The Psychological Record, 2012, 62, 55–68

A BEHAVIOR ANALYTICALLY MODIFIED IMPLICIT ASSOCIATION TEST FOR MEASURING SEXUAL CATEGORIZATION OF CHILDREN

Amanda Gavin
Teesside University, UK

Bryan Roche
National University of Ireland, Maynooth

Maria R. Ruiz
Rollins College, Florida

Maria Hogan and Anthony O'Reilly
National University of Ireland, Maynooth

*The current study assessed the sexual categorization of children among a random sample of adults from the general population. Twenty-seven males and 27 females (*N *= 54) were exposed to a categorization task that assessed their ability to discriminate adult- from child-related words and sexual from nonsexual words. Then, in a modified Implicit Association Test they were required to respond with a particular key press to individual child- and adult-related stimuli paired with either sexual or nonsexual stimuli. In another block of testing the pairs of stimuli requiring a common key response were juxtaposed. There was more effective acquisition of common response functions for child/nonsexual than for child/sexual stimulus pairs for all participants combined. This effect was also observed for female participants separately but not for males. These findings support the utility of behavior-analytic variations of the Implicit Association Test but raise important considerations regarding their use as forensic and diagnostic tools.*

Key words: Implicit Association Test, implicit cognition, sexual categorization, pedophilia

The Implicit Association Test (IAT; Greenwald, McGhee, & Schwarz, 1998) is currently a popular psychological test format originally designed to reveal unconscious bias (or attitudes) in the context of race (Baron & Banaji, 2006; Greenwald et al., 1998; Greenwald, Oakes, & Hoffman, 2003; Nosek, 2005), gender (Aidman & Carroll, 2003; Greenwald & Farnham, 2000; Nosek, 2005; van Well, Kolk, & Oei, 2007), and other socially sensitive domains. When taking this computer-based test, a participant responds

Correspondence concerning this article should be addressed to Amanda Gavin, Psychology Section, School of Social Sciences and Law, Teesside University, Middlesbrough, TS1 3BA, UK. E-mail: A.Gavin@tees.ac.uk, or Bryan Roche, E-mail: Bryan.T.Roche@nuim.ie

to several items from each of four categories: usually two concepts (e.g., “African American” and “European”) and two attributes (e.g., “Good” and “Bad”). Participants are required to respond quickly with a right-hand key press to items representing one concept and one attribute (e.g., African American and Bad) and with a left-hand key press to items from the remaining two categories (e.g., European and Good). Importantly, participants are then required to perform the same task with the response rules juxtaposed in a counter-cultural fashion (e.g., respond with a right-hand key press to African American/Good items and with a left-hand key press to European/Bad items). The IAT reports bias where it finds shorter response latencies to tasks performed under the culturally consistent rules (referred to as *consistent tasks*) compared to those performed under the culturally inconsistent rules (referred to as *inconsistent tasks*).

The IAT has captured the attention of so many researchers because it is thought to be superior to many explicit attitude tests insofar as it may be capable of overcoming experimental demand characteristics and precluding fake responses (see Banse, Seise, & Zerbes, 2001; Kim, 2003). However, many serious issues remain regarding assumptions of core process and the opaque nature of the scoring process. These criticisms have been leveled by social/cognitive researchers (De Houwer, 2006; Fiedler, Messner, & Bluemke, 2006; Govan & Williams, 2004; Karpinski & Hilton, 2001; Olson & Fazio, 2003; Rothermund & Wentura, 2004; Steffens & Plewe, 2001) as well as behavior analysts (e.g., Gavin, Roche, & Ruiz, 2008; Roche, Ruiz, O’Riordan, & Hand, 2005).

Specifically, there is a great deal of uncertainty within the social-cognitive literature regarding what core processes are at work in the IAT (e.g., Rothermund & Wentura, 2004) and the accuracy of its measurement techniques (Blanton & Jaccard, 2006; Blanton, Jaccard, Gonzales, & Christie, 2006). While the creators of the IAT have claimed their test measure to be valid, others have called its validity into question (e.g., De Houwer, 2001, 2006). Indeed, Blanton and Jaccard (2006) raised concerns regarding the arbitrariness of the IAT metric system (i.e., the test-effect calculation method). Specifically, while arbitrariness in measurement systems is acceptable in the domain of theoretical research, IAT researchers may not be justified in drawing inferences about the traits or states of individuals on tacit psychological dimensions. Blanton and Jaccard called for a clearer operational definition of the IAT test effect. Despite a lack of clarity over what exactly the IAT measures and what researchers mean by the term “implicit” (De Houwer, 2006), psychologists continue to apply the IAT from within a social-cognitive paradigm (see Fiedler et al., 2006).

Most behavior analysts will be uncomfortable with the mentalistic terminology embedded in IAT narratives and will object to the use of statistical techniques (e.g., Greenwald, Nosek, & Banaji, 2003) that distort trial-by-trial behavior rates and actual reaction times to create an opaque, hybrid IAT score that accurately reflects neither response rate nor response time (see Gavin et al., 2008). However, a behavioral model of the IAT was recently developed by Roche et al. (2005) and was first empirically tested by Gavin et al. (2008). According to this model, the IAT can be conceived as a measure of participants’ fluencies with the relevant verbal categories employed in the test and their degree of experience at juxtaposing members of those verbal categories (i.e., the extent of contextual control over the categorization of the relevant social/verbal stimuli). Gavin et al. (2008) explained the model more specifically as follows: The verbal categories employed in an IAT are conceived as equivalence classes containing words. Higher order equivalence relations (see Wulfert, Greenway, & Dougher, 1994), or relations between equivalence relations or word categories (see Stewart, Barnes-Holmes, Roche, & Smeets, 2002), are often obtained in the natural environment. For instance, for an African-American racist the verbal categories White and Bad may participate in a higher order equivalence relation that we might refer to here as “Things I don’t like.” The IAT works by measuring the ease with which a common response function (e.g., press a left-hand key) can be established for two or more members of this higher order equivalence relation compared to members of different and unrelated equivalence relations for a given individual (e.g., White and Good). Thus, by measuring the facilitative or retarding effect of preexperimentally established verbal relations on the acquisition of laboratory-controlled functional response classes, the relative strengths of the relations between words in the various verbal categories can be ascertained (see Gavin et al., 2008, for empirical evidence). From a behavior-analytic perspective, therefore, the IAT measures verbal stimulus relations that are implicit (i.e., indirect) in the verbal repertoire, rather than the mental apparatus, of the individual.

Recently, researchers have begun to suggest that the IAT and other tests employing the IAT's basic process (e.g., the Implicit Relational Assessment Procedure; IRAP; Barnes-Holmes et al., 2006) might be useful in clinical or forensic contexts for identifying histories of behavior, such as sexual offending against children, which is rarely self-reported due to fear of sanctions. For example, Gray, Brown, MacCulloch, Smith, and Snowden (2005) explored the extent to which the IAT could distinguish a group of pedophilic offenders from offenders with no history of pedophilia. Using a Child-Sex Association IAT (CSA-IAT), the child sex offenders produced significantly faster response latencies on trials in which word exemplars of sex shared a response key with word exemplars of child than the control (or nonpedophilic offenders) group. Another study by Dawson, Barnes-Holmes, Gresswell, Hart, and Gore (2009) employed the IRAP to identify differences in the implicit beliefs of sexual offenders and nonoffenders. Their findings suggest that although both groups were able to discriminate between adults as sexual and children as nonsexual, this ability was significantly impaired in the offender group.

Interestingly, the behavioral model of the IAT mentioned above emerged directly from research on differentiating sex offenders. Specifically, Roche et al. (2005) first tested a behavior analytically modified IAT (i.e., employing traditional behavioral stimulus presentation formats and free of opaque scoring techniques) on a sample of incarcerated sex offenders against adults and children, as well as on a random sample of males and females from the general population. That preliminary study reported differences in test performances across groups that may allow researchers to differentiate sex offenders against children from sex offenders against adults. Perhaps more interestingly, however, although it was not discussed explicitly by the authors, the Roche et al. (2005) study also found a relatively high rate of "offender profile" response patterns among "normal" males and non-sex-offender prisoners compared to normal females (i.e., for normal males and sex offenders against children, functional response classes were often easily formed for "child" and "sexual" stimulus words. This was rarely observed for female control subjects, who typically showed the reverse effect). Indeed, Dawson et al. (2009) also reported a relatively high number of false positive implicit test results identifying control subjects as pedophiles using an implicit test. However, these researchers did not provide separate analyses of male and female control participants' response patterns on the implicit test.

We might interpret the observation of pedophile-profile response patterns (or false positive test results) in the general population first and foremost as failures of the relevant implicit test to detect a clearly delineated social group. Indeed, Dawson et al. (2009) accounted for the number of false positives uncovered in their study, saying that "although the IRAP may have some discriminative validity, the results presented here suggest that at present, similar to the IAT, it is an imprecise tool" (p. 71). However, the idea that false positive test results are based on test inadequacies belies two important and dangerous assumptions. The first of these assumptions is that sex offenders and normal males should not verbally categorize children in the same way. There are no grounds on which to base this assumption, because we do not yet understand the specific and relative role played by the sexual categorization of children in pedophilia. The two phenomena may be relatively unrelated, or causality between them may operate in an unexpected direction (i.e., sex offending may more often lead to inappropriate sexual categorization than vice versa). Secondly, this idea assumes that sex offenders against children comprise a distinct group with distinct syndromal characteristics, expressed via their implicit categorization of children. Both of these assumptions are problematic to the behavior analyst because in behavior analysis distributions of behavioral patterns across and within socially identified groups are empirical matters to be determined, not assumed. Moreover, behaviorists generally avoid syndromal classification, instead preferring a functional-analytic approach to behavior that emphasizes behavioral process over symptomology (see Sturmey, Ward-Horner, Marroquin, & Doran, 2007).

Given the foregoing, we should at least consider the possibility that many people selected at random from the general population may respond regularly to verbal relations more or less characteristic of child sex offenders. This may be because sexual categorization of children is not a reliable indicator of pedophilia or because, as some research indeed suggests, many more people than we may like to consider display pedophilic tendencies (for empirical evidence based on phallometric assessment see Freund & Costello, 1970; Freund & Watson, 1991; Hall, Hirschman, & Oliver, 1995; Quinsey, Steinman, Bergersen, & Holmes, 1975; see also Green, 2002).

Irrespective of which of these possibilities turns out to be the case in ongoing research, any program attempting to delineate two distinct social groups *must* start with the question of how frequently socially inappropriate categorizations are observed in normal populations. Indeed the extent to which this occurs will directly inform the validity of implicit tests designed to detect real or potential sex offenders as a clearly delineated social group.

Before we outline the current study it is important to review the modification we made to the traditional IAT format, consistent with our behavior-analytic orientation. The first modification was to remove feedback typically delivered by the IAT following incorrect responses only. The effect and, indeed, the purpose of the imbalanced feedback technique is to artificially lengthen response times recorded for the inconsistent (i.e., difficult) trials only (i.e., in line with hypotheses) and thereby exaggerate or even create *ab initio* an IAT effect. More specifically, in a traditional IAT, correct responses are in fact forced on a trial-by-trial basis by the requirement for participants to produce a correct response. This is achieved through the delivery of feedback, effectively requiring a timed observation response, following incorrect responses only. However, incorrect response times are not in fact measured to the point of the first incorrect response but from the point of stimulus presentation to the production of the forced second, *correct* response. In effect, response time measures include the time taken to produce the first response, *plus* the time taken to respond privately to the on-screen feedback that an incorrect response has been made, *plus* the time taken to produce the altered response. Because errors are more often made on inconsistent than on consistent trials, more artificially lengthened response times are recorded for the inconsistent trial block, thereby exaggerating or even creating a reaction-time-based IAT effect where one was weak or absent using a simple and transparent definition of response time (i.e., time taken to respond to the initial stimulus). As an interesting historical note for behavior analysts, early versions of the IAT that did not involve response feedback employed an arbitrary time penalty to be added to the reaction times recorded for incorrect responses. However, this arbitrary penalty has been replaced by the corrective feedback procedure, which leads to largely the same ranges of test results.

Given the foregoing, it should now be clear that it is response accuracy that actually underlies the widely reported reaction-time-based IAT effects. A similar criticism also applies to the recently developed IRAP (Barnes-Holmes et al., 2006) because it employs several IAT-style stimulus-presentation, feedback-presentation, and response-time-calculation techniques. In contrast, a more transparent and behavior analytically oriented implicit relations test eliminates the corrective feedback from any and all trials and emphasizes response accuracy over convoluted response time measures, as is tradition in the experimental analysis of complex human behavior.

In addition to imbalanced (i.e., poor) stimulus control across trials, the IAT and the IRAP both typically employ scoring algorithms to normalize data and remove outliers prior to inferential statistical analysis (i.e., data may sometimes be normalized twice if a *t* test or similar method is employed to compare difference scores across groups). Such algorithms typically involve the recoding of raw data, such as truncating response times above 3,000 ms to 3,000 ms and those below 300 ms to 300 ms. However, response windows are in fact usually infinite on each trial, and the effect of time taken beyond 3,000 ms to respond on individual trials (i.e., practice) has an unknown effect on the probability and speed of responses on subsequent trials (i.e., learning). Thus, rather than contrive a narrow range of reaction times statistically, or rely merely on experimental instructions to create rapid responding, a behavior-analytic implicit test employs a finite response window.

The current study employed a modified IAT to assess the sexual categorization of adults and children by a random sample of volunteers from the general population. Specifically, 27 male and 27 female participants (*N* = 54) were first exposed to a pretest categorization task that established their ability to discriminate adult- from child-related words and sexual from nonsexual words. They were then exposed to a test in which they responded to individual child, adult, sexual, and nonsexual stimuli in one of two ways. That is, for two sets of stimuli (e.g., child and sexual) participants were required to respond with a red key press, whereas for the other two sets of stimuli (e.g., adult and nonsexual) participants were required to respond with a blue key press (i.e., inconsistent block). In another block of testing the requirements were altered so that the stimulus pairs requiring a common key response were juxtaposed (i.e., “press red for child and nonsexual,” “press blue for adult and sexual”; consistent task block). This technique allowed the experimenters to assess the relational congruence of the child and adult verbal categories with each of the sexual and nonsexual verbal categories by comparing response accuracies across the two task blocks (i.e., consistent and inconsistent). It was expected that higher response accuracy would be observed on the consistent task block compared to the inconsistent task block for all participants (i.e., male and females combined). In addition, it was predicted that males and females, considered separately, would also show this pattern of responding.

Method

Participants

Fifty-four (27 male and 27 female) participants were recruited through personal contacts and by being approached on a university campus by the experimenter. Male participants had an age range of 18 to 62 years (*M* = 24.29 years, *SD* = 9.05); female participants had an age range of 19 to 56 years (*M* = 26.3 years, *SD* = 11.6). Of the 54 participants, four had not completed second-level education (i.e., high school), seven had completed second-level education but no higher, and 43 had completed (or were currently completing) third-level education (i.e., a university degree). All participants reported that they had no difficulties with reading and were capable of using a standard computer mouse and QWERTY keyboard.

Materials and Procedure

All three phases of the experiment were presented to participants on a laptop computer with a 15-in. display. Stimulus presentations were controlled using the software package Microsoft Visual Basic v.6.0, which also recorded all response accuracies. Sixteen stimuli in total were employed, all comprising words in the English language. These were assigned to one of four groups: adult, child, sexual, and nonsexual (see Table 1). It is important to appreciate that in an IAT, the categories employed need not be orthogonal to each other. Thus, individual stimuli need not be related along some continuum (e.g., good/bad, American/non-American, etc.). Stimuli were chosen here only on the basis that they should represent a recognizable instance of one of the four stimulus categories for most verbally able adults. This was established in Phases 1 and 2 (see the following section).

Table 1
Stimuli Employed as Exemplars of Each of the Four Verbal Categories in the Implicit Test

Adult	Child	Sexual	Nonsexual
Senior	Minor	Erection	Lamp
Grown-up	Infant	Horny	Stone
Mature	Kid	Foreplay	Tree
Old	Young	Aroused	Cloud

General Experimental Sequence

The current experiment consisted of three phases. Phases 1 and 2 were pretest categorization tasks and were presented in sequence. The aim of the categorization task was to ensure preexperimental familiarity with all stimuli. Participants sat at a standard computer desk and viewed the computer screen at eye level from a distance of 70 cm. Phase 3 was the main test phase. The three phases in total took approximately 10 min to complete.

Phase 1. For Phase 1 a set of instructions was presented on screen that read as follows:

> In a moment some words will appear on this screen. Your task is to choose which one of the words presented on the bottom of the screen goes with the word presented at the top of the screen. It is important that you try to make as many correct choices as possible. Please click “continue” when you are ready to proceed.

During this first categorization test, participants were presented with a word that verbally represented either the child or the adult category at the top of the screen. Participants categorized each stimulus by selecting an on-screen button labeled either “child” or “adult.” These child or adult category labels appeared as gray shaded rectangles in either the bottom left or bottom right side of the screen (these positions were counterbalanced across trials), and selection was operationalized by left-clicking the mouse on the chosen rectangle (see Figure 1).

All of the stimuli were presented in a quasirandom order, with participants exposed to each stimulus twice across 16 trials. There were no time constraints on these initial categorization tasks. The aim of this process was simply to establish whether or not participants were already familiar with the stimuli and could discriminate them from each other. A preset criterion of 14 out of 16 correct responses was set for progression to the next phase. No participant failed to satisfy this criterion.

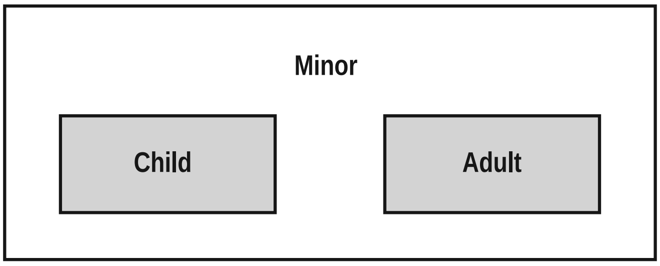


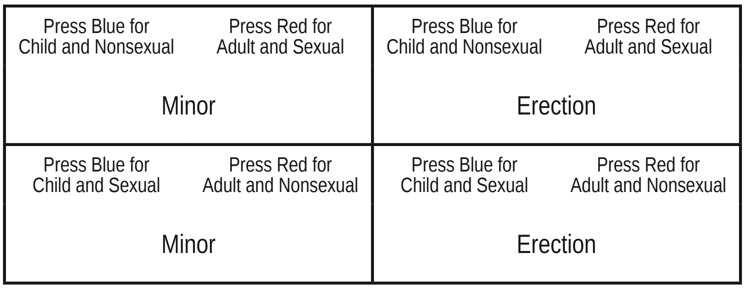
Figure 1. A sample task from Phases 1 and 2 (stimulus categorization).

Phase 2. This categorization task was identical to Phase 1 except that the stimuli were replaced with exemplars of the sexual and nonsexual categories. The instructions were once again presented at the beginning of this phase. Participants were required to discriminate between exemplars of the two stimuli by selecting the buttons labeled “sexual” or “nonsexual” displayed as gray boxes on the lower left and right of the screen using the left key press on the mouse. No participant failed to satisfy the response criterion.

Phase 3. Phase 3 consisted of 160 trials administered across two task blocks (80 trials in each block). One task block was expected to be consistent with the verbal history of the participants (i.e., press blue for child and nonsexual, press red for adult and sexual) and the second task block was predicted to be inconsistent with the verbal history of the participants (i.e., press blue for child and sexual, press red for adult and nonsexual; see Figure 2). Participants responded to the on-screen stimuli by pressing a key on the computer keyboard that was color coded. The specific keys that were colored blue and red were the “Z” and “M” keys, respectively.

The consistent tasks in the current experiment involved the child and nonsexual words sharing a response key (blue) and the adult and sexual words sharing a response key (red). For the inconsistent tasks, the child and sexual words shared a response key (blue) and the adult and nonsexual words shared a response key (red).

Each task block included four task types that involved the presentations of one of the following stimuli: child word, adult word, sexual word, or nonsexual word. These four tasks were presented once each in a random order in a block of four trials. There were 20 successive presentations of these four-trial blocks (i.e., 80 trials).



Press Blue for Child and Nonsexual Press Red for Adult and Sexual

Minor

Press Blue for Child and Nonsexual Press Red for Adult and Sexual

Erection

Press Blue for Child and Sexual Press Red for Adult and Nonsexual

Minor

Press Blue for Child and Sexual Press Red for Adult and Nonsexual

Erection

Figure 2. Four sample tasks presented to participants during Phase 3; the upper panels show consistent task types while the lower panels show inconsistent task types.

The consistent and inconsistent task blocks were presented in a randomized order, determined at the outset of the experiment by the computer software. Participants responded with either a blue or red key press within a 3,000-ms response window. If participants did not respond within the response window, the trial ended and the next trial began immediately. In this instance, the response was recorded as incorrect. Feedback was not given during the test trials. Participants received the following experimental instructions for both the consistent and inconsistent task blocks:

> In a moment some items will appear on this screen. Your task is to learn to press a blue or a red key on the keyboard when you see each of these items. Check the keyboard now to make sure you know where they are.
>
> You should use the instructions that will be presented at the top of this screen to help you decide which key to press.
>
> So, you should first look at the item in the centre of the screen and then use the rule at the top of the screen to help you make the correct response (i.e., press the blue or red key).
>
> Your object is to make as many correct responses as possible. You have only three seconds to respond to each item or your response will be recorded as incorrect, so you need to work fast!
>
> If you have any questions please ask the experimenter now.

Once the participants read and understood the instructions, they clicked on a gray rectangle labeled “Begin” to proceed with the task. There was no intertrial interval; tasks were presented immediately upon the production of a response or at the end of the 3,000-ms response window, whichever came first.

Results

All participants successfully reached criterion in Phase 1 and 2 categorization tasks (i.e., 14 or more correct responses out of 16 on each phase). These data were not analyzed in further detail. In Phase 3, all 54 participants completed the required 160-trial test (80 consistent tasks and 80 inconsistent tasks). Initially, the data for all participants (males and females combined) were examined to assess overall test effects.

The implicit test effect is best understood as a within-group difference in response accuracy across task blocks. Where such a difference is found, it points to a measured bias in the formation of one functional response class over another. Viewed this way, the direct cross-group comparisons of performances on individual blocks (e.g., using analysis of variance) are not meaningful insofar as performances on individual task blocks within the test have not been attributed with psychological status by either social cognitivists or behavior analysts. In the current study we conceive of implicit test outcomes in terms of planned within-group comparisons of response accuracies across task blocks (i.e., the size of the accuracy differential across task blocks). For all three *t* tests conducted, alpha was set at $p \leq .008$ in accordance with Holm’s (1979) sequential Bonferroni adjustment for conducting multiple *t* tests.

Forty-two of the 54 participants produced more correct responses on the consistent task block (press blue for child and nonsexual, press red for adult and sexual; *M* = 71.04, *SD* = 7.9) than on the inconsistent task block (press blue for child and sexual, press red for adult and nonsexual; *M* = 63.11, *SD* = 14.5). In effect, the implicit relations test results suggest that at a group level (i.e., males and females combined) functional response classes were more easily established for adult-related and sexual terms than for child-related and sexual terms. Similarly, functional response classes were more easily established for child-related and nonsexual terms than for adult-related and nonsexual terms. A paired-samples two-tailed *t* test revealed a significant difference between consistent and inconsistent response accuracies for all participants in the expected direction (i.e., greater accuracy on consistent than on inconsistent blocks), where $t(53) = 3.75$, $p < .001$, and $d = 0.21$, indicating a small effect size.

Interestingly, of the 12 participants who responded with greater accuracy on the inconsistent task block (i.e., a socially inappropriate performance), five were female (P1, P5, P10, P13, P25; see Figure 3), while seven were male (P7, P9, P13, P14, P17, P22, P24; see Figure 4). In addition, it should be pointed out that the magnitude of the accuracy differences across the two task blocks was notably smaller for these five females than for the seven males. Specifically, the mean female response accuracy was 2.4 greater on the inconsistent block compared to the consistent block. For males, however, this difference was 13.7, suggesting a stronger bias toward forming child–sexual and adult–nonsexual functional classes.

Data was separated according to sex in order to determine if response accuracy differences existed across the consistent and inconsistent task blocks for either or both of the gender groups considered individually. For females, a paired-samples two-tailed *t* test revealed a significant difference in response accuracy between the consistent (*M* = 72.52, *SD* = 5.4) and inconsistent (*M* = 61.48, *SD* = 15.4) task blocks, in the expected direction (i.e., culturally appropriate), where $t(26) = 4.21$, $p < .001$, and $d = 0.41$, indicating a small effect size.

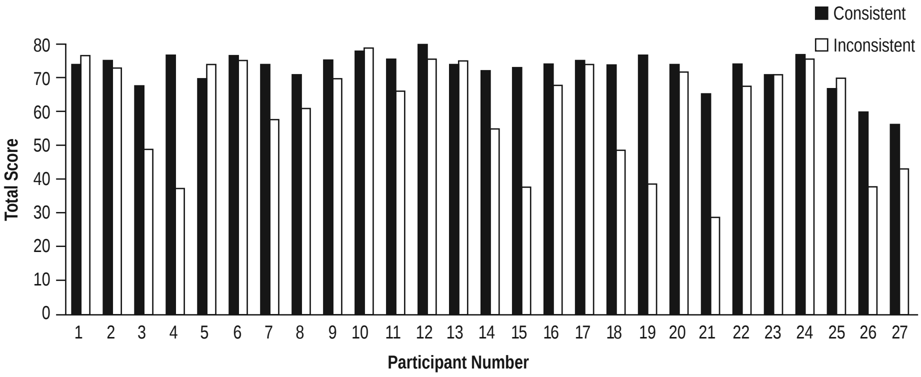


Figure 3. Response accuracies out of 80 on consistent and inconsistent task blocks for female participants.

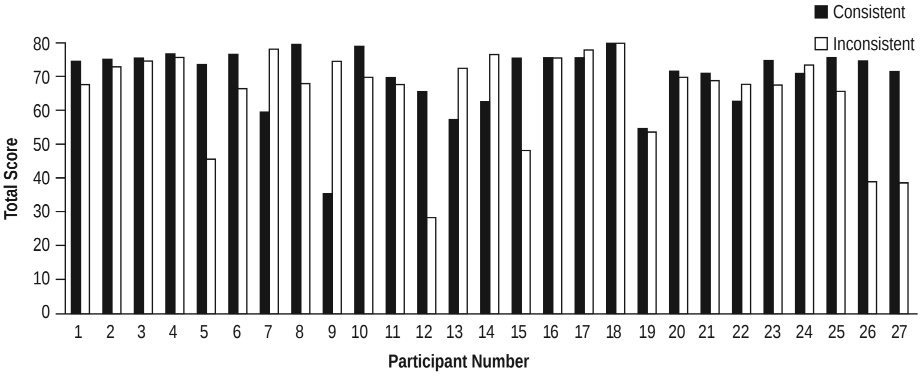


Figure 4. Response accuracies out of 80 on consistent and inconsistent task blocks for male participants.

Males showed no significant differences in response accuracy across consistent and inconsistent task blocks. In effect, female participants produced a statistically significant test effect in the culturally appropriate direction, whereas male participants did not.

Discussion

The current study used a behavior-analytic variation of the IAT to assess rates of acquisition of common response functions for words considered compatible for a normal population (child and nonsexual, adult and sexual) compared to words considered incompatible for a normal population (child and sexual, adult and nonsexual). Overall, there was more effective acquisition of common response functions on consistent task blocks than on inconsistent task blocks, as expected for a sample of participants from the general population. That is, participants responded with greater accuracy when child-related and nonsexual terms shared a common response key and when adult-related and sexual terms shared a common response key compared to tasks in which child-related and sexual terms shared a common response key and adult-related and nonsexual terms shared a common response key. Given that this outcome was expected (i.e., it is culturally appropriate) at a group level, these findings suggest that the current functionally transparent implicit relations test has some utility in identifying the verbal/social categorization practices of individuals. This supports the suggestions of Roche et al. (2005) and Gavin et al. (2008) that behavior analytically modified IATs may be of practical use to behavior analysts and psychologists generally as tests for implicit verbal relations.

When performances of male participants and female participants were analyzed separately, it emerged that only the female participants' test effect was statistically significant. That is, females responded with significantly greater accuracy when instructed to press blue for child/nonsexual and press red for adult/sexual than when instructed to press blue for child/sexual and press red for adult/nonsexual. Put simply, female participants displayed a culturally appropriate implicit test effect, whereas male participants did not. While most male participants showed individual test effects in the expected direction, as a group they did not do so to a significant degree. A visual analysis of the data suggests that there was a slightly higher number of male participants who produced a socially inappropriate (child sex offender profile) response pattern than there were females. Moreover, where this occurred, the differences in response accuracies across task blocks in the culturally inappropriate direction were considerably larger for these male participants than for their female counterparts. However, it is important to remember that the statistical analysis found any such trend to be nonsignificant for the male group as a whole. Thus, while male participants did not show a pedophilic response pattern, they did not show a "normal" one either. In other words, for male participants differences in response accuracies across test blocks were at chance levels.

Although it is difficult to ascertain at this point why an expected culturally appropriate performance was not observed for male participants, the current findings should serve as a cautionary note for those intending to identify sex offenders using implicit testing methodologies. While a small number of studies have reported differences in the sexual categorization of children by child sex offenders and nonoffenders (e.g., Dawson et al., 2009; Gray et al., 2005; Kamphuis, De Ruiter, Janssen, & Spiering, 2005; Mihailides, Devilly, & Ward, 2004; Nunes, Firestone, & Baldwin; 2007; Roche et al., 2005), none have done so with a particularly noteworthy level of discriminative validity. For instance, Dawson et al. (2009) reported that 43.7% of their control participants were incorrectly identified as sex offenders using an IRAP. Similarly, using an IAT, Gray et al. (2005) incorrectly identified 42% of control participants as child sex offenders. Interestingly, the current study found that only seven of 27 (25.9%) males and five of 27 (18.5%) females from the nonincarcerated population showed a child-offender profile response pattern (i.e., a reversed test effect). Of course, given that the sexual histories of the current participants are unknown, we have no way at present to determine if these figures represent false positives or not. However, it is precisely this type of research, replicated with a range of stimuli and participant populations, that will allow us to make such assessments over time.

It is worth noting that while it is prudent to treat the absence of a significant test effect for male participants in the current study as a chance result, not all researchers may view implicit test outcomes this way. In the Dawson et al. (2009) study, for instance, both child sex offenders and control participants correctly categorized children as nonsexual on an IRAP test. However, the offender group scores on the relevant trial type (child–sex) were found to be *not* significantly different from zero using a planned *t*-test analysis. The authors interpreted this absence of a difference from zero as a meaningful finding. The offenders in that study did not, therefore, show a prototypical offender profile response in categorizing children as sexual. Rather, "the offenders appeared to be unable to discriminate . . . between children as sexual versus nonsexual" (Dawson et al., 2009, p. 68). Given the foregoing rationale, the absence of a significant test effect for males in the current study may well be treated as psychologically significant by other researchers. Only continued research will clarify whether a significant test effect (i.e., in reverse direction to a normal test outcome) is indeed characteristic of child sex offenders, or whether the mere absence of a normal test effect affords in itself some predictive validity to implicit tests for pedophilia.

Given the current findings, it is worth considering the possibility that a larger difference in implicit test profiles between child sex offenders and control participants would be observed when the control group consists of males and females combined than when only males are included as controls. In effect, it is likely that while sex offenders may be marginally discriminable from a mixed-sex nonoffender population using an IAT-type test, they may not be so discriminable from any random sample of males. This possibility may in turn go some way toward providing an understanding of why previous studies using the IAT and IRAP to identify child sex offenders have yielded high false positive rates of offender identification.

One possible way in which the discriminative validity of implicit relations tests for child sex offenders might be enhanced was suggested by Roche et al. (2005). Specifically, Roche et al. suggested that the IAT and its variants reveal only histories of verbal categorization established by a verbal community. These tests cannot reveal unconscious desires or sexual predilections per se, unless these in turn covary reliably with identified verbal categorization patterns. Research on the reliability of any such covariations has yet to begin, although laboratory research into derived relational responding and the derived transformation of functions has suggested that sexual responses may indeed emerge directly from verbal contingencies (Roche & Barnes, 1997, 1998; Roche, Barnes-Holmes, Smeets, Barnes-Holmes, & McGeady, 2000; Roche & Dymond, 2008). While much research remains to be conducted to fully explore this issue, the understanding that implicit relations tests such as the IAT measure verbal category organization and not emotions, intentions, or beliefs per se points the pragmatic researcher toward improved forms of stimulus control over verbal behavior in order to improve the discriminative validity of these test formats. More specifically, what is required, according to Roche et al. (2005), is the use of stimuli whose discriminative functions are known and that differ across demographic groups of interest. In other words, it is not so much how participants categorize stimulus words compared to each other that is at issue, but the degree to which one group of participants is able to categorize the stimulus words compared to the other. For instance, pedophiles active on the Internet participate in a verbal culture with its own set of words specific to that subculture (see Roche et al., 2005). An implicit relations test employing these terms should show strong and clear categorization effects for those individuals whose verbal behavior is controlled by those words (i.e., a test effect is only possible because of known compatibilities and incompatibilities across word pairs and category pairs). In contrast, no such effect should be observed for those for whom the words have no special sexual or criminal meaning. In effect, by using an appropriate set of demographically distinguishing verbal stimuli, an implicit relations test should be better able to identify membership in a cultural group (e.g., users of online child pornography).

Interestingly, previous research has already suggested that individual stimuli play a key role in the overall IAT test effect (Dasgupta & Greenwald, 2001; De Houwer, 2001; Lane, Banaji, Nosek, & Greenwald, 2007). Specifically, differences in responses to a set of verbal stimuli in the IAT are not necessarily obtained across semantically related categories (Dasgupta & Greenwald, 2001) but depend on the choice of individual category members as stimuli appropriate to the task. Research indicates that the emotional valence of words may also be a crucial factor in the overall test effect (Govan & Williams, 2004). In this regard the current study could be fairly accused of having compromised stimulus control. The emotional valences of stimuli were unknown and the number of exemplars employed to represent the four categories were minimal (i.e., four each). It is reasonable to suggest that the use of different stimuli to represent these same verbal categories may have led to somewhat altered, if not broadly similar, results. Future research will require behavior analysts to functionally assess the stimuli employed in such tests in order to ensure stimulus control across the task blocks.

In summary, the current behavioral implicit relations test appears to function generally well as a functional-analytic alternative to the traditional IAT format for assessing patterns of verbal categorization. Despite modifications to the test format and the dependent measure employed, the test successfully identified a clear pattern of culturally appropriate sexual categorization of children and adults, which was observable at a group level and for most of the individual participants. However, the failure to find a significant test effect for male participants has also alerted us to important considerations regarding the reliability of implicit tests as clinical or forensic screening devices for identifying sexual offenders against children.

References

AIDMAN, E. V., & CARROLL, S. M. (2003). Implicit individual differences: Relationships between implicit self-esteem, gender identity, and gender attitudes. *European Journal of Personality, 17*, 19–36. doi:10.1002/per.465

BANSE, R., SEISE, J., & ZERBES, N. (2001). Implicit attitudes toward homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie, 48*, 145–160. doi:10.1026//0949-3946.48.2.145

BARNES-HOLMES, D., BARNES-HOLMES, Y., POWER, P., HAYDEN, E., MILNE, R., & STEWART, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit belief. *Irish Psychologist, 32*, 169–177.

BARON, A. S., & BANAJI, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 to 10 and adulthood. *Psychological Science, 17*, 53–58. doi:10.1111/j.1467-9280.2005.01664.x

BLANTON, H., & JACCARD, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*, 27–41. doi:10.1037/0003-066X.61.1.27

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. doi:10.1016/j.jesp.2005.07.003

DASGUPTA, N., & GREENWALD, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81*, 800–814. doi:10.1037/0022-3514.81.5.800

DAWSON, D. L., BARNES-HOLMES, D., GRESSWELL, D. M., HART, A. J. P., & GORE, N. J. (2009). Assessing the implicit beliefs of sexual offenders using the Implicit Relational Assessment Procedure: A first study. *Sexual Abuse: A Journal of Research and Treatment, 21*, 57–75. doi:10.1177/1079063208326928

DE HOUWER, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37*, 443–451. doi:10.1006/jesp.2000.1464

DE HOUWER, J. (2006). What are implicit measures and why are we using them? In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11–28). Thousand Oaks, CA: Sage.

FIEDLER, K., MESSNER, C., & BLUEMKE, M. (2006). Unresolved problems with the “I,” the “A” and the “T”: Logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology, 17*, 74–147.

FREUND, K., & COSTELLO, R. (1970). The structure of erotic preference in the nondeviant male. *Behavior Research and Therapy, 8*, 15–20.

FREUND, K., & WATSON, R. (1991). Assessment of the sensitivity and specificity of a phallometric test: An update of “Phallometric diagnosis of pedophilia.” *Psychological Assessment, 3*, 254–260. doi:10.1037/1040-3590.3.2.254

GAVIN, A., ROCHE, B., & RUIZ, M. R. (2008). Competing contingencies over derived relational responding: A behavioral model of the Implicit Association Test. *The Psychological Record, 58*, 427–441.

GOVAN, C., & WILLIAMS, K. (2004). Changing the affective valence of the stimulus items influences the I.A.T. by redefining the category labels. *Journal of Experimental Social Psychology, 40*, 357–365. doi:10.1016/j.jesp.2003.07.002

GRAY, N. S., BROWN, A. S., MACCULLOCH, M. J., SMITH, J., & SNOWDEN, R. J. (2005). An implicit test of the associations between children and sex in paedophiles. *Journal of Abnormal Psychology, 114*, 304–308. doi:10.1037/0021-843X.114.2.304

GREEN, R. (2002). Is paedophilia a mental disorder? *Archives of Sexual Behavior, 31,* 467–471. doi:10.1023/A:1020699013309

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,* 1022–1038. doi:10.1037/0022-3514.79.6.1022

GREENWALD, A. G., MCGHEE, D. E., & SCHWARZ, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. doi:10.1037/0022-3514.74.6.1464

GREENWALD, A. G., NOSEK, B. A., & BANAJI, M. R. (2003). Understanding and using the Implicit Association Test: An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216. doi:10.1037/0022-3514.85.2.197

GREENWALD, A. G., OAKES, M. A., & HOFFMAN, H. G. (2003). Targets of discrimination: Effects of race on responses to weapons holders. *Journal of Experimental Social Psychology, 39,* 399–405. doi:10.1016/S0022-1031(03)00020-9

HALL, G. C. N., HIRSCHMAN, R., & OLIVER, L. L. (1995). Sexual arousal and arousability to pedophilic stimuli in a community sample of "normal" men. *Behavior Therapy, 26,* 681–694. doi:10.1016/S0005-7894(05)80039-5

HOLM, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

KAMPHUIS, J. H., DE RUITER, C., JANSSEN, B., & SPIERING, M. (2005). Preliminary evidence for an automatic link between sex and power among men who molest children. *Journal of Interpersonal Violence, 20,* 1351–1365. doi:10.1177/0886260505278719

KARPINSKI, A., & HILTON, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 774–788. doi:10.1037/0022-3514.81.5.774

KIM, D. Y. (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly, 66,* 83–96. doi:10.2307/3090143

LANE, K. A., BANAJI, M. R., NOSEK, B. A., & GREENWALD, A.G. (2007). Understanding and using the Implicit Association Test: IV. Procedures and validity. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes: Procedures and controversies* (pp. 59–102). New York: Guilford Press.

MIHAILIDES, S., DEVILLY, G. J., & WARD, T. (2004). Implicit cognitive distortions and sexual offending. *Sexual Abuse: A Journal of Research and Treatment, 16,* 333–350. doi:10.1177/107906320401600406

NOSEK, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134(4),* 565–584. doi:10.1037/0096-3445.134.4.565

NUNES, K. L., FIRESTONE, P., & BALDWIN, M. W. (2007). Indirect assessment of cognitions of child sexual abusers with the Implicit Association Test. *Criminal Justice and Behavior, 34,* 454–474. doi:10.1177/0093854806291703

OLSON, M. A., & FAZIO, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14,* 36–39. doi:10.1046/j.0956-7976.2003.psci_1477.x

QUINSEY, V. L., STEINMAN, C. M., BERGERSEN, S. G., & HOLMES, T. F. (1975). Penile circumference, skin conductance, and ranking responses of child molesters and "normals" to sexual and nonsexual visual stimuli. *Behavior Therapy, 6,* 213–219. doi:10.1016/S0005-7894(75)80143-2

ROCHE, B., & BARNES, D. (1997). A transformation of respondently conditioned stimulus function in accordance with arbitrarily applicable relations. *Journal of the Experimental Analysis of Behavior, 67,* 275–301. doi:10.1901/jeab.1997.67-275

ROCHE, B., & BARNES, D. (1998). The experimental analysis of human sexual arousal: Some recent developments. *The Behavior Analyst, 21*, 37–52.

ROCHE, B., BARNES-HOLMES, D., SMEETS, P. M., BARNES-HOLMES, Y., & MCGEADY, S. (2000). Contextual control over the derived transformation of discriminative and sexual arousal functions. *The Psychological Record, 50,* 267–291.

ROCHE, B., & DYMOND, S. (2008). A transformation of functions in accordance with the non-arbitrary relational properties of sexual stimuli. *The Psychological Record, 58,* 71–90.

ROCHE, B., RUIZ, M., O'RIORDAN, M., & HAND, K. (2005). A relational frame approach to the psychological assessment of sex offenders. In M. Taylor & E. Quayle (Eds.), *Viewing child pornography on the Internet: Understanding the offence, managing the offender, and helping the victims* (pp. 109–125). Dorset: Russell House.

ROTHERMUND, K., & WENTURA, D. (2004). Underlying processes in the Implicit Association Test (I.A.T.): Dissociating salience from associations. *Journal of Experimental Psychology: General, 133*, 139–165. doi:10.1037/0096-3445.133.2.139

STEFFENS, M. C., & PLEWE, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. *Zeitschrift fur Experimentelle Psychologie, 48*, 123–134. doi:10.1026//0949-3946.48.2.123

STEWART, I., BARNES-HOLMES, D., ROCHE, B., & SMEETS, P. M. (2002). Stimulus equivalence and non-arbitrary relations. *The Psychological Record, 52*, 77–88.

STURMEY, P., WARD-HORNER, J., MARROQUIN, M., & DORAN, E. (2007). *Structural and functional approaches to psychopathology and case formulation*. San Diego, CA: Elsevier Academic Press. doi:10.1016/B978-012372544-8/50002-1

VAN WELL, S., KOLK, A. M. M., & OEI, N. (2007). Direct and indirect assessment of gender role identification. *Sex Roles: A Journal of Research*, 56, 617–628. doi:10.1007/s11199-007-9203-7

WULFERT, E., GREENWAY, D. E., & DOUGHER, M. J. (1994). Third-order equivalence classes. *The Psychological Record, 44*, 411–439.