Supplement B to 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.

Supplementary Study B: Comparison of IAT effects calculated separately for attribute and category exemplars

Study 1 from Nosek, Greenwald and Banaji (2004) demonstrated that IAT effects calculated from response times for items from one category (e.g., Old), the other category (e.g., Young), or both categories (Old-Young) were essentially measuring a single evaluative attribute, that does not mean that responses to all stimuli within an IAT are measuring the same construct, to the same extent. Items that comprise an IAT usually represent two categories for each of two dimensions. For example, a measure of automatic preference for Old people relative to Young people would consist of exemplars from two concept categories (Old and Young) and exemplars from two evaluative categories (Good and Bad). Calculation of an IAT effect requires inclusion of response latencies for all four categories (Old, Young, Good, Bad) in two response-pairing conditions (Old with Good, Young with Bad; and Old with Bad, Young with Good). It is not known whether the category from which the response latencies are recorded moderates the magnitude of the IAT effect, the reliability of that effect, or the relationship between the effect and criterion variables. In other words, is the IAT effect expressed the same way in response latencies toward exemplars from all categories, or do response latencies toward some stimuli reflect the IAT effect more than responses toward other stimuli? No research to date has examined this question, and there is little basis for expecting that one dimension (category or attribute) or category (Old, Young, Good, Bad) would be more or less robust or have predictive utility than the others. This study was conducted in an exploratory fashion to observe differences, if any exist, in the magnitude, reliability, and predictive utility of IAT effects when using exemplars representing only the attribute (e.g., evaluation – good or bad) or the concepts (e.g., age – old or young).[1]

Method

Materials

The same data from the four tasks from Study 1 in Nosek et al. (2004; Bush-Gore attitude, Old-Young attitude, Black-White attitude, and Gender-Science stereotype from the Yale demonstration website) were used for this study.

Analysis Strategy

Two IAT effects were calculated for each of the four tasks following the identical procedure described in Nosek et al., 2004 (Study 1). The category IAT effects were calculated after first deleting the attribute data (i.e., good-bad responses for the attitude tasks; science-liberal arts responses for the stereotype task), and the attribute IAT effects were calculated after first deleting the category data (i.e., Bush-Gore, Old-Young, Black-

[1] Similar analyses were conducted examining the four IAT categories separately. Those analyses suggested that the only meaningful differences emerged between attribute and concept categories. So, individual category results are not reported here.

White and Male-Female responses).[2] As a result, the category and attribute IAT effects were calculated with about half of the data that typically comprises an IAT result. While overall reliability is likely decreased as a consequence, we restrict ourselves to comparisons between the category and attribute effects.

Results and Discussion

The Table presents data for each of the four IATs summarizing the average effect magnitude, relationship between the IAT effect and self-reported attitudes or stereotypes, split-half reliability of the IAT effect, zero-order correlations between the attribute and category IAT effects, and zero-order correlations with three extraneous influences for IAT effects calculated separately for attribute and concept stimuli. With the exception of the effect size of the Black-White IAT, attribute stimuli consistently elicited larger IAT effects, with the average difference between effect magnitudes for attribute versus concept categories being a moderate effect, $\underline{d}$ = .32 (Unweighted means of mean IAT effects: attribute $\underline{M}$ = .65, category $\underline{M}$ = .48). It is also notable that standard deviations for attribute effects were consistently larger than standard deviations for category effects (Unweighted means of standard deviations: attribute $\underline{M}$ = .63, category $\underline{M}$ = .51). Responses to attribute stimuli, representing categories such as 'Good' and 'Bad' appear to elicit stronger and more variable effects. Little theory exists to predict why this effect should emerge, and there is even less reason to predict why the Black-White task would deviate from this trend (at least in overall effect size).[3]

Relationships between IAT effects and self-report followed a similar trend as the effect magnitudes. With the exception of the Gender-Science stereotype, self-reported attitudes related more strongly to the attribute IAT effect (Unweighted average $\underline{r}$ = .35) than to the category IAT effect (Unweighted average $\underline{r}$ = .29), an average effect size difference of $\underline{q}$ = .08. The attribute and category IAT effects for the Gender-Science task showed equally strong relationships with self-reported attitudes.

Attribute IAT effects were more reliable than category IAT effects for two of the four tasks (Bush-Gore, Gender-Science), but were virtually identical for the other two tasks (Black-White, Old-Young). This lead to an average reliability for attribute IAT effects ($\underline{r}$ = .51) that exceeded the average reliability for category IAT effects ($\underline{r}$ = .45), an effect size difference of $\underline{q}$ = .09. However, because of the inconsistency of this effect across tasks, this result should be considered tentative. Also, the strength of the relationship between attribute and category IAT effects varied across the four tasks with the Bush-Gore task ($\underline{r}$ = .74) showing much greater consistency between attribute and category effects than the other three tasks ($\underline{r}$'s = .62, .60, .59). Finally, attribute and concept stimuli are not differentiated by resistance to extraneous influences with attribute stimuli being only slightly more resistant to overall speed (average $\underline{q}$ = -.035)

---

[2] For the Gender-Science stereotype task, it is not necessarily the case that Male-Female should be considered the categories and Science-Liberal Arts the attributes – i.e., genders have academic associates. One could argue instead that academic domains have gender.

[3] Follow-up analyses suggested that the difference between effect magnitudes for attribute versus concept stimuli appeared to vary as a function of self-reported attitudes. For example, respondents who reported equal liking for Black and White people showed no difference in effect magnitude for attribute and concept stimuli, but respondents who reported liking White people more than Black people showed larger effects with attribute stimuli than with concept stimuli. This observation does little by way of theoretical handles on which to interpret the effect.

and pairing order (average $\underline{q}$ = -.07) artifacts, and concept stimuli being slightly more resistant to experience with the IAT (average $\underline{q}$ = .03).

While little theory exists to predict differences in the magnitude, reliability, and relations to criterion (self-reported attitudes) of IAT effects calculated with just attribute or category exemplars, some differences emerged in these analyses. Across tasks, a tentative conclusion suggests that responses to attribute stimuli elicit larger and somewhat more reliable effects that are more strongly related to self-reported preferences than category stimuli do. The impact of this conclusion is tempered by the lack of theory available to explain it, and the occasional inconsistencies across tasks. Future research investigating the underlying process model that gives rise to IAT effects should consider the possible differences between category and attribute responses to provide a complete account of these effects. However, recent theorizing by Conrey et al. (2004) concerning the influence of automatic and controlled processes in IAT effects suggests that subjects may attempt to exert greater control over responses on concept trials because they more clearly implicate potential bias. As such, to the extent that subjects are successful at overcoming automatic biases, it may be particularly enhanced in responses to concept exemplars compared to attribute exemplars.

Reference

Conrey, F. R., Sherman, J. W., Gawronski, B. Hugenberg, K., & Groom, C. (2004). Beyond automaticity and control: The quad-model of behavioral response. Unpublished manuscript. Northwestern University.

Table. Calculation of two IAT effects for each of four tasks (Bush/Gore, Black/White, Gender-Science, Old/Young) using only response latencies for trials of exemplars from either the attribute (e.g., good/bad) or concept (e.g., Bush/Gore) dimension, and the relationship between implicit and explicit measures, reliability of the IAT effects, and relationships between attribute and concept IAT effects.

Task	Response categories	IAT D mean effect	IAT D SD	Effect size (d)	Implicit-Explicit Corr	Split-half reliability	Attribute-Concept Corr	overall speed	pairing order	experience with IAT
Bush-Gore attitude[1]	Good/Bad	.78	.63	1.23	.72	.70	.74	.05	-.09	.00
	Bush/Gore	.41	.43	.95	.62	.51		.05	-.03	.00
Black-White attitude	Good/Bad	.55	.65	.86	.33	.46	.62	-.04	.08	-.04
	Black/White	.54	.57	.95	.25	.47		.01	.13	-.10
Gender-Science stereotype	Science/Arts	.52	.65	.80	.20	.46	.60	.06	.06	-.09
	Male/Female	.37	.51	.73	.20	.38		.08	.20	-.10
Old-Young attitude	Good/Bad	.73	.57	1.29	.14	.41	.59	.07	.11	-.15
	Old/Young	.58	.51	1.13	.09	.43		.14	.14	-.20
Average effects	Attribute	.65	.63	1.05	.35	.51	.64	.04	.04	-.07
	Concept	.48	.51	.94	.29	.45		.07	.11	-.10

[1] IAT D mean effect, SD, and effect size calculated by combining strong Bush supporters (-2; N = 1747) and strong Gore supporters (+2; N = 2163) after reverse scoring Bush supporters' IAT effects.