Supplement A 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.

This study was mentioned in Footnote 8 of the original paper.

Supplementary Study A: Do the exemplars that represent a category influence IAT effects?

To date, our position regarding the selection of concept and attribute labels and stimuli necessary for the IAT has been that the labels should represent the attitude object of interest simply and clearly, and that the stimulus items should be selected to clearly belong to one, and only one, of the superordinate categories. When stimulus items are easily categorizable as more than one of the superordinate categories, the task demands of identification and categorization can interfere with the assessment of association between concepts and threaten the interpretability of the resulting effects.

However, more complex issues regarding these two dimensions, labels and stimuli, have been raised. The IAT requires the identification of individual exemplars of concept categories as belonging to one of two categories and variation is possible at both levels: the manner in which the concept/attribute category itself is specified (e. g., Black vs. White or African American vs. European American), and the types of exemplars used to represent the category (e.g., famous and liked individuals from the two groups, or unknown individuals). Do these design decisions have implications for the resulting task? For example, Mitchell et al. (2003) measured implicit attitudes toward Black athletes compared to White politicians with two separate IATs. One emphasized the race of the targets by using 'Black' and 'White' as category labels. The second used the same stimulus items, but emphasized the occupation of the targets by using 'Athlete' and 'Politician' as category labels. Despite the fact that the stimulus items were identical, the resulting implicit attitude varied as a function of the labeling of the categories. When race was emphasized, White politicians were preferred over the Black athletes, but when occupation was emphasized, Black athletes were preferred over the White politicians. These results suggest that category labels set the interpretive frame for stimulus items and that it's choice can determine the attitude effect that is observed. Category labels should represent the concepts of interest, be easily understood by the respondents, and provide clear distinctions between the four categories comprising an IAT.

While the specification of category labels in determining the IAT effects now seems obvious, the influence of the particular exemplars representing the concept categories is less understood and the goal of Study 3 is to address this issue more substantively. Mitchell et al. (2003) showed that the IAT race effect, when using the labels Black and White but with liked Black exemplars and disliked White exemplars produced a much weaker race bias (d = .23) than when the Black exemplars were disliked and the White exemplars were liked (d = 1.68). Yet, some researchers have argued that IAT effects are a function of the category labels, and not the individual exemplars (De Houwer, 2001; Fazio & Olson, 2003). Others have shown that exemplars do affect implicit responses when the meanings of stimulus items are confounded so that they represent more than one of the superordinate categories (Steffens & Plewe, 2001), and the question remains an open one with the data suggesting that both can play a role in the resulting IAT effect. The current study examines this question substantively with three types of stimulus variations to test whether selection of stimuli has important

consequences for the magnitude, reliability, predictive utility, and resistance to extraneous influences of IAT effects.

It is clear from previous research that dramatic changes in stimulus qualities will influence IAT scores (Mitchell et al., 2003; Steffens & Plewe, 2001), so this study examines some conditions in which more subtle variations could have an effect. For three tasks, using subsets of a homogeneous set of items produced a subtle manipulation of stimulus features (e.g., different old and young faces; Black-White, Old-Young, and Gender-Science tasks). This manipulation should not elicit different IAT effects unless the task is extremely sensitive to irrelevant stimulus variations. A more significant variation of stimuli involved generating two unique sets of stimuli that were conceptually similar but physically distinct (pictures versus words; Native-White stereotype). IAT effects might vary with this task if stimulus modality is an important factor. And finally, a third test involved switching of just two exemplars from the stimulus set with slightly different conceptual meanings (Gay-Straight attitude). This task might show variation in IAT effects if subtle shifts in semantic meaning of exemplars can influence IAT effects.

Method

Materials

Implicit measures. For three of the tasks (Black-White, Gender-Science, and Old-Young at the Demonstration Web site) multiple versions of the tasks were created with subsets of exemplars taken from a single set of items. The two versions of the Black-White task used the same Black and White faces, but each used four of the eight good and bad terms from the original race IAT for the categories 'Good' and 'Bad.' The eight text stimuli for each of the four categories in the Gender-Science task were split in half to create two versions of the task with no stimuli in common.[1] Finally, the six faces representing the categories 'Old' and 'Young' in the Old-Young task were randomly divided into three sets of two to create three versions of that task. The eight stimuli representing the categories 'Good' and 'Bad' in the Old-Young task were held constant across versions of the task.

For the Native-American task, two tasks were created that were designed to represent the same concepts using very different stimulus modalities (pictures versus words). Both versions used 'historical' White and Native American faces to represent the categories 'White American' and 'Native American', but used different sets of stimuli to represent the categories 'American' and 'Foreign'. In one case, those categories were represented by names of American or foreign places (e.g., Utah, Miami, Oslo, Italy), and in the other case the categories were represented by pictures of natural scenes of American or foreign landmarks (e.g., Grand Canyon, Sahara Desert).[2]

The two versions of the Gay-Straight task subtly varied the emphasis of the category 'Gay' to refer to lesbians or gay men. In both tasks, the same eight stimuli were

---

[1] Additional conditions described in the previous study with only two stimuli per category for the Black-White and Gender-Science tasks were also run. For simplicity, those data are not reported, but the effects observed with those versions are consistent with the effects reported here.

[2] Though it is not relevant for the current purposes, for the word task, American places were deliberately selected to be names of Native origins (e.g., Missouri). Also, for the picture task, the name of the landmark also appeared below the image in case respondents did not recognize the image on first viewing. So, the picture task is more appropriately termed a picture+word task. In any case, stimulus modality was very different than the word task.

used to represent the categories 'Good' and 'Bad', the same four stimuli represented the category 'Straight', and two of the four stimuli for the category 'Gay' were held constant across tasks (i.e., the words 'Gay' and 'Homosexual'). The only difference between the two versions of the Gay-Straight task involved two exemplars representing the category 'Gay'. In one case, the two images represented lesbians (i.e., two wedding cake bride figurines together, and two female 'stick figures' side-by-side), and in the other case, the two images represented gay men (i.e., two wedding cake groom figurines together, and two male 'stick figures' side-by-side). In sum, the tasks were nearly identical except for the slight shift in emphasis of gender for two items representing the category 'Gay'.

For all tasks, task-version conditions were manipulated between subjects, and respondents were randomly assigned to condition.

Results and Discussion

The table presents effect magnitudes, implicit-explicit correlations, internal consistencies, correlations between attributes and concepts, and correlations with known extraneous influences on five tasks each with 2 or 3 conditions in which different sets of exemplars represented the concepts of interest. For three tasks (Black-White, Gender-Science, Old-Young), the different versions of the tasks were created by using subsets of stimuli from the same stimulus set, which should maximize the task similarity except for the specific items used to represent the tasks. For the Native-American task, the two versions of the task were designed to capture the same concepts but varied in stimulus modality (picture or word). And, for the Gay-Straight task, the two versions of the task were nearly identical except that two stimuli in the category 'Gay' shifted the salience of lesbians or gay men. As can be observed in the Table, across tasks, variations in stimuli had very little impact on most of the evaluation criteria: implicit-explicit correlations (average $q = .02$), internal consistency (average $q = .05$), attribute-concept correlations (average $q = .07$), and resistance to extraneous influences (overall speed average $q = .07$, pairing order average $q = .06$, experience with the IAT average $q = .02$).

From the above effects, we might be tempted to conclude that exemplars play little to no role in the production of IAT effects. And it is clear that modality differences (pictures versus words) or different samples of exemplars selected from the same population produce similar IAT effects. This result is reassuring because it suggests that the IAT effect is not sensitive to the particular items that are chosen from a larger set that is homogeneous. It also suggests that IAT effects across different studies and from different laboratories using comparable stimuli from a superordinate set can be meaningfully compared.

However, when the exemplars vary in the content such that their meaning evokes a different evaluation, the stimuli can be said to be measuring different attitudes and the resulting IAT effect should differ. For three of the tasks, variations in stimuli produced a very small effect on the magnitude of IAT effects (Black-White $\underline{d} = .03$, Old-Young $\underline{d} = .09$, Native-American $\underline{d} = .09$). But, for the Gender-Science ($\underline{d} = .19$) and especially for the Gay-Straight task ($\underline{d} = .28$), changes in the exemplars representing the categories led to noticeable differences in the magnitudes of the resulting effects (see Table 3). This is perhaps most surprising for the Gay-Straight task because only two of the total of 24 stimuli were different between the two tasks. Even so, negativity toward 'Gay' relative to 'Straight' was much stronger when the category 'Gay' made salient gay men rather than lesbians, fitting with much existing theory that anti-gay attitudes are stronger for gay men than lesbians. While it is notable that the implicit-explicit attitude relationship did not also change as a function of the gender salience manipulation, that effect may be a function of the particular construct of interest (i.e., Gay-Straight attitudes) and not a

suggestion that similar effects will be observed for other attitude objects. While attitudes toward lesbians might be less negative than attitudes toward gay men, individual variation in liking might be very consistent. The effects of exemplars on correlations and prediction need to be explored further.

The results of this study clearly show two results of interest. First, trivial stimulus shifts from homogeneous sets of stimuli do not seem to influence IAT effects. Small variations in IAT scores were observed across subsets of good and bad exemplars (Black-White task), faces (Old-Young task), male and female pronouns, and science and humanities disciplines (Gender-Science task). The fact that stimulus modality (picture or word) did not have an impact on IAT effects is noteworthy. However, previous research has indicated an influence of stimulus modality suggesting that this effect should be examined further (Nosek et al., 2002). Second, when the exemplars tap a subcategory of the category label that is evaluatively distinct from other subcategories (Gay-Straight task), the IAT score may reflect that shift, as it well should. Together these results suggest that stimulus items may influence IAT scores when the variations in stimulus features are evaluatively relevant, but may have little effect when the features are evaluatively irrelevant.

Conclusion. Selection of exemplars for categories in the IAT should be made carefully with an emphasis on selecting stimuli that are (a) easily identified as members of their superordinate category, (b) not members of multiple superordinate categories, (c) reinforce the concept or attitude object of interest, and (d) as a set, fully represent all aspects of the concept or attitude object of interest, and not inadvertently emphasize only a part of the concept (e.g., only lesbians if the more general category 'Gay People' is of interest). These data also suggest that very minor stimulus variations (Black-White, Old-Young, and Gender-Science tasks) and use of pictures versus words (Native-American task) do not necessarily alter IAT effects.[3]

---

[3] Various pretests in our laboratories have been performed to examine the boundary conditions of stimulus selection for IAT effects. For example, we have conducted successful tests of IATs in which all stimuli were eliminated in favor of temporarily representing one category (Math) with X's as exemplars and the other category (Arts) with O's as exemplars to show that stimuli can, in principle, exist simply as placeholders for the superordinate category. Because the representativeness of stimuli of their superordinate categories can enhance the quality of IAT effects, we recommend against such applications in IAT use except for parameter testing.

References

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. <u>Journal of Experimental Social Psychology, 37</u>, 443-451.

Fazio, R. H. & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. Annual Review of Psychology, 54, 297-327.

Mitchell, J. A., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. <u>Journal of Experimental Psychology: General</u>, <u>132(3)</u>, 455-469.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting intergroup implicit attitudes and beliefs from a demonstration Web site. <u>Group Dynamics</u>, <u>6(1)</u>, 101-115.

Steffens, M. C., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. <u>Zeitschrift Fuer Experimentelle Psychologie, 48(2)</u>, 123-134.

Table. Effects of altering the exemplars representing a category for three tasks (Black/White, Gender-Science, Old/Young)

Task	Exemplar conditions	Effect Magnitude				Internal Consistency		Extraneous Influences		
		IAT D mean effect	IAT D SD	Effect size (d)	Implicit-Explicit Corr	Practice – Test Corr	Attribute-Concept Corr	overall speed (r)	pairing order (r)	IAT experience (r)
Black-White attitude	4a	.53	.51	1.04	.35	.54	.61	-.01	.10	-.10
	4b	.54	.51	1.06	.35	.57	.64	-.03	.15	-.08
Gender-Science stereotype	4a	.43	.48	.90	.24	.52	.63	-.01	.13	-.09
	4b	.53	.50	1.05	.24	.56	.61	.04	.17	-.10
Old-Young attitude	2a	.64	.47	1.36	.13	.54	.61	.14	.19	-.19
	2b	.60	.46	1.29	.07	.50	.57	.05	.22	-.19
	2c	.57	.47	1.21	.13	.52	.62	.15	.18	-.19
Native-American stereotype	Pictures	.18	.55	.32	.17	.64	.71	.04	.15	-.04
	Words	.23	.64	.36	.15	.68	.76	-.08	.28	.00
Gay-Straight attitude	Gay men	.52	.55	.95	.48	.64	.67	-.01	-.07	-.03
	Lesbians	.36	.58	.63	.51	.66	.71	.07	-.04	-.04

Note: Letter designations refer to multiple versions of the task using subsets of full stimulus set, but the same number of exemplars per set.