. A. (2006). Biases in Mean Response Latency Differences. Unpublished manuscript. University of Virginia: Charlottesville, Virginia.

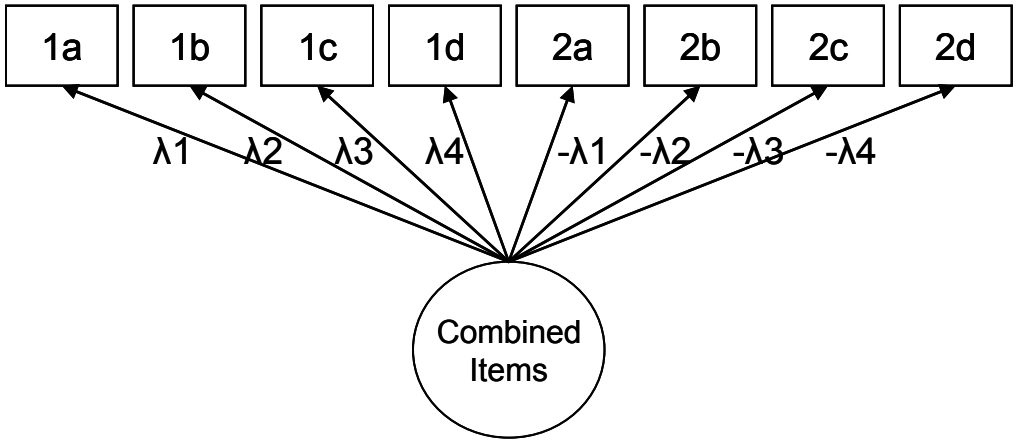
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 11(4481)*, 453-458.

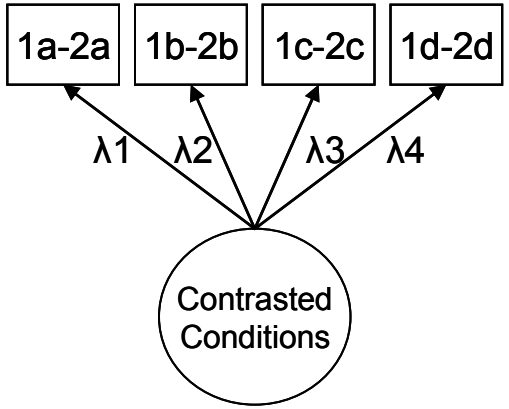
Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76(1),* 31-48.

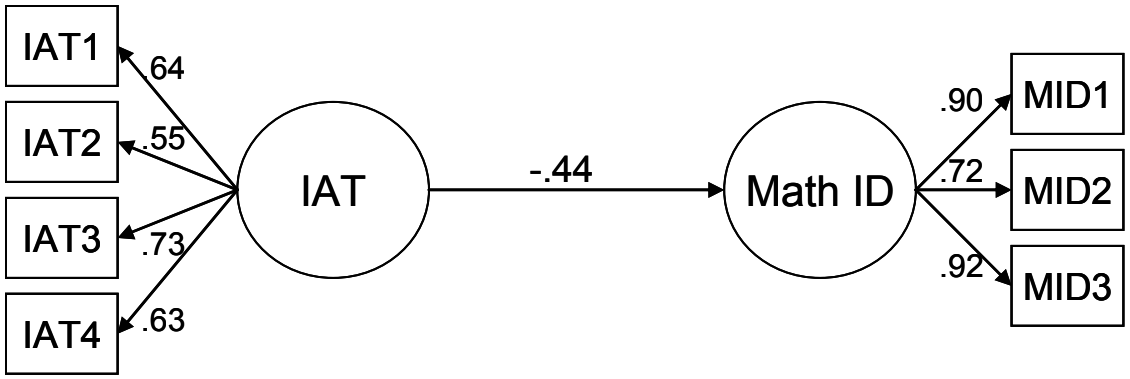
Figure 1. Structural models for combining items (top) and contrasting conditions (bottom) with four indicators (a,b,c,d) for each condition (1, 2), where the effect is represented as a difference score (Condition1 – Condition2). The combining items model assumes that (a) the subtraction reverse codes Condition2 scores to match scale interpretation with Condition1 (e.g., higher values=more of construct), (b) conditions can be decomposed, and (c) both conditions are indicators of a single construct. The contrasted conditions model assumes that (a) scale interpretation is matched before subtraction occurs, (b) the conditions cannot be decomposed meaningfully, and (c) the construct is revealed by the comparison of interdependent conditions.

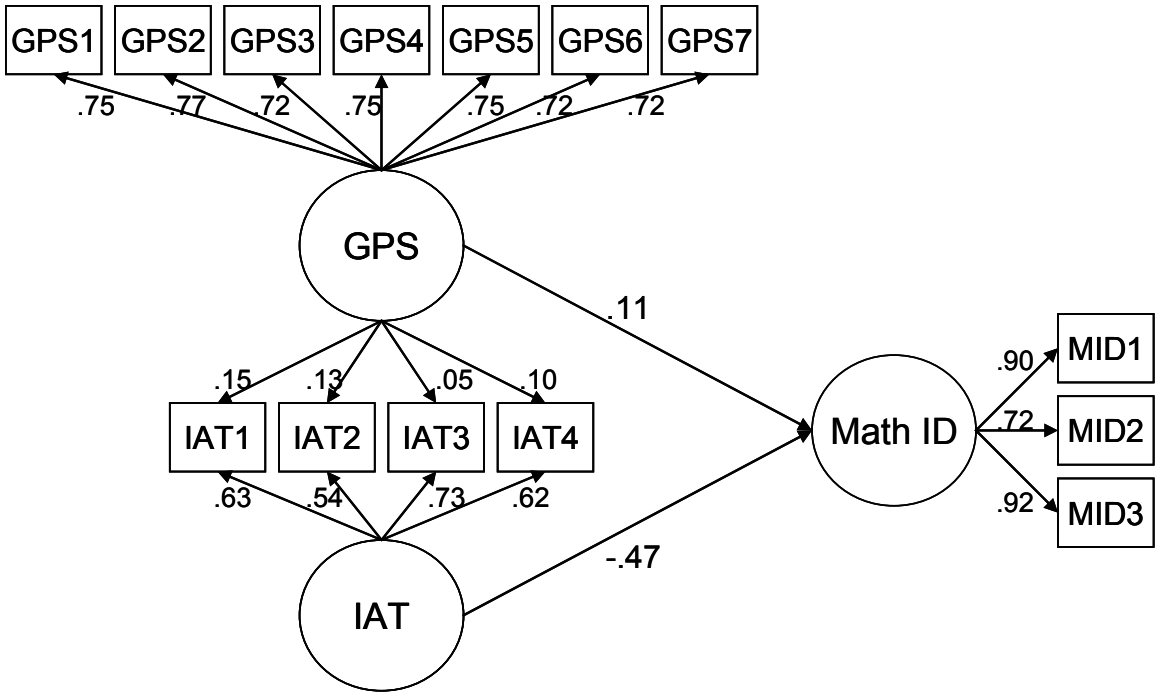
Figure 2. Structural equation model of a latent math-arts attitude IAT factor predicting math identity.

Figure 3. Structural equation model of a latent math-arts attitude IAT factor predicting math identity accounting for variance in IAT indicators shared with general processing speed.


1a
1b
1c
1d
2a
2b
2c
2d
λ1
λ2
λ3
λ4
-λ1
-λ2
-λ3
-λ4
Combined Items


1a-2a
1b-2b
1c-2c
1d-2d
λ1
λ2
λ3
λ4
Contrasted Conditions


IAT1
IAT2
IAT3
IAT4
.64
.55
.73
.63
IAT
-.44
Math ID
.90
.72
.92
MID1
MID2
MID3


GPS1
GPS2
GPS3
GPS4
GPS5
GPS6
GPS7
.75
.77
.72
.75
.75
.72
.72
GPS
.11
.15
.13
.05
.10
IAT1
IAT2
IAT3
IAT4
.63
.54
.73
.62
IAT
-.47
Math ID
.90
.72
.92
MID1
MID2
MID3