Test-Retest Reliability and Predictive Validity of the Implicit Association Test in Children

James R. Rae and Kristina R. Olson

University of Washington

Rae, J. R., & Olson, K. R. (2018). Test–retest reliability and predictive validity of the Implicit Association Test in children. *Developmental Psychology, 54*(2), 308-330.

Correspondence concerning this article should be addressed to James R. Rae, who is now at the Department of Experimental Psychology, University of Oxford, 15 South Parks Road, Oxford OX1 3UD, United Kingdom. E-mail: james.rae@psy.ox.ac.uk

Abstract

The Implicit Association Test (IAT) is increasingly used in developmental research despite minimal evidence of whether children's IAT scores are reliable across time or predictive of behavior. When test-retest reliability and predictive validity have been assessed, the results have been mixed, and because these studies have differed on many factors simultaneously (lag-time between testing administrations, domain, etc.), it is difficult to discern what factors may explain variability in existing test−retest reliability and predictive validity estimates. Across five studies (total N = 519; ages 6-11 years old), we manipulated two factors that have varied in previous developmental research – lag-time and domain. An internal meta-analysis of these studies revealed that, across three different methods of analyzing the data, mean test-retest (*r*s of .48, .38, and .34) and predictive validity (*r*s of .46, .20, and .10) effect sizes were significantly greater than zero. While lag-time did not moderate the magnitude of test-retest coefficients, whether we observed domain differences in test-retest reliability and predictive validity estimates was contingent on other factors, such as how we scored the IAT or whether we included estimates from a unique sample (i.e., one containing gender typical and gender diverse children). Recommendations are made for developmental researchers that utilize the IAT in their research.

Keywords: Implicit Association Test, test−retest reliability, predictive validity, implicit attitudes, implicit identity

Test-Retest Reliability and Predictive Validity of the Implicit Association Test in Children

The Implicit Association Test (IAT; Greenwald, Schwartz, & McGhee, 1998) is the most widely used measure of implicit cognition (Payne & Gawronski, 2010), and is increasingly used in developmental research (Dunham & Emory, 2014). Indeed, a search of the Google Scholar and the PsychInfo databases (using terms such as "IAT", "implicit association test", "children", "childhood", etc.) and cross-referencing published articles yielded 71 reports, papers, and theses/dissertations reporting IAT results with child samples (defined as mean age of under 12 years).[1,2] The number of identified reports per year is plotted in Figure 1, demonstrating a clear upward trajectory. However, despite the practical and theoretical implications of this work (Gonzalez, Steele, & Baron, 2016; Olson & Dunham, 2010), we know surprisingly little about whether or not this increasingly utilized measure is reliable or valid when used with young children. In the current work, we assess whether the IAT is reliable across testing administrations (test-retest reliability) and predictive of behavior (predictive validity)—cornerstones of good assessment tools.

Why Does Test-Retest Reliability and Predictive Validity Matter?

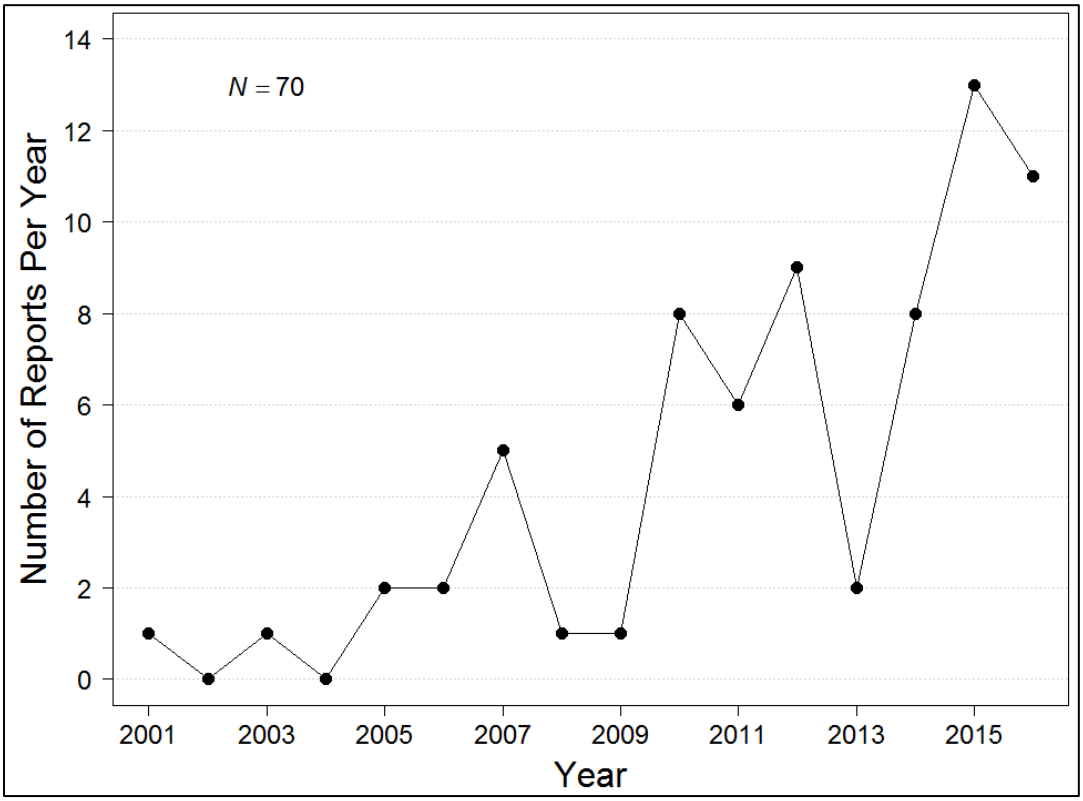
Test- Retest Reliability. Establishing adequate test−retest reliability is critical insofar as researchers believe themselves to be assessing meaningful, stable individual differences, rather than momentarily-accessible associations. Indeed, the IAT was developed as an individual difference measure of implicit cognition in adults (Greenwald et al., 1998), and in the literature

[1] For our literature search, we included any report containing a sample with a mean age below 12.0 years. For longitudinal research, we considered the mean age of the sample at first time point. We also included reports from samples in which the mean was not reported, but the midpoint (or weighted average) of the age range was below 12.0 years. For example, if a report indicated that twenty 11-year olds, five 12-year-olds, and five 13-year-olds were in the sample, this report would be included in that the weighted average of the age range is (.5×10) + (.25×12) + (.25×13) = 11.25 years.

[2] The cut-off date for the search was January 1, 2017. See Reference section for all identified reports.

to date, developmental researchers have often assumed that the IAT with children *is* a reliable indicator of a trait-like individual difference. For example, several lines of research have documented how children's racial attitudes (indexed via the IAT) vary as a function of their social status (e.g., Dunham, Newheiser, Hoosain, Merrill, & Olson, 2014) or preference for high-status groups (e.g., Newheiser, Dunham, Merrill, Hoosain, & Olson, 2014). This work – and much other work using the IAT with child samples - assumes that there is meaningful child-specific variance in IAT scores that can be explained by individual difference factors (e.g., experience or preferences).

Figure 1. Number of publications, reports, and theses using the IAT in child samples (12 years or younger) plotted against the publication year (2000-2016).



Note: One in press report identified by our search is not included in Figure 1.

Even though developmental researchers have treated the IAT as an individual difference measure and test−retest reliability is the primary test for sensitivity to individual differences (Greenwald & Nosek, 2001), there have been only a handful of studies examining test-retest reliability using the IAT with children to date. Unfortunately, the findings are mixed and difficult to interpret. Table 1 presents the test−retest reliability coefficients from six reports using the IAT with children reported in previous developmental research. Critically, these values range from −.17 (Corenblum & Armstrong, 2012) to +.62 (Bruni & Schultz, 2010). Further, directly comparing these results is nearly impossible as the studies themselves drastically differed in the structure and presentation of the IAT (e.g., number of trials, response type, stimulus presentation), domain of study, age of participants, and many other factors (e.g., sample-related differences). Therefore, it is difficult to discern what differences may explain variability in test−retest reliability estimates.

Predictive Validity. Predictive validity is another key feature of good measures (Kimberlin & Winterstein, 2008), and is particularly crucial for the IAT because predicting behavior is a key motivation behind the use of implicit measures (e.g., McConnell & Leibold, 2001). While a meta-analysis of a large number of adult studies found that the IAT is predictive of behavior in numerous domains (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; cf. Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013), we have relatively few tests of the predictive validity of the IAT in child participants, and what evidence we have is quite mixed. Table 1 presents the 17 reports of predictive validity of the IAT in the developmental literature. As the table indicates the literature has reported results ranging from no relationship between the IAT and behavior (r =−.02; Pieters, van der Vorst, Engels, & Wiers, 2010) to a strong positive relationship between the IAT and behavior (r =.52, Cvencek, Meltzoff & Greenwald, 2011).

Again, however, interpretation is difficult because of the numerous factors (structure of the IAT, age of respondents, etc.) that varied across studies reporting predictive validity estimates.

What Explains Variability in Test-Retest Reliability and Predictive Validity Estimates?

As alluded to above, the variation observed in past work on test-retest reliability and predictive validity could be explained by a wide range of factors, which we explore below.

Domain Differences. Studies differing in test−retest reliability and predictive validity estimates in Table 1 have come from very different domains. Within adult populations, domain moderates the magnitude of IAT test-retest reliability and predictive validity estimates. For example, IAT scores assessing political attitudes have much stronger test-retest reliability than IAT scores indexing attitudes toward the self (i.e., self-esteem; Bar-Anan & Nosek, 2014). Similarly, scores on the Race Attitude IAT have been shown to predict interracial behavior (e.g., Greenwald et al., 2009), while scores on the Self-Esteem IAT have not been found to correlate with theoretically relevant criterion variables (e.g., Bosson, Swann, & Pennebaker, 2000; Buhrmester, Blanton, & Swann, 2011). Thus, domain differences may explain variability in the test-retest reliability and predictive validity estimates presented in Table 1.

Age of Respondents. Discrepancies in results across the reviewed studies could also be explained by differences in the age of respondents. While all studies reported in Table 1 had a mean age of less than 12 years, there was still substantial age variability across reports with mean ages ranging from 4.46 years (Cvencek et al., 2011) to 11.67 years (Lemmer, Gollwitzer, & Banse, 2011). As such, differences across studies could be driven by developmental changes in the cognitive abilities that the IAT relies upon, such as task switching, response inhibition, and reaction time (Dunham & Emory, 2014). Although there has been no direct evidence that these early developmental changes affect performance on the IAT, changes occurring later in life

Table 1. Authors, publication year, sample size, domain, age, lag-time (across Time 1 and 2) and parameter estimates from identified articles, reports, and theses that reported either test-retest reliability or predictive validity estimates of the IAT in child samples.

Authors (year)	*N*	Domain	Age (years)	Lag-Time	Test-retest reliability (*r*)	Predictive Validity (|*r*|)
Bruni (2007) Study 1	52	Nature identity	*M* = 9.69	Same session	.45	
Bruni and Schultz (2010) Study 3	30	Nature identity	Range =10 - 11	Same session	.62	.41, .45, .36[1]
Pieters et al. (2010) Study 1	99	Alcohol attitudes	*M* = 10.17			.02
Pieters et al. (2010) Study 2	35	Alcohol attitudes	*M* = 11.36			.39
van Goethem et al. (2010) - Version 1	240	Bullying attitudes	*M* = 11.41			.07, .02, .06
van Goethem et al. (2010) - Version 2	240	Bullying attitudes	*M* = 11.41			.07, .06, .10
Cvencek et al. (2011) Study 2	75	Gender attitudes	*M* = 4.46			.52
Dunham et al. (2011) Study 1	33	Intergroup attitudes	*M* = 5.40			.50
Dunham et al. (2011) Study 2	43	Intergroup attitudes	*M* = 5.50			.41
Grumm et al. (2011)	115	Aggression identity	*M* = 9.70			.23[2]
Corenblum and Armstrong (2012)	196	Self-esteem; Race attitudes	*M* = 8.84	1-year	.18, –.17[3]	
O'Connor et al. (2012)	376	Alcohol attitudes	*M* = 11.10			.06[4]
Galdi et al. (2014) – Females[5]	120	Math-Gender stereotypes				.27
Galdi et al. (2014) - Males	120	Math-Gender stereotypes				.05
Cvencek et al. (2015)	299	Math Self-concept Math-Gender stereotypes Gender identity	*M* = 9.35			.15, 16, .06
Diesendruck and Menahem (2015)	48	Ethnic attitudes	*M* = 6.50			.51[6]
Lemmer et al. (2014)[7]	574	Aggression identity	*M* = 11.67	5-months – 1.5-years	.14 - .36[8]	
	317	Aggression identity	*M* = 11.61			.20, .20, .14, .17,.12, .19, .06, .11
Leeuwis et al. (2015)	330	Self-esteem	*M* = 11.20	1-year	.29	.05, .15, .06, .05, .07, .05, .04, .05
Meyer and Gelman (2016)	76	Gender stereotypes	*M* = 6.45			.08
Vander Heyden et al. (2016)	237	Gender stereotypes	*M* = 10.82			ns[9]
Williams and Steele (2016) – Study 2	144	Race attitudes	*M* = 7.98	Same session	.24	

(Table 1 continued)

Notes:

1. There were nine null IAT-behavior correlations, though the magnitude of these correlations were not reported.
2. Estimate is a standardized regression coefficient.
3. Estimate is correlation between "error-free" latent variables.
4. We converted the reported Cohen's *d* between those that had and had not tried alcohol to the *r* metric using the effect size conversion spreadsheet provided by Lakens (2013).
5. Mean age for the entire sample was 6.47 years. Age by gender was not reported.
6. The authors reported a significant correlation for male participants, but did not report the magnitude of a non-significant correlation for female participants.
7. Only a subset of the sample completed measures used for tests of predictive validity.
8. The authors assessed stability of the IAT across four time points, but only reported the range of estimates.
9. The authors reported that IAT-criterion correlations < .18 for male participants, and did not report the magnitude of any correlations for female participants. All correlations were non-significant.

(age-related slowing among elderly participants) must be considered when analyzing and interpreting IAT scores (e.g., Hummert, Garstka, O'Brien, Greenwald, & Mellott, 2002). Thus, it seems plausible that changes in cognitive ability during childhood may similarly affect IAT performance.

Structure of IAT. Another factor that likely affects the test-retest reliability and predictive validity of children's IAT scores is the structure of the IAT administered in any given report. Despite using the same name (the IAT), researchers have taken a number of liberties in the design of the IAT. While these design decisions cannot be concisely summarized in Table 1, as just a sample of these differences, the number of trials (range = 80 – 228), the modality of stimulus presentation (only visual, only audio, or both visual and audio), scoring algorithm, and response type (e.g., paper and pencil, keyboard press, external buttons, computer mouse movements, etc.) differed across studies. As previous work with adult samples has suggested that even minor procedural differences can impact the reliability and validity of the IAT (Bar-Anan & Nosek, 2014), it seems possible that these design decisions may similarly affect the extent to which the IAT is reliable and valid in child samples. Lastly, while not reliant on the IAT structure per se, past developmental research has used IAT scoring algorithms that differ in the

criterion used to remove trial latencies deemed as outliers (e.g., Bruni & Schultz, 2010; Newheiser et al., 2014) and/or participants with too many error responses (e.g., Cvencek et al., 2011; Newheiser et al., 2014).

Lag-Time. Within the studies exploring test-retest reliability, the amount of time between testing administrations varies considerably (as can be seen in Table 1). The strongest test-retest reliability occurs in studies in which IATs were completed during the same testing session (e.g., Bruni & Schultz, 2010) while weaker evidence comes from work in which administrations were separated by one year (e.g., Corenblum & Armstrong, 2012), which is not surprising in that lower reliability coefficients are to be expected with longer lag-times (Fraley & Roberts, 2005; Ozer, 1999). However, lag-time may especially impact test−retest reliability in developmental samples, as changes in IAT scores over a one-year period could reflect problems with the measure, but could also reflect real developmental changes in the underlying construct (rather than unreliability of IAT scores; Carmines & Zeller, 1979).

Can Specific Factors Be Isolated? While we have made the case that four factors may explain variability in test−retest reliability and predictive validity estimates from previous developmental research, the critical question is whether the unique effects of any one factor can be isolated. For example, if two estimates of test−retest reliability come from reports varying only on domain assessed while the other three factors (e.g., age of respondents, structure of the IAT, lag-time) are held constant, then there is at least preliminary evidence that domain differences may explain some of the observed variability in test-retest coefficients. In contrast, if more than one factor differs across reports (e.g., domains assessed and lag-time), such inferences are not possible. Unfortunately, this latter scenario exemplifies the research reviewed here. For example, studies providing strong (e.g., Bruni & Schultz, 2010) and weak (e.g., Corenblum &

Armstrong, 2012) evidence of test–retest reliability differ on nearly every dimension—domain (nature identification vs. race attitudes), age of participants (10-11 years vs. 8.84 years), length (168 trials vs. 180 trials), stimulus presentation (moving vertically from the top of the screen to the bottom vs. presented center screen), lag-time between testing administration (same testing session vs. one year), etc. As such, identifying the key factor that explains why test–retest reliability differed across these studies is impossible.

Overview of the Present Research

Across five studies, we sought to evaluate test-retest reliability and predictive validity of the IAT with young children by either holding constant or systematically varying (across studies) the factors that complicate interpretation of the previously reviewed literature: domain assessed, age of participants, structure of the IAT, and lag-time. In all studies, we used one version of a child-adapted IAT with the same exclusion criteria and tested only children between 6 and 11 years of age (at the time of the first testing administration). We obtained results for two domains (race attitudes and gender identity), and at three time lags (10-minutes, 1-month, and 1-year [gender-identity only]). Comparing results across lag-times in the same domain allowed us to estimate the effects of time on test-retest reliability, while comparing across studies in different domains allowed us to evaluate the effect of content domain on both predictive validity and test-retest reliability. Study 1 assessed the predictive validity and test-retest reliability of the Race Attitude IAT across a lag-time of 10-minutes while Study 2 assessed these constructs with a 1-month time lag. Studies 3 and 4 replicated Studies 1 and 2 using the Gender Identity IAT. Finally, Study 5 examined the test-retest reliability and predictive validity of the Gender Identity IAT across 1-year in a sample that included children with diverse gender identities (e.g., transgender and gender-nonconforming children).

Race Attitude IAT in Children

Studies 1 and 2 investigated test-retest reliability of the Race Attitude IAT and an explicit race attitude measure across a 10-minute lag-time (Study 1) and 1-month lag-time (Study 2). This 10-minute lag-time likely provides an upper limit on the test-retest reliability of each measure. Both studies also tested whether the IAT and/or an explicit attitude measure predicted a race-related behavior (selection of an art contest winner). One methodological difference between the two studies was our explicit measure. We used an absolute measure of explicit attitudes in Study 1 (see below), but after discovering work showing that correlations between the IAT and explicit measures tend to be stronger when explicit measures are relative in nature (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005), we used a relative explicit race attitude measure in Study 2.

Study 1: 10-Minute Lag-Time

Method. Data, analysis scripts, and materials[3] for Study 1 (and Studies 2-4) are available from the first author or via the Open Science Framework (OSF): https://osf.io/c85nr/?view_only=113d4aa7e5b2474d8d262e2a03ec5833.

Participants. One hundred and eight White American children between the ages of 6 and 11 years participated in the study at a research lab located at the University of Washington. Only children between 6 and 11 years of age were eligible to participate (also true of all subsequently presented studies). One participant was identified by parental report as Asian-American and another did not complete any of the study procedures, and therefore, data from both was

[3] The IAT script used in all studies is publicly available via OSF, as are the data from Studies 1-4 (there are constraints about posting data from Study 5 due to the sensitive nature of the participant population and concerns about identifiability). The images used in the IAT and explicit measure, as well as drawings used in the behavioral measures, can be obtained by either contacting the first author or requesting access to a private component to this research posted on OSF.

excluded from all analyses. An additional five participants did not complete a subset of study measures, and an IAT score (from Time 2) from one other participant was removed via established IAT exclusion criteria (see below). Partial data from these participants is reported whenever possible. Table 2 shows the sex[4], race, and age characteristics of participants for whom any data is reported in Study 1 (and all subsequently presented studies).

Measures.

Race-related behavior. To portray the activity as unrelated to the research study for which the lab visit was scheduled, participants were told a cover story that children visiting the lab space often colored pictures while waiting for their activities to start. Participants were then told of an ongoing art contest open to children that had previously visited the lab and asked if they would be willing to vote for a winner between two finalists. All participants agreed to vote. Participants were then presented two pictures, both of which were the same template of a house taken from a coloring book. Each house differed in color but had the same color scheme (one house had a brown roof and blue walls while the other house had blue walls and a brown roof). Children younger than 8-years-old were shown a more basic house template, while children 8-years and older were shown a more elaborate house taken from the same coloring book.

On the art contest entries, we paper-clipped colored photographs of White and Black children (approximately 3 × 3 inch size), with the White child attached to one entry and the Black child attached to the other entry. The photographs were approximately matched to the

[4] Due to the gender diversity of participants in Study 5, we differentiate participants' *sex* (based upon one's "physical and biological traits"; American Psychological Association, 2015) and *gender* (based upon one's internal sense of identity). When referring to participant sex, we use the terms "male" and "female". In all studies, we use this as the primary distinction for data analysis and categorize children based on their sex at birth. In Study 5, we use the terms "boy" and "girl" (or "non-binary") to refer to the child's gender in everyday life. For example, a child who was identified by doctors at birth as a female and identifies as a girl and a child who was identified by doctors at birth as a male but identifies as a girl would both be termed "girls" (as their gender) even though their sexes would be different.

participant's sex and age. Photographs and house pictures were counterbalanced across participants. While presenting the art contest entries, the experimenter identified the children as the creator of each picture by pointing at the target before nonchalantly unclipping the photo and placing it next to the picture. Finally, participants were asked which child should win the art contest and receive a prize.

Explicit racial attitudes. Participants used a 6-point smiley-face scale (1= "really don't like"*;* 6 = "really like") to indicate how much they liked 8 children depicted serially in photographs on a laptop computer. Four of the photographs were of Black children and the other four were of White children (race of evaluated child alternated on each trial). All photographs were in color, edited so that each child's head and shoulders were presented on a white background, and approximately matched to the participant on both sex and age.[5] An explicit preference index for Whites relative to Blacks was computed by subtracting the mean liking of the four photographs of White children from the mean liking of the four photographs of Black children. Scores greater than zero indicate preference for Whites over Blacks, scores of zero indicate no preference, and scores less than zero indicate preference for Blacks over Whites. Participants rated different sets of photographs at Time 1 and Time 2, and the order in which the two sets of photographs were rated was counterbalanced across participants. One photograph set presented a White child on the first trial while the other set presented a Black child on the first trial.

[5] A racially and ethnically diverse group of undergraduate research assistants located the photographs used in the explicit attitude measure and rated all photographs on dimensions of perceived age, warmth, and attractiveness. Each prospective photograph was rated by between 8 and 15 raters. Internal consistencies for both the age ($\alpha = .97$) and warmth ratings ($\alpha = .92$) were excellent, but internal consistency for the attractiveness ratings was acceptable ($\alpha = .65$). Two sets of photographs were created for each sex and age group by assembling photographs of White and Black children that were similar on each dimension. Given the low internal consistency of the initial attractiveness ratings, three additional undergraduate research assistants (that did not provide the initial evaluations) rated the White and Black children within each set as equally attractive both *within* and *between* the final sets of photographs.

Table 2. Sex, race/ethnicity, and age of all participants for whom any data are reported. Age was calculated at Time 1.

	Study 1 (*N*=106)	**Study 2 (*N*=107)**	**Study 3 (*N*=102)**	**Study 4 (*N*=107)**	**Study 5 Gender Diverse (*N*=41)**	**Study 5 Gender Typical (*N*=56)**
Sex						
Male	56	56	55	54	24	24
Female	50	51	47	53	17	32
Race/Ethnicity						
American Indian/Alaska Native	0	0	0	0	0	0
Asian	0	0	15	9	2	2
Black/African American	0	0	0	6	0	0
Hispanic	0	0	3	4	0	1
Native Hawaiian/Other Pacific Islander	0	0	2	2	0	0
White	106	107	63	65	31	38
Two or more races/ethnicities	0	0	9	14	8	15
Other or not reported	0	0	10	7	0	0
Age						
6 years	26	24	13	31	9	11
7 years	27	19	16	18	3	6
8 years	1	16	14	20	10	15
9 years	3	13	16	24	6	7
10 years	27	17	25	5	9	12
11 years	22	18	18	9	4	5

Implicit racial attitudes. The IAT is a latency-based measure that indexes the relative association between pairs of target and attribute concepts. The child-adapted IAT used throughout this paper (but also in previous research Newheiser & Olson, 2012; Olson, Key, & Eaton, 2015) used images of children that were unambiguous members of their categories (utilizing criteria outlined by Lane et al, 2007) and approximately age-matched to participants, required participants to respond to trials by pushing one of two computer keys that were marked with stickers, and contained 70 trials (see Table 3 for full design). The Race Attitude version of this IAT measures the speed with which participants pair pictures of White and Black children (of both sexes) with pictures of pleasant (i.e., a wrapped gift, a gumball machine, a litter of puppies, and a portion of ice cream) and unpleasant stimuli (i.e., a house on fire, a car crash, a tarantula, and a broken house-window). Children that respond more rapidly when photographs of White children are paired with pleasant images relative to unpleasant images are taken to hold an implicit preference for Whites over Blacks.

A summary score for the IAT (called a *D*-score), which is conceptually similar to a Cohen's *d*, is calculated by dividing the mean latency difference between the third and fifth block using the "inclusive" standard deviation (i.e., all retained latencies completed within the third and fifth block). IAT *D*-scores can take on values from +2 to −2. Consistent with recommended scoring procedures, we excluded all trials greater than 10,000-ms and data from participants completing more than 10% of trials in under 300-ms (Greenwald, Nosek, & Banaji, 2003). The Race Attitude IAT was scored such that positive values indicate higher preference for Whites over Blacks. In line with the most common approach to assessing the internal consistency of the IAT (e.g., Greenwald et al., 2003; Schmukle & Egloff, 2004), we computed the split-half correlation (Spearman-Brown adjusted) between *D*-scores computed on two subsets (10 trials

each) of IAT trials. These adjusted correlations were .59 and .64 for the IAT administered at Time 1 and Time 2, respectively.

Table 3. Structure of child-adapted Race Attitude IAT used in Studies 1 and 2. The Gender Identity IAT used in Studies 3 through 5 used the target categories "Male" and "Female" and the attribute categories of "Me" and "Not Me".

			Response key assignment	
Block	***N* trials**	**Task**	**Left key ('D')**	**Right key ('K')**
1	10	Target discrimination	White	Black
2	10	Attribute discrimination	Good	Bad
3	20	Initial combined task	White + Good	Black + Bad
4	10	Reverse target discrimination	Black	White
5	20	Reversed combined task	Black + Good	White + Bad

Procedure. All participants were tested individually and completed the study tasks in a fixed order (also true of Studies 2 through 4). After obtaining consent from parents as well as verbal assent from children (and written assent from participants age 9 and above), an experimenter requested the participant vote in an "art contest" (the race-related behavioral measure) before beginning their activities. Next, a second experimenter (in order to maintain the appearance that the art contest was unrelated to the research study) entered the testing room and administered the explicit attitude measure and Race Attitude IAT.[6] Because we anticipated that presenting the explicit measure first would produce smaller carry-over effects between measures, the explicit attitude measure was always administered prior to the Race Attitude IAT. After a five-minute filler task, participants again completed both attitude measures. With the filler task and the time it took to complete each measure, there were approximately 10 minutes between the start of the first and second testing administration of each measure. The testing session concluded with an opportunity for the participants to ask questions about the research.

[6] For two participants, one experimenter administered all study measures because a second experimenter was not available.

Results. Degrees of freedom vary across statistical tests due to incomplete data on various study measures (also true in subsequent studies). We streamlined our results by using averaged scores from measures collected at Time 1 and Time 2 in our predictive validity analyses. However, results computed separately for measures collected at Time 1 and Time 2 are available in the Supplemental Material. If participants did not have scores for both Time 1 and Time 2 measures (e.g., a participant completed the IAT at Time 1 but not at Time 2), predictive validity results were computed using the score they did have (rather than dropping them from the analyses altogether).

Table 4 shows means, standard deviations, and zero-order correlations for both demographic and measured study variables. Based upon conventions for small, moderate, and large effect sizes (*d*s of .20, .50, and .80, respectively; Cohen, 1992), tests comparing measured variables to zero (IAT and explicit measure) or chance responding (art contest) revealed that participants showed a large implicit (d =1.22), but relatively small explicit (d=.26), preference for Whites over Blacks, and were at chance-responding in voting for the White and Black child to win the art contest. Correlations in Table 4 indicate (a) scores on the IAT and explicit measure were not significantly correlated, (b) female participants (relative to male participants) tended to have higher implicit preference for Whites over Blacks and were more likely to vote for a White child to win the art contest, and (c) older participants tended to have higher implicit – but lower explicit - preference for Whites over Blacks and were less likely to vote for the White child to win the art contest.

Test-retest reliability. Test-retest reliability was assessed via the zero-order correlation between scores on the attitude measures collected before and after the filler task. Results indicated that scores on the IAT, $r(99)= .34$, $p < .001$, 95% CI [.15, .50], and explicit measure,

$r(101) = .63$, $p < .001$, 95% CI [.50, .73], were correlated across the 10-minute lag-time. To test whether participant age predicted the stability of scores on either attitude measure, we correlated age with the absolute difference in scores on the IAT and explicit measure between Time 1 and Time 2. Results indicated that participant age was not significantly correlated with the absolute difference between IAT scores, $r(99) = .00$, $p = .994$, 95% CI [−.20, .20], or explicit attitude scores, $r(101) = -.09$, $p = .348$, 95% CI [−.28, .11].

Predictive validity. As shown by the zero-order correlations in Table 4, participants with higher implicit, $r(103) = .28$, $p = .003$, 95% CI [.09, .45], or explicit preference, $r(104) = .25$, $p = .010$, 95% CI [.06, .42], for Whites over Blacks were more likely to vote for a White child to win the art contest. We next conducted multivariate analyses to test two further questions: First, we tested the IAT and explicit measure as simultaneous predictors of performance on the art contest to see if implicit and explicit attitudes were unique predictors of race-related behavior. We tested this question using multiple logistic regression (with art contest performance as the dichotomous dependent variable; Black child=0; White child=1). To simplify interpretation, coefficients were transformed into odds ratios (via the exponential function raised to the power of the raw regression coefficient). Values significantly larger than one indicate a positive association between the predictor and dependent variable, whereas values significantly less than one indicate a negative association between the predictor and the dependent variable. As shown in Table 5, scores on the IAT and explicit measure simultaneously predicted performance on the art contest, such that participants with stronger implicit or explicit preference for Whites over Blacks were more likely vote for a White child compared to a Black child (see Model 1).

A second multiple logistic regression model further introduced participant age and sex as predictors, which allowed us to test whether implicit and explicit race attitudes predicted

Table 4. Means, standard deviations, and zero-order correlations for demographic and measured variables in Study 1. Scores on the explicit attitude measure and IAT were averaged across Time 1 and Time 2. The *p*-values were computed with *N*'s ranging from 105 to 106.

	Variable	Mean	SD	*t*	*d*	1	2	3	4
1	Sex (Female = 1)	.47	.50	-	-	-	-	-	-
2	Age	8.80	1.99	-	-	.10	-	-	-
3	Explicit Measure (average)	.19	.74	2.58*	.26	.11	−.53***	-	-
4	IAT (average)	.44	.36	12.57***	1.22	.32***	.20*	−.08	-
5	Art Winner (White child = 1)	.44	0.50	1.14[1]	.12[2]	.22*	−.28**	.25*	.28**

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$, † = $p < .10$

1. The percentage of White children chosen to win the art contest was tested from chance using a one-sample proportion test. The test-statistic reported for the test is a χ^2.
2. The reported effect size for the proportion of children choosing a White child as the winner on the art contest (from chance responding of 50%) is Cohen's *h*.

Table 5. Estimates, standard errors, and 95% CIs of *standardized* odds ratios (all predictors other than sex were standardized) of hierarchical multiple logistic regression models regressing race of the child selected as the winner on the art contest (Black = 0; White =1) on (a) racial attitude measures (Model 1) and (b) race attitude measures and demographic variables in Study 1 (Model 2).

	Model 1			Model 2		
Variable	OR	SE (OR)	95% CI	OR	SE (OR)	95% CI
Explicit Measure (average)	1.86**	1.26	1.21 - 2.99	1.26	1.31	.74 - 2.16
IAT (average)	2.11**	1.28	1.34 - 3.52	2.23**	1.31	1.35 - 3.94
Sex (Female =1)	-	-	-	2.09	1.61	.82 - 5.44
Age	-	-	-	.47**	1.32	.27 - .80
Nagelkerke's R^2		.20			.30	

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$, † = $p < .10$

performance on the art contest above and beyond participant demographic characteristics. As shown in Table 5, results from this model indicated that scores on the IAT (but not explicit measure) continued to predict performance on the art contest (see Model 2). Additionally, while participant sex did not predict performance on the art contest, older participants were less likely to vote for a White child compared to a Black child.

A more intuitive sense of the predictive utility of the IAT was obtained by comparing the predicted probability of choosing the White child to win the art contest at ± 1 standard deviation on the IAT while holding other values at their respective means and participant sex at zero (female participants). For a participant at −1.0 standard deviations below the IAT mean, the predicted probability of voting for the White child to win the art contest was 32.9%. By contrast, the predicted probability of voting for the White child to win the art contest at +1.0 standard deviations above the IAT mean was 70.3%.

Study 2: 1-Month Lag-Time

Method. Unless otherwise noted, all methods and measures in Study 2 were identical to those used in Study 1.

Participants. One hundred and nine White American children between the ages of 6 and 11 years old participated in the study in either a quiet space in their school (*N*=96) or a research lab (*N*=13) located at the University of Washington. Due to experimenter error, two participants younger than 6 years completed the study and data from these participants are excluded from all analyses. At Time 1, three participants were absent and two others failed to complete all study procedures. At Time 2, five participants were absent and one other failed to complete all IAT procedures. Partial data from these participants is reported whenever possible.

Measures.

Explicit racial attitudes. Across eight trials, participants were shown pairs of children (one White and the other Black) and asked to choose which child they liked better. Using the same pre-rated photographs from Study 1, we created target pairs that were (a) sex-matched to the participant and (b) approximately the same age in middle childhood. To index relative preference for Whites over Blacks, we summed the number of trials in which the participant indicated liking the White child more than the Black child. We then subtracted 4 from all scores so that positive scores indicated a pro-White preference, zero indicated no preference, and negative scores indicated a pro-Black preference.

Implicit racial attitudes. The Spearman-Brown corrected split-half reliability of the Race Attitude IAT at Time 1 and Time 2 was .66 and .61, respectively.

Lag-time. Participants completed study measures approximately 3 to 4 weeks apart (M = 28 days, SD = 4 days), with the art contest administered at the end of Time 2.

Results. Table 6 shows means, standard deviations, and zero-order correlations for both demographic and measured study variables. Effect sizes from tests comparing measured variables to zero (IAT and explicit measure) or chance responding (art contest) indicated that participants showed a large implicit preference (d= 1.29) and a medium explicit preference (d= .53) for Whites over Blacks, but were at chance in voting for the White and Black child to win the art contest. Correlations in Table 6 reveal that (a) scores on the IAT and explicit measure were positively correlated, (b) female participants (relative to male participants) tended to have stronger implicit preference for Whites over Blacks, and (c) older participants tended to have lower explicit preference for Whites over Blacks and were less likely to vote for the White child to win the art contest.

Table 6. Means, standard deviations, and zero-order correlations for demographic and measured variables in Study 2. Scores on the explicit attitude measure and IAT were averaged across Time 1 and Time 2. The *p*-values were computed with *N*'s ranging from 101 to 107.

	Variable	Mean	SD	*t*	*d*	1	2	3	4
1	Sex (Female= 1)	.48	.50	-	-	-	-	-	-
2	Age	8.70	1.85	-	-	.05	-	-	-
3	Explicit Measure (average)	1.02	1.91	5.55***	.53	.15	−.40***	-	-
4	IAT (average)	.46	.36	13.05***	1.29	.23*	.03	.31**	-
5	Art Winner (White child = 1)	.51	0.50	.04[1]	.02[2]	.15	−.44***	.28**	.17†

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$, † = $p < .10$

1. The percentage of White children chosen to win the art contest was tested from chance using a one-sample proportion test. The test-statistic reported for the test is a χ^2.
2. The reported effect size for the proportion of children choosing a White child as the winner on the art contest (from chance responding of 50%) is Cohen's *h*.

Table 7. Estimates, standard errors, and 95% CIs of *standardized* odds ratios of hierarchical multiple logistic regression models regressing race of the child selected as the winner on the art contest (Black = 0; White =1) on (a) racial attitude measures (Model 1) and (b) race attitude measures and demographic variables in Study 2 (Model 2).

	Model 1			Model 2		
Variable	OR	SE (OR)	95% CI	OR	SE (OR)	95% CI
Explicit Measure (average)	1.74**	1.27	1.10 - 2.83	1.06	1.33	.60 - 1.88
IAT (average)	1.18	1.27	.75 - 1.90	1.34	1.29	.82 - .2.24
Sex (Female = 1)	-	-	-	2.13	1.62	.84 - 5.65
Age	-	-	-	.35***	1.32	.19 - .59
Nagelkerke's R^2	.11			.31		

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$, † = $p < .10$

Test-retest reliability. Scores on the IAT, $r(96) = .25$, $p = .012$, 95% CI [.05, .43], and the explicit measure, $r(95) = .67$, $p < .001$, 95% CI [.54, .77], were correlated across the 1-month lag-time. Participant age was not correlated with the absolute difference in scores on the IAT, $r(96) = -.01$, $p = .889$, 95% CI [−.21, .19], or explicit measure, $r(95) = .09$, $p = .366$, 95% CI [−.11, .29], across Time 1 and Time 2.

Predictive validity. As shown in Table 6, scores on the explicit measure predicted performance on the art contest, $r(99) = .28$, $p = .005$, 95% CI [.09, .45], but IAT scores did not, $r(99) = .17$, $p = .098$, 95% CI [−.03, .35]. Table 7 shows that when both attitude measures were tested as simultaneous predictors of art contest performance in a multiple logistic regression model, scores on the explicit measure (but not the IAT) predicted performance on the art contest (see Model 1). However, after introducing participant age and sex as predictors into the model, only age was a significant predictor art contest performance (see Model 2).

Discussion of Race Attitude Studies

Results from Studies 1 and 2 indicated that White children's scores on the Race Attitude IAT were significantly correlated across lag-times of 10-minutes ($r = .34$) and 1-month ($r = .25$). Importantly, these studies provide insight into whether IAT scores capture stable associations in young children that would not have been possible had we examined test-retest reliability at only one lag-time. Indeed, while researchers frequently use test-retest coefficients from one lag-time to conclude that a construct is stable (Fraley & Roberts, 2005), data from at least two different lag-times (preferably more) are necessary to examine whether the strength of test-retest reliability decays over time. Thus, the similarity of our estimates across different lag-times provides some evidence that the version of the Race Attitude IAT tested here may index more durable trait-like associations in young children. Near the end of the paper we present the results

of an internal meta-analysis formally testing the effects of lag-time on the magnitude of test-retest reliability estimates.

Studies 1 and 2 also provided the first tests of whether the Race Attitude IAT is predictive of a race-related behavior in children. Results were mixed: On one hand, results in Study 1 revealed that participants with stronger implicit preference for Whites over Blacks tended to vote for a White child to win the art contest (r = .28). Moreover, we found that scores on the IAT (but not explicit measure) predicted performance on the art contest even when controlling for demographic variables. On the other hand, results from Study 2 indicated that scores on the explicit attitude measure (r = .28; but not IAT, r =.17) predicted performance on the art contest, though neither implicit or explicit attitudes predicted scores on our behavioral measure after controlling for demographic variables. Although the predictive validity results across Studies 1 and 2 were inconsistent, the internal meta-analysis near the end of the paper presents a summary of the predictive validity evidence across all studies.

Studies 1 and 2 produced several secondary findings. First, we found that participants showed particularly strong implicit preference for Whites over Blacks (D-scores = .44, .45; cf. Baron & Banaji, 2006). While this result may stem from participants living in racially homogenous environments (comparable results were obtained from White children attending a majority White school in Connecticut, Newheiser & Olson, 2012), Studies 1 and 2 differed from previous work in other factors as well (e.g., IAT design features). Second, consistent with research showing that relative (compared to absolute) explicit measures tend to correlate more strongly with IAT scores (Hofmann et al., 2005), we found that IAT scores correlated with a relative explicit race attitude measure (Study 2) but not an absolute explicit race attitude measure (Study 1). However, an alternative explanation is that the relationship may just be a small and/or

unreliable one and therefore will sometimes be observed and sometimes will not be observed. Third, while men generally hold stronger implicit social preferences than women (Nosek et al., 2007), we found across Studies 1 and 2 that female participants (relative to male participants) tended to have stronger implicit preference for Whites over Blacks. Finally, age did not consistently predict stability (e.g., differences in IAT scores between Time 1 and Time 2) or magnitude of IAT scores.

Gender Identity IAT in Children

While Studies 1 and 2 provided initial evidence for the test–retest reliability and – to a lesser extent - predictive validity of children's Race Attitude IAT scores, it is unclear whether these results hold for IATs assessing other constructs. As previously discussed, research from adult participants shows clear domain differences in the test–retest reliability (Bar-Anan & Nosek, 2014) and predictive validity of IAT scores (Greenwald et al., 2009). Studies 3 through 5 focused on the domain of gender identity rather than racial attitudes, investigating the test-retest reliability across a 10-minute lag-time (Study 3), a one-month lag-time (Study 4), and a one-year lag-time. These studies also assessed the predictive validity (all studies) of a Gender Identity IAT. In Study 5, we also tested whether test-retest reliability and/or predictive validity estimates were stronger within a more diverse sample by including not only gender "typical" children, but also children who are transgender and/or gender nonconforming.

We chose the domain of gender because it is often the focus of developmental IAT research (e.g., Cvencek et al., 2015; Meyer & Gelman, 2016; Olson et al., 2015). Further, the gender domain allowed us to easily assess the impact of the inclusion of distinct subgroups (in this case, male and female children) on our results. As correlations are generally stronger when there is more variability among observations (Aron & Aron, 2003), reliability and validity

estimates should covary with the extent to which accounting subgroup differences produces more/less variability in IAT scores. In the case of gender identity, for example, the IAT could be scored to indicate association of self with either male or female (e.g., Cvencek et al., 2015; henceforth "intergroup scoring"). Insofar that male and female children show very different rates of identification with either male or female (e.g., female children identify as female, male children identify as male), this approach is likely to produce a distribution of IAT scores with considerable variability, and thus, strengthen correlations. Alternatively, the IAT could be scored to indicate association of self with either one's own sex or the opposite sex (Olson et al., 2015; henceforth "ingroup scoring"). Insofar as male and female children show roughly equal rates and distributions of ingroup identification, this approach will likely produce a less variable distribution of IAT scores (compared to intergroup scoring), and thus, attenuate correlations.

As a concrete example, imagine that male and female participants twice complete an IAT indexing gender identity to assess the test-retest reliability of the measure. Further, imagine the IAT is scored using intergroup scoring procedures (e.g., positive values indicate identification with female and negative values indicate identification with male), and that most female participants implicitly identify as female and most male participants implicitly identify as male at both time points. The fact that female participants have higher scores on the IAT at both time points, and that the two distributions do not overlap (much or at all), will likely produce a significant test-retest coefficient. That is, even if there is not much variability of IAT scores within each group (i.e., all female participants show about the same identification with female and all male participants show about the same identification with male), IAT scores will be likely correlated across time because the large between-group differences for male and female participants.

Now imagine that the same IAT data is rescored using ingroup scoring (e.g., positive values indicate identification with one's sex and negative values indicate identification with the opposite sex). In the recoded data, both male and female participants will have positive values indicating identification with their own sex, and the distribution of IAT scores for male and female participants will likely have a large degree of overlap. Thus, insofar that IAT scores within the same sex are highly similar and that ingroup scoring largely eliminates the between-sex variability in IAT scores, this less variable distribution of IAT scores will likely result in an attenuated test-retest coefficient (relative to when using intergroup scoring procedures). While each scoring procedure can be useful for answering different research questions (see General Discussion), we report correlations both using intergroup coding and ingroup coding in our gender studies to explore how conclusions about the reliability and validity of the IAT could differ for researchers using each approach.

Study 3: 10-Minute Lag-Time

Method. Unless otherwise noted, all methods in Study 3 were identical to those used in Study 1.

Participants. One hundred and five participants completed the study at a research lab located at the University of Washington. Due to experimenter error, three participants who were 12 years or older participated in the study. Data from these three participants are excluded from all analyses. Partial data is reported from one six-year-old who failed to complete the IAT at Time 1 or Time 2 due to difficulty reading word stimuli in the IAT, two other participants that requested to discontinue the testing session after completing the IAT at Time 1, and one participant whose Time 2 IAT data was removed due to the same established exclusion criteria cited above.

Measures.

*Gender related behavior - clothing***.** As one measure of gender-related behavior, two experimenters rated participant's clothing (what they wore to the appointment without notice that it would be coded) on a scale from 1 = "very masculine" to 5 = "very feminine". An outfit that would be scored 1 was a football jersey with cargo shorts, black socks and charcoal gray shoes. An examine outfit that would be scored 5 would be a tutu paired with a sparkly pink shirt and ballet slippers. We assumed that at these ages, children likely play a non-trivial role in the selection of their outfits—an assumption supported by asking a handful of parents. Interrater agreement for clothing ratings was excellent (interclass correlation coefficient [ICC] = .93).[7]

*Gender related behavior – coloring page prize***.** As a second measure of gender-related behavior, participants were able to choose a coloring page as a prize at the conclusion of the testing session. The experimenter placed four coloring pages on the table in front of the participant in random order. Coloring pages were chosen via a pilot study (results of this pilot study are reported in the Supplemental Material) such that they differed in degree of perceived masculinity/femininity (in order: a truck, an astronaut, a horse and carriage, and a unicorn). Coloring book pages were coded with values from 1 (truck) to 4 (unicorn).

*Explicit gender identity***.** Explicit gender identity was assessed via two sequential questions in which children were shown a picture of a White male and a White female (of approximately equal age in middle childhood) on a laptop computer (Cvencek, Meltzoff, & Greenwald, 2011). While pointing at each picture, the experimenter explained the name and gender of each child (e.g., "On the left is Paul. He is a boy" and "On the right is Amanda. She is

[7] Due to experimenter error, 2 participants did not receive any clothing ratings. An additional 5 participants only had a rating from one experimenter.

a girl."). Children were first asked to indicate which child they were more like (e.g., "Are you more like Amanda or are you more like Paul?"). After making a selection, the experimenter then covered the unselected child with his/her hand and pointed to two circles (one small and one large) located below the selected child and asked the child to indicate the degree of their similarity with the selected target (e.g., "How much like [Paul/Amanda] are you? A little [while pointing at the small circle] or a lot [while pointing at the large circle]?"). At Time 2, participants rated different boys and girls (Emily and David, also White). The order in which the different photo sets were rated, as well as the side of the screen on which the male and female targets were presented, was counterbalanced across participants. The measure was scored from 1 (indicating the participant felt "a lot" like the boy target) to 4 (indicating the participant felt "a lot" like the girl target).

Implicit gender identity. Implicit gender identity was assessed with the same child-adapted IAT as described in Study 1. However, the target concepts of "White" and "Black" were replaced with "Male" and "Female" and the attribute concepts of "Good" and "Bad" were replaced with the categories of "Me" and "Not Me." The target concepts were represented by the photographs of both male and female children whereas attribute concepts were represented with "Me" words (e.g., I, me, mine, myself) and "Not Me" words (e.g., others, them, theirs, they). Children that respond more rapidly when "Female" photographs are paired with "Me" words relative to when they are paired with "Not Me" words are taken to implicitly identify as female. Using intergroup scoring procedures, the IAT was scored such that positive values indicate higher identification with female relative to male, and Spearman-Brown corrected split-half reliability at Time 1 and Time 2 was .73 and .82, respectively.

Ingroup Recoding. For analyses using ingroup scoring, measures were rescored so that higher scores for all participants meant they exhibited behavior typically associated with their own sex or identified with their own sex.

***Procedure*.** After completing both gender identity measures at Time 1 and Time 2, participants were given the opportunity to choose a coloring book page as a prize for study participation. Along with ratings of participants' clothing, the content of the selected coloring book page (e.g., masculine or feminine content) served as a measure of gender behavior.

Results.

***Intergroup scoring*.** Table 8 shows means, standard deviations, and zero-order correlations for both demographic and measured study variables. Results in Table 8 indicate (a) scores on the IAT and explicit measure were positively correlated, (b) participant age was not correlated with any identity or behavior measures, and (c) female participants (relative to male participants) tended to have stronger implicit and explicit identification with female and have more feminine clothing/coloring book prize selections (all rs $\geq .65$). Table 9 reveals large differences in measured variables between female and male participants, such that female participants had much higher scores than male participants on all measures (all ds ≥ 1.68).

Test-retest reliability. Scores on the IAT, $r(96) = .63$, $p < .001$, 95% CI [.49, .74], explicit measure, $r(99) = .56$, $p < .001$, 95% CI [.41, .68], were correlated across the 10-minute lag-time. Moreover, age was not correlated with the absolute difference in IAT scores, $r(96) = .16$, $p = .114$, 95% CI [−.04, .35], or explicit gender identity scores, $r(99) = .01$, $p = .897$, 95% CI [−.19, .21], across Time 1 and Time 2.

Table 8. Means, standard deviations, and zero-order correlations for demographic and measured variables in Study 3. Scores on the explicit identity measure and IAT were averaged across Time 1 and Time 2. The *p*-values were computed with *N*'s ranging from 94 to 102.

	Variable	Mean	SD	1	2	3	4	5
1	Sex (Female = 1)	.46	.50	-	-	-	-	-
2	Age	9.18	1.69	−.04	-	-	-	-
3	Explicit Measure (average)	2.43	.68	.82***	−.07	-	-	-
4	IAT (average)	−.04	.50	.65***	.03	.54***	-	-
5	Coloring Book Selected	2.51	1.17	.70***	−.08	.65***	.47***	-
6	Clothing	2.81	1.15	.90***	−.15	.73***	.57***	.67***

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$, † = $p < .10$

Table 9. Mean and standard deviations of measured variables for male and female participants in Study 3. The *t*-values and *p*-values are from two-tailed *t*-tests (*df* ranging from 72.08 to 95.10). Effect sizes are presented as Cohen's *d*s.

Variable	Females	Males	*t*	*p*	*d*
Explicit Measure (average)	3.03 (.41)	1.92 (.38)	14.13	< .001	2.82
IAT (average)	.32 (.41)	−.33 (.36)	8.32	< .001	1.68
Coloring Book Selected	3.40 (.82)	1.77 (.85)	9.44	< .001	1.94
Clothing	3.94 (.60)	1.87 (.38)	20.09	<.001	4.21

Predictive validity. As shown in Table 8, participants that implicitly, $r(98) = .57$, $p < .001$, 95% CI [.42, .69], or explicitly, $r(98) = .73$, $p < .001$, 95% CI [.62, .81], associated themselves with female tended to wear more feminine clothing. Similarly, implicit, $r(93) = .47$, $p < .001$, 95% CI [.30, .61], and explicit, $r(93) = .65$, $p < .001$, 95% CI [.52, .75], identification with female was associated with selecting a more feminine coloring page prize at the conclusion of the experiment. As shown in Table 10, multiple regression analyses revealed that scores on the IAT and explicit measure uniquely predicted clothing scores (Model 1a), but that only the explicit measure predicted coloring page selection (Model 1b). In a second set of multiple regression analyses, we further introduced participant sex and age into the model containing scores on the IAT and explicit measure. As shown in Table 10, after controlling for participant demographics

Table 10. The unstandardized coefficients, standard error of the unstandardized coefficients, and standardized coefficients (β) of hierarchical multiple regression models regressing clothing and coloring page ratings on (a) gender identity measures (Models 1a and 1b) and (b) gender identity measures and demographic variables (Models 2a and 2b). Scores on identity measures were averaged across Time 1 and Time 2.

	Clothing Ratings						**Coloring Book Page Ratings**					
	Model 1a			**Model 2a**			**Model 1b**			**Model 2b**		
Variable	**B**	**SE (B)**	**β**	**B**	**SE (B)**	**β**	**B**	**SE (B)**	**β**	**B**	**SE (B)**	**β**
Explicit Measure (average)	1.03***	.13	.61	−.06	.12	−.03	1.00***	.17	.58	.42†	.22	.24
IAT (average)	.60***	.18	.26	−.01	.13	−.01	.30	.23	.13	−.08	.24	−.04
Sex (Female=1)	-	-	-	2.14***	.18	1.86	-	-	-	1.23***	.33	1.05
Age	-	-	-	−.09**	.03	−.14	-	-	-	−.05	.05	−.07
R^2	.59			.84			.44			.51		

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$, † = $p < .10$

(especially sex), neither implicit nor explicit gender identity was predictive of either participant's clothing (Model 2a) or coloring book selection (Model 2b).

Ingroup coding. After recoding our measures to reflect ingroup identity/behavior, there were no significant differences in measured variables between male and female participants (see Table 11). In terms of test-retest reliability, scores on the IAT, $r(96) = .45$, $p < .001$, 95% CI [.28, .60], but not the explicit measure, $r(99) = .07$, $p = .468$, 95% CI [−.13, .26], were correlated across the 10-minute lag-time. With regards to predictive validity, scores on the IAT, $r(98) = -.02$, $p = .813$, 95% CI [−.22, .18], and explicit measure, $r(98) = -.01$, $p = .959$, 95% CI [−.21, .19], did not predict clothing ratings. Similarly, scores on the IAT, $r(93) = -.02$, $p = .873$, 95% CI [−.22, .18], and explicit measure, $r(93) = .19$, $p = .059$, 95% CI [−.01, .38], were not associated with participant's coloring book prize selection.

Table 11. Mean and standard deviations of measured variables for male and female participants in Study 3. Variables are coded to reflect ingroup identity/behavior. The *t*-values and *p*-values are from two-tailed *t*-tests (*df* ranging from 72.08 to 95.10).

Variable	Females	Males	*t*	*p*	*d*
Explicit Measure (average)	3.03 (.41)	3.08 (.38)	0.63	0.528	0.13
IAT (average)	.32 (.41)	.33 (.36)	0.18	0.861	0.04
Coloring Book Selected	3.40 (.82)	3.23 (.85)	0.96	0.342	0.20
Clothing	3.94 (.60)	4.13 (.38)	1.78	0.080	0.37

Study 4:1-Month Lag

Method. Unless otherwise noted, all methods in Study 4 were identical to those used in Study 3.

Participants. One hundred and seven participants between the ages of 6 and 11 years participated in the study at either a research lab located at the University of Washington ($N = 9$) or a quiet testing space in their school ($N = 98$; all schools located in Washington state).

Incomplete data occurred for a variety of reasons. Due to experimenter error, IAT data was lost from both Time 1 (*N*=10) and Time 2 (*N*=1), and some coloring page selections (*N* = 7) and clothing ratings from Time 2 (*N*=15) were not recorded. Additionally, four participants requested to skip part of the study procedure, and clothing ratings for 12 participants were not recorded at Time 1 or Time 2 because they wore a required school uniform during the experiment. Finally, due to absence (school participants) or inability to re-contact participants for a second lab visit, an additional eight participants provided no data at Time 2. Partial data from participants with incomplete data is reported whenever possible.

***Measures*.**

Gender-related behavior- clothing. Interrater agreement of clothing ratings was excellent at Time 1 (ICC = .89) and Time 2 (ICC = .93).[8]

Implicit gender identity. The Spearman-Brown corrected split-half reliability of the IAT at Time 1 and Time 2 was .80 and .72, respectively.

***Procedure*.** Participants completed the explicit and implicit gender identity measures twice approximately 3 to 4 weeks apart (*M* = 27 days, *SD* = 3 days). After completing the identity measures at Time 2, participants could choose a coloring page prize.

Results.

***Intergroup coding*.** Table 12 shows means, standard deviations, and zero-order correlations for both demographic and measured study variables. Results in Table 12 indicate (a) scores on the IAT and explicit measure were positively correlated, (b) older participants tended to wear less feminine clothing, and (c) female participants tended to have stronger implicit and

[8] Due to having only one experimenter present during the testing session, seven participants only received one rating at Time 1 and two others had only one rating at Time 2.

explicit identification with female and have more feminine clothing/coloring book prize selections (all *r*s ≥ .67). Table 13 shows large differences in measured variables between male and female participants (all *d*s ≥ 1.78).

Table 12. Means, standard deviations, and zero-order correlations for demographic and measured variables in Study 4. Scores on the explicit identity measure, IAT, and clothing ratings were averaged across Time 1 and Time 2 (the coloring book measure was administered at Time 2). The *p*-values were computed with *N*'s ranging from 79 to 107.

	Variable	Mean	SD	1	2	3	4	5
1	Sex (Female = 1)	.50	.50	-	-	-	-	-
2	Age	8.22	1.59	−.24*	-	-	-	-
3	Explicit Measure (average)	2.57	.82	.79***	−.10	-	-	-
4	IAT (average)	.03	.50	.67***	−.05	.52***	-	-
5	Coloring Book Selected	2.59	1.19	.74***	−.17†	.63***	.55***	-
6	Clothing	2.92	1.25	.93***	−.27**	.82***	.62***	.78***

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$, † = $p < .10$

Table 13. Mean and standard deviations of measured variables for male and female participants in Study 4. The *t*-values and *p*-values are from two-tailed *t*-tests (*df* ranging from 87.78 to 105).

Variable	Females	Males	*t*	*p*	*d*
Explicit Measure (average)	3.23 (.51)	1.94 (.51)	13.10	< .001	2.53
IAT (average)	.37 (.39)	−.30 (.36)	9.09	< .001	1.78
Coloring Book Selected	3.45 (.80)	1.68 (.80)	10.50	< .001	2.20
Clothing	4.10 (.50)	1.77 (.40)	24.93	<.001	5.13

Test-retest reliability. Scores on the IAT, $r(88) = .56$, $p < .001$, 95% CI [.40, .69], and explicit measure, $r(94) = .76$, $p < .001$, 95% CI [.66, .83], were correlated across the 1-month lag-time. Participant age was not correlated with the absolute difference in scores on the IAT, $r(88) = -.03$, $p = .782$, 95% CI [−.24, .18], or explicit measure, $r(94) = -.03$, $p = .745$, 95% CI [−.23, .17], across Time 1 and Time 2.

Table 14. The unstandardized coefficients, standard error of the unstandardized coefficients, and standardized coefficients (β) of hierarchical multiple regression models regressing clothing and coloring page ratings (collected at Time 2) on (a) gender identity measures (Models 1a and 1b) and (b) gender identity measures and demographic variables (Models 2a and 2b). Clothing ratings and scores on identity measures were averaged across Time 1 and Time 2.

	Clothing Ratings						**Coloring Book Page Ratings**					
	Model 1a			**Model 2a**			**Model 1b**			**Model 2b**		
Variable	**B**	**SE (B)**	**β**	**B**	**SE (B)**	**β**	**B**	**SE (B)**	**β**	**B**	**SE (B)**	**β**
Explicit Measure (average)	1.04***	.10	.68	.27**	.10	.17	.70***	.14	.48	.22	.17	.15
IAT (average)	.60***	.16	.24	.05	.12	.02	.71**	.22	.30	.22	.23	.09
Sex (Female=1)	-	-	-	1.87***	.19	1.49	-	-	-	1.33***	.31	1.11
Age	-	-	-	−.08**	.03	−.11	-	-	-	−.02	.06	−.03
R^2	.72			.89			.46			.57		

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Predictive validity. As shown in Table 12, participants that implicitly, $r(91)= .62$, $p < .001$, 95% CI [.48, .73], or explicitly, $r(93)= .82$, $p < .001$, 95% CI [.74, .88], associated themselves with female tended to wear more feminine clothing. Similarly, implicit, $r(89)= .55$, $p < .001$, 95% CI [.39, .68], and explicit, $r(89)= .63$, $p < .001$, 95% CI [.49, .74], identification with female was associated with selecting a more feminine coloring page prize at the conclusion of the experiment. As shown in Table 14, multiple regression models testing scores on the IAT and explicit measure as simultaneous predictors of gender-related behavior revealed that both identity measures predicted clothing ratings (see Model 1a) and the coloring page prize selected at the end of the experiment (see Model 1b). However, a second set of multiple regression models indicated that except for a significant association between explicit gender identity and clothing ratings (see Model 2a), implicit and explicit gender identity did not predict gender-related behavior after controlling for participant demographics (see Table 14).

Ingroup coding. After scoring measured variables to reflect ingroup identity/behavior, there were no significant differences in between male and female participants (see Table 15). We found non-zero test-retest reliability estimates for both identity measures, such that scores on the IAT, $r(88)= .32$, $p = .002$, 95% CI [.12, .49], and explicit measure, $r(94)= .51$, $p < .001$, 95% CI [.34, .64], were correlated across Time 1 and Time 2. Tests of predictive validity revealed that scores on the explicit measure, $r(93)= .23$, $p = .023$, 95% CI [.03, .41], but not the IAT, $r(91)= -.03$, $p = .787$, 95% CI [−.23, .17], predicted clothing ratings. Neither scores on the IAT, $r(89)= .11$, $p = .318$, 95% CI [−.10, .31], or explicit measure, $r(89)= .16$, $p = .137$, 95% CI [−.05, .35], predicted children's coloring book prize selection.

Table 15. Mean and standard deviations of measured variables for male and female participants in Study 4. Variables are coded to reflect ingroup identity/behavior. The *t*-values and *p*-values are from two-tailed *t*-tests (*df* ranging from 87.78 to 105).

Variable	Females	Males	*t*	*p*	*d*
Explicit Measure (average)	3.23 (.51)	3.06 (.51)	1.64	0.104	0.32
IAT (average)	.37 (.39)	.30 (.36)	1.00	0.319	0.20
Coloring Book Selected	3.45 (.80)	3.32 (.80)	0.77	0.446	0.16
Clothing	4.10 (.50)	4.23 (.40)	1.40	0.165	0.29

Study 5: 1-Year Lag Time in Gender Diverse Sample

Method. Data came from an ongoing longitudinal study of gender development in transgender and gender nonconforming children. We use the term "transgender*"* to refer to children who identify with the *gender* "opposite" to their *sex* (as determined by a doctor at birth) and identify with their gender identity in everyday life (e.g., they use the pronouns associated with their gender, not the ones associated with their sex). In addition, this study included several children whose gender could not be as easily categorized—for example, children who feel they are boys and girls or children whose gender changed over the course of the study. We use the term "gender nonconforming" to refer to these children. Collectively, we refer to the transgender and gender nonconforming children as "gender diverse children". To clearly differentiate between gender and sex, we use the terms "boys" and "girls" to describe participants' *gender* (using "nonbinary" for all gender nonconforming children) and "male" and "female" to classify participants' *sex* (all children were clearly identified as male or female by doctors at birth).

Participants. Participants included 41 gender diverse children: 31 participants who were transgender children at both time points and 10 participants who were gender nonconforming. The 10 gender nonconforming children included (a) 5 participants who used pronouns associated with their sex but whose gender expression aligned with the opposite gender, (b) 3 participants who used pronouns associated with their sex and whose gender expression aligned with the

opposite gender at Time 1, but were then living as transgender children at Time 2, and (c) 2 participants who identified as both boys and girls or as somewhere between boy and girl, and preferred neutral or a mix of gender pronouns at both time points. Our sample also included 20 siblings of children who are gender diverse and 36 unrelated gender-typical children who are used in the longitudinal studies as a control group, for a total of 97 participants.[9] The latter two groups were combined in analyses as they were all gender "typical".

Participants completed the study measures at conferences, support meetings, their family's home, or at a research lab at the University of Washington. Only children who first completed the Gender Identity IAT when they were between the ages of 6 and 11 years were eligible for inclusion in this study. Additionally, children had to have returned for a follow-up IAT before January 1, 2017 to be included (this study is ongoing). At Time 1, two participants skipped the IAT and one other failed to complete all IAT trials. Moreover, due to experimenter error, IAT data for five participants was lost from Time 2, and one participant did not have IAT data from either Time 1 or Time 2. Finally, 44 participants did not have any clothing ratings at Time 1 and 3 others did not have any clothing ratings at Time 2.

Measures. For the current study, only IAT scores and ratings of participant's clothing (both measures were identical to those used in Studies 3 and 4) were analyzed; the other measures utilized in Studies 3 and 4 were not asked of these children.

Gender-related behavior- clothing. Interrater agreement of clothing ratings was excellent at Time 1 (ICC = .96) and Time 2 (ICC = .89).[10]

[9] IAT data (from Time 1 only) from 50 participants is reported by Olson et al. (2015).
[10] Due to having only one experimenter present during the testing session, 25 participants only received one rating at Time 1 and 24 others had only one rating at Time 2.

Implicit gender identity. The Spearman-Brown corrected split-half reliability of the IAT at Time 1 and Time 2 was .73 and .79, respectively.

***Procedure*.** Participants completed the IAT in an initial testing session, and approximately 1-year later (M = 440 days, SD = 97 days), completed the IAT for a second time.

Results.

Intergroup scoring. Table 16 shows means, standard deviations, and zero-order correlations for both demographic and measured study variables. Results indicate that (a) female participants tended to have stronger implicit identification with female and (b) older respondents tended to wear less feminine clothing. Table 17 shows mean IAT scores and clothing ratings by participant sex (male vs. female) and gender status (diverse vs. typical). To remain consistent with the reporting of Studies 3 and 4, we compared measured variables between male and female participants. As a reminder, the sample of, for example, male participants in this study included male children who identified and behaved in ways stereotypically-associated with boys (gender typical children, as in past studies), but also male children who felt they were girls or who felt their gender was neither that of a boy or a girl (gender diverse children). Female participants (M = .18, SD = .49) had higher IAT scores than male participants (M = −.04, SD = .49), t(93.78)=2.12, p=.037, d=.43, 95% CI of the difference [.01, .41], but clothing ratings did not significantly differ between female (M = 3.08, SD = 1.23) and male participants (M = 2.78, SD = 1.17), t(92.94)=1.22, p=.225, d=.25, 95% CI of the difference [−.19, .79]. A full comparison of means is available in the Supplemental Material.

*Test-retest reliability***.** IAT scores were correlated across the 1-year lag-time, r(86)=.56, p<.001, 95% CI [.40, .69]. Participant age was not correlated with the absolute difference in IAT scores, r(86)= −.08, p = .433, 95% CI [−.28, .13], across Time 1 and Time 2.

*Predictive validity***.** As shown in Table 16, IAT scores were predictive of participants' clothing ratings, $r(92) = .64$, $p < .001$, 95% CI [.50, .75]. Results of a multiple regression analysis indicated that IAT scores remained a significant predictor of clothing ratings ($\beta = .60$, $p < .001$) even after controlling for participant age and sex. Moreover, we found that age ($\beta = -.17$, $p = .045$), but not sex ($\beta = .03$, $p = .848$), predicted participant's clothing ratings.

Table 16. Means, standard deviations, and zero-order correlations for demographic and measured variables in Study 5. Scores on the IAT and clothing ratings were averaged across Time 1 and Time 2. The *p*-values were computed with *N*'s ranging from 94 to 97.

	Variable	Mean	SD	1	2	3
1	Sex (Female = 1)	.49	.50	-	-	-
2	Age	8.74	1.64	.03	-	-
3	IAT (average)	.07	.50	.21*	−.19†	-
4	Clothing	2.93	1.20	.13	−.30**	.64***

Note: *** = $p < .001$, ** = $p < .01$, * = $p < .05$, † = $p < .10$

Table 17. Mean and standard deviations of IAT scores and clothing ratings in Study 5 stratified by participant sex (male vs. female) and gender (typical vs. diverse).

	Male		Female	
	Gender Typical	Gender Diverse	Gender Typical	Gender Diverse
Variable	(*N* = 24)	(*N* = 24)	(*N* = 32)	(*N* = 17)
IAT (averaged)	−.40 (.31)	.35 (.32)	.41 (.32)	−.27 (.44)
Clothing	1.79 (.37)	3.82 (.71)	3.83 (.79)	1.71 (.41)

Table 18. Mean and standard deviations of IAT scores and clothing ratings in Study 5 stratified by participant sex (male vs. female) and gender (typical vs. diverse). Variables are coded to reflect ingroup identity/behavior.

	Male		Female	
	Gender Typical	Gender Diverse	Gender Typical	Gender Diverse
Variable	(*N* = 24)	(*N* = 24)	(*N* = 32)	(*N* = 17)
IAT (averaged)	.40 (.31)	−.35 (.32)	.41 (.32)	−.27 (.44)
Clothing	4.21 (.37)	2.17 (.71)	3.83 (.79)	1.71 (.41)

Ingroup coding. After scoring measured variables to reflect ingroup identity/behavior (again, to be consistent with Studies 3-4 this was coded as association with one's own sex), there were no significant differences between male and female participants on IAT scores and clothing ratings (means presented in Table 18; *p*s ≥ .161; see Supplemental Material). We found non-zero test-retest reliability and predictive validity estimates, such that IAT scores were correlated across the 1-year lag-time, $r(86)=.54$, $p<.001$, 95% CI [.37, .67], and associated with ratings of participant's clothing, $r(92)= .61$, $p <.001$, 95% CI [.46, .72].

Discussion of Gender Identity Studies

Studies 3 through 5 investigated the test-retest reliability and predictive validity of the Gender Identity IAT. The extent to which the IAT was reliable and valid depended on the sample and scoring approach. Studies 3 and 4 sampled only gender typical children, which meant that intergroup IAT scoring procedures (in which scores indicate association of self with either male or female) produced a distribution of IAT scores with large-between sex differences. We found that the IAT in these studies predicted itself (*r*s of .63 and .56 across lag-times of 10-minutes and 1-month, respectively) and gender-related behavior (*r*s ranging from .47 to .62) quite well. However, after accounting for participant sex (by including sex as a covariate in a multiple regression) or using coding that eliminated between-sex differences (ingroup IAT scoring procedures), test-retest coefficients were reduced and predictive validity estimates were no longer significant. Study 5 sampled both gender typical and gender atypical children, which meant that there was substantial within-sex and between-sex variability in our measures. Thus, in contrast to Studies 3 and 4, estimates using both intergroup and ingroup IAT scoring procedures were similar in magnitude. Finally, it is worth noting that age did not predict the stability or magnitude of IAT scores in Studies 3 through 5.

What explains the overall pattern of results we observed in Studies 3 through 5? As previously mentioned, all else being equal, correlations tend to be stronger when there is more variability among observations than when there is less variability (Aron & Aron, 2003). Accordingly, changes in reliability and validity estimates corresponded to the extent to which accounting for between-sex variability reduced the total variability in IAT scores and behavioral measures. For example, when using intergroup scoring in Studies 3 and 4, between 40% to 64% of the variability in IAT scores and between 65% to 93% of the variability in behavioral measures was between male and female participants. As ingroup coding eliminated between-sex differences and significantly reduced the amount of variability in scores on both the IAT or gender-related behavior measures (see Supplemental Material), it is not surprising that this scoring approach drastically reduced the magnitude of reliability and validity estimates compared to our results under intergroup scoring procedures. By contrast, our more diverse sample in Study 5 contained large within-sex variability, which meant that between 5% and 8% of the variability in IAT scores and 1% of the variability in the behavior measure was between male and female participants. As such, ingroup coding by participant sex did not significantly reduce variability in scores on either the IAT or gender-related behavior measures (see Supplemental Material), and thus, recoding hardly changed the magnitude of test-retest or predictive validity estimates.[11]

[11] The Supplemental Material reports a re-analysis of the data from our gender identity studies using an alternative procedure that breaks down overall test-retest reliability and predictive validity correlations (i.e., those ignoring participant sex) into *within-sex* (i.e., correlation among only male or female participants) and *between-sex* correlations (i.e., correlation among means computed for male and female participants; George & James, 1993; Pedhazus, 1997). Results using this approach are consistent with the results reported in-text such that within-sex (i.e., devoid of between-sex variability) test-retest coefficients were greater than zero but predictive validity estimates were not.

One limitation of the analysis and results of Studies 3 through 5 is that we may have reached different conclusions had we used an alternative analytic approach. For example, at the request of an anonymous reviewer, we re-analyzed our data from Study 5 using ingroup scoring procedures, but then reverse coded all gender diverse participants. While it may seem as if this approach switched our coding from one based on sex (the above analyses) to one based on gender, in fact, interpretation was challenging and therefore we did not include it here. For example, some participants identified as both boys and girls, so this approach to switching their category assignment was arbitrary and not reflective of their gender identities. Nonetheless, because the largest group of gender diverse children in the sample were transgender (those who identify as the gender opposite their natal sex), this approach did reduce variability in our measures, and in turn, either reduced (test-retest reliability) or eliminated (predictive validity) effects obtained using intergroup coding (see Supplemental Material). As such, it is difficult to provide clear-cut scenarios in which the Gender Identity IAT is likely to be more/less reliable and valid (e.g., when accounting for participant sex). Instead, we believe that the results from these studies are better considered as a demonstration of a more general principle that reducing (e.g., by using coding schemes that eliminate between-group differences) or increasing variability (e.g., by obtaining a more diverse sample) has clear consequences on the extent to which IAT scores will correlate with themselves or criterion variables. In the General Discussion, we further discuss this issue and provide suggestions for researchers interested in using the IAT to predict behavioral outcomes in their own research.

Internal Meta-Analyses

To summarize the overall test-retest reliability and predictive validity evidence for the IAT in the present studies, we next conducted an internal meta-analysis of the test-retest

coefficients and IAT-behavior correlations for Studies 1 through 5. For transparency, we present our results in two ways. First, we present results using test-retest and predictive validity estimates using intergroup scoring procedures. For studies using the Gender Identity IAT, measures reflected identification and behavior consistent with either male or female. Second, we conducted parallel analyses in which the Gender Identity IAT and behavior measures were ingroup scored (i.e., scores reflect identification and behavior congruent with each participant's sex). Recall that because only White children participated in studies using the Race Attitude IAT, the intergroup IAT scoring procedure is already ingroup coded such that higher scores indicate stronger associations of participants' own race with positive valence. Thus, estimates from studies using the Race Attitude IAT were the same in both analyses. In all analyses, correlations were normalized via Fisher's *z*-transformation (as per the recommendation of Borenstein, Hedges, Higgins, & Rothstein, 2009), and then transformed back to the *r* metric for presentation.

Results

Intergroup scoring. A fixed-effects test[12] of the mean effect size (*M r*) for the correlation between IAT scores collected at Time 1 and Time 2 was significantly greater than zero, *M r* = .48, 95% CI [.40, .55], $Z = 11.15$, $p < .001$. We also found a positive mean effect size for the correlation between IAT and behavior measures *M r* = .46, 95% CI [.39, .53], $Z = 11.01$, $p < .001$.[13] Figures 2a (test-retest results) and 3a (predictive validity results) present Forest Plots of

[12] With a small number of samples, random effects tests are very conservative (Goh, Hall, & Rosenthal, 2016) and lack precision in estimating between-studies variance (Borenstein, Hedges, Higgins, & Rothstein, 2010). Thus, we report results using fixed effects tests. However, all results are presented using random effects tests in the Supplemental Material. Results were largely unchanged.

[13] Recall that Studies 1 and 2 assessed multiple measures of gender-related behavior. Thus, to prevent dependency among our predictive validity estimates (from using several outcome variables taken from the same sample of participants), for these studies we used correlations between average IAT scores and the average of standardized scores from the clothing and coloring page prize measures. All other predictive validity estimates are those reported in-text.

the effect sizes (along with 95 % CIs) for each study along with the mean effect size across all studies.

Next, we tested whether lag-time (measured in days) and domain (dummy coded with race attitudes =0; gender identity =1) moderated the strength of test-retest reliability, and whether domain moderated the strength of association between IAT and criterion variables. These results indicated that test-retest coefficients did not significantly differ as a function of the lag-time between Time 1 and Time 2, $B = .00$, $Z = 0.50$, $p = .612$. By contrast, both test-retest reliability estimates, $B = 0.36$, $Z = 3.86$, $p < .001$, and predictive validity estimates, $B = 0.45$, $Z = 4.88$, $p < .001$, were significantly stronger in the domain of gender identity (compared to race attitudes).

Ingroup scoring. Using estimates from Studies 3 through 5 in which measures index ingroup identity/behavior, we found a significant mean effect size for test-retest reliability, $M\ r = .38$, 95% CI [.30, .46], $Z = 8.53$, $p < .001$, and predictive validity, $M\ r = .20$, 95% CI [.11, .29], $Z = 4.50$, $p < .001$. Figures 2b (test-retest results) and 3b (predictive validity results) present Forest Plots of the effect sizes for each sample along with the mean effect size across all samples. Tests of moderator variables (domain and lag-time) indicated that test-retest coefficients did not significantly differ as a function of lag-time between Time 1 and Time 2, $B = .00$, $Z = 1.91$, $p = .057$. Moreover, test-retest reliability, $B = 0.16$, $Z = 1.52$, $p = .127$, and predictive validity estimates, $B = -0.02$, $Z = -0.06$, $p = .954$, did not significantly vary across domain.

Excluding Study 5 estimates. Our ingroup coded effect size estimates (especially for predictive validity; see Figure 3b) appeared to be somewhat driven by the estimates from Study 5. To test the sensitivity of our results to the exclusion of these estimates, we re-ran our ingroup coded meta-analyses excluding the results from Study 5. The mean effect size for test-retest

Figure 2. Shows coefficients using intergroup (panel A) and ingroup (panel B) coded *test-retest* coefficients (along with 95% CIs) for the studies using the Race Attitude IAT (Studies 1 & 2) and Gender Identity IATs (Studies 3 through 5). The mean effect size represents the weighted averages across all samples (N = 475). All estimates are scaled on the r metric.

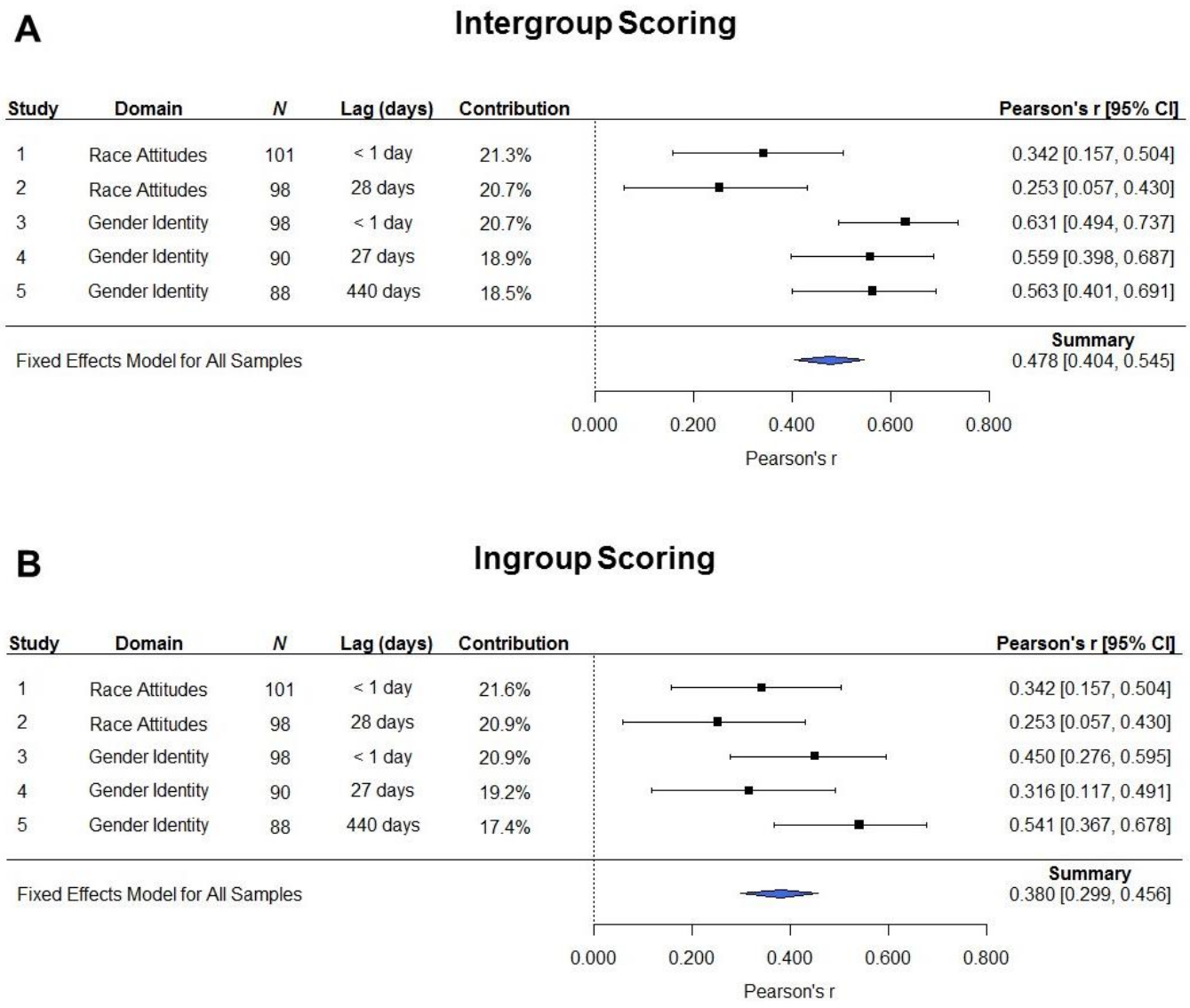
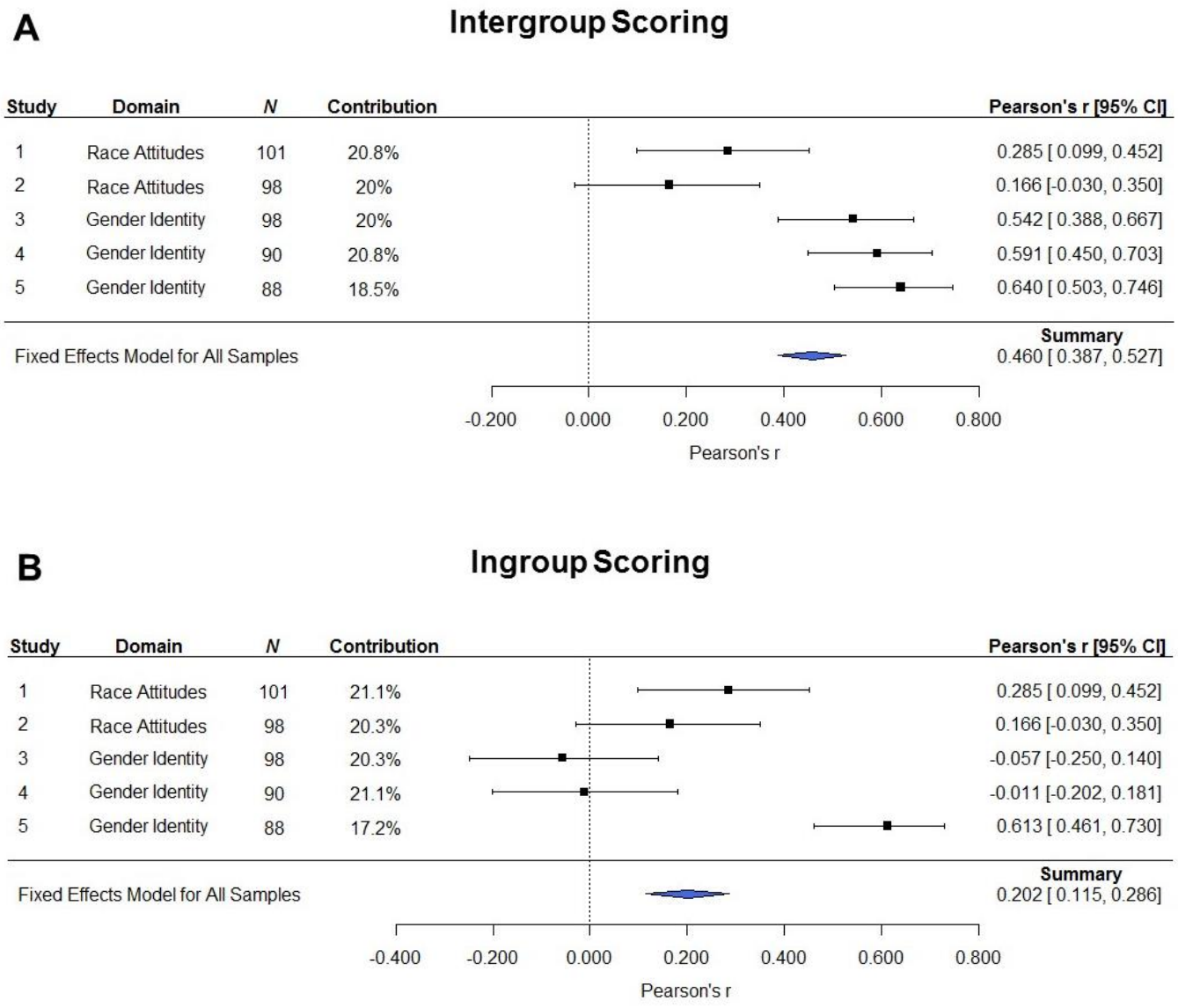


Figure 3. Shows coefficients using intergroup (panel A) and ingroup (panel B) coded *predictive validity* estimates (along with 95% CIs) for the studies using the Race Attitude IAT (Studies 1 & 2) and Gender Identity IAT (Studies 3 through 5). The mean effect size represents the weighted averages across all samples (N = 506). All estimates are scaled on the r metric.



reliability was similar in magnitude, *M r* = .34, 95% CI [.25, .43], *Z* = 6.92, *p* < .001, and our results indicated that neither domain, *B* = 0.10, *Z* = 0.98, *p* = .304, nor lag-time, *B* = −0.00, *Z* =−1.25, *p* = .211, moderated the magnitude of test-retest coefficients. We also found a significant mean effect size for the correlation between IAT and behavior measures, *M r* = .10, 95% CI [.001, .20], *Z* = 1.98, *p* =.048, that was significantly weaker in the domain of gender identity (compared to race attitudes), *B* = −.27, *Z* = −2.65, *p* = .008. More specifically, IAT scores were predictive of behavior in the domain of race attitudes, *M r* = .23, 95% CI [.09, .37], *Z* = 3.28, *p* =.001, but not gender identity, *M r* = −.03, 95% CI [−.17, .11], *Z* = −0.47, *p* =.635.

Discussion

An internal meta-analysis of the results of Studies 1 through 5 produced several interesting findings. First, regardless of how we analyzed the data, we found that mean test-retest and predictive validity effect sizes were significantly greater than zero. Moreover, in all cases, we found that lag-time did not moderate the magnitude of test-retest coefficients, which provides particularly compelling evidence that the IAT indexes trait-like associations in children (Fraley & Roberts, 2005). However, our analyses using intergroup coding (but not ingroup coding) indicated that test-retest and predictive validity estimates were stronger in the domain of gender identity (compared to race attitudes), whereas excluding the results from Study 5 from our ingroup coded analyses revealed that the IAT was a better predictor of behavior in the domain of race attitudes (compared to gender identity).

We previously discussed the role of measure variability in influencing the magnitude of test-retest and predictive validity estimates within Studies 3 through 5. However, this factor may also explain the meta-analytic domain differences in test-retest reliability that we observed. Indeed, when comparing the variance of IAT scores using intergroup coding (see Supplemental

Material), Gender Identity IAT scores were more variable than those on the Race Attitude IAT. By contrast, IAT variances were not statistically different when the Gender Identity IAT was coded using ingroup scoring (including Study 5 estimates). Taken together, we predict we would have found stronger test-retest coefficients in our studies using the Race Attitude IAT had we sampled both White and Black children, and thus, not observed domain differences using intergroup IAT scoring procedures. Regardless, the robustness of our results provides evidence that the IAT may be sensitive to trait-like individual differences - even when tested within otherwise homogenous groups.

The second interesting finding from our meta-analysis was that the mean effect size for predictive validity estimates – while always significantly above zero varied substantially between our analyses using intergroup scoring (average r = .46), ingroup scoring (r = .20), or ingroup scoring but excluding estimates from Study 5 (average r = .10). Moreover, the moderating role of domain functioned differently between our two sets of analyses: While our analyses performed with intergroup scoring indicated that Race Attitude IAT scores had *weaker* predictive validity compared to the Gender Identity IAT, our analyses that used ingroup scoring but excluded results from Study 5 revealed the opposite pattern of results such that Race Attitude IAT scores were *stronger* predictors of behavior than the Gender Identity IAT.

Our analysis of the variance in IAT scores provides some insight into why we may have found stronger predictive validity evidence in the domain of gender – at least when using intergroup IAT scoring procedures. However, it does not explain why we may have found stronger predictive validity in the domain of race when using ingroup scoring and excluding Study 5 estimates (there were not significant differences in IAT variances across domain). Of course, evidence for predictive validity has as much to do with the behavioral measures' own

variance, as the IAT's variance. For example, in Studies 1 and 2, around half of White children selected the Black child in the art contest (i.e., behavior consistent with favoring an outgroup member). By contrast, in Studies 3 and 4, only 2% of girls received a clothing score in the masculine range and no boys ever received a clothing score in the feminine range while between 7% and 21% of children selected the cross-sex coloring page. Thus, the IAT had greater opportunity to predict cross-group behavioral responses in the domain of race than in the domain of gender. A further discussion of this issue is provided in the General Discussion.

General Discussion

We conducted five studies with over 500 children to investigate the test-retest reliability and predictive validity of a child-friendly IAT. An internal meta-analysis of all studies revealed that regardless of how we analyzed the data, mean test-retest and predictive validity effect sizes were significantly greater than zero. Across studies, we also manipulated two factors that have varied in previous research investigating the test-retest reliability and predictive validity of the IAT in children: lag-time (10-minutes vs. 1-month vs. 1-year) and domain (gender identity vs. race attitudes). Our internal meta-analysis yielded a straight-forward conclusion about the role of lag-time on test-retest reliability; irrespective of how we analyzed the data, lag-time did not moderate the magnitude of test-retest coefficients. However, domain differences (or lack thereof) in test-retest reliability and predictive validity estimates were contingent on other factors, such as how we scored the IAT or whether or not we included estimates from a unique sample (i.e., one containing both gender typical and gender diverse children). Next, we provide a more thorough discussion of our findings with an eye toward how our findings might not only serve as a springboard for future work investigating the reliability and validity of the IAT in child samples, but also inform the practices of developmental researchers that utilize the IAT in their research.

Primary Findings

Developmental researchers have treated the IAT as an index of trait-like individual differences (e.g., Newheiser & Olson, 2012), yet the best test of whether a measure taps into individual differences (i.e., test-retest reliability; Greenwald & Nosek, 2001), has been understudied in child samples. Our internal meta-analyses yielded mean test-retest reliability effect sizes that were always above zero (mean *r*s of .48, .38, and .34) and that did not diminish as a function of lag-time (Fraley & Roberts, 2005). Moreover, after accounting for the fact that our gender studies contained more sample diversity than our race studies (i.e., by comparing ingroup scored estimates), we observed no domain differences in test-retest reliability. Taken together, it seems reasonable to conclude that the child-adapted IAT used here indexes general trait-like associations in elementary-aged children.

Moving forward, we believe our findings should inform both methodological and substantive developmental research. Indeed, as our findings do not provide definitive explanations for the wide range of test-retest coefficients observed in previous developmental research (summarized in Table 1), future methodological investigations might consider how factors not investigated here affect the test-retest reliability of the IAT. For example, the test-retest coefficients we observed were quite modest, which might simply reflect the fact that response latency measures like the IAT often have lower reliability than self-report measures (Gawronski & Payne, 2010; Nosek et al., 2007). Another possibility is that other IATs adapted for use with children may produce more reliable scores than the one used here, and testing this possibility should be a focus of future research. Our findings also have practical implications insofar as reliability sets an upper-bound on the extent to which the IAT will correlate with other variables of interest (e.g., Nunnally, 1978). As structural equation modeling accounts (SEM) for

measurement error in IAT scores (e.g., Cunningham, Preacher, & Banaji, 2001; Williams & Steele, 2016), developmental researchers might consider analyzing IAT data using SEM to circumvent issues of low reliability.

Our findings about the degree to which IAT scores predict behavior and whether the predictive validity of children's IAT scores differ by domain are less straightforward in their conclusions. Indeed, depending on the sample and scoring procedure used, our predictive validity meta-analysis yielded both strong and weak evidence that the IAT predicts behavior (mean *r*s of .46, .20, and .10). We believe two main factors influenced the pattern of results we observed, and thus, should be given consideration by researchers at the onset of a new research project using the IAT: between-group differences and reliability/variability of the outcome measure.

Between-group differences. Our predictive validity results demonstrate that the extent to which IAT scores correlate with behavior may depend on whether or not a researcher accounts for between-group differences that exist within their sample and the extent to which between-group differences explain variability in the IAT and/or behavioral measure. In our gender identity studies, for example, the IAT strongly correlated with measures of gender-related behavior (all *r*s $\geq$ .54) when measures were intergroup scored (i.e., on a scale from "female identity/behavior" to "male identity/behavior") and we did not statistically account for whether participants were male or female. However, except for the unique sample obtained in Study 5, a substantial proportion of the variability in IAT scores ($\geq$ 40%) and behavioral measures ($\geq$ 65%) was between male and female participants. Thus, removing between-sex variability by either using participant sex as a covariate in a multiple regression model or rescoring measures using

ingroup coding (i.e., a scale from "own-sex identity/behavior" to "opposite-sex identity/behavior) eliminated the association between scores on the IAT and behavioral measure.

Moving forward, it seems clear that researchers should consider the role that between-group differences may play in driving their results. However, our recommendations for when researchers should utilize intergroup or ingroup scoring procedures are less straightforward. Instead of advocating for any one approach, we suggest that this decision be guided by the research objectives at hand. Indeed, ingroup scoring may be useful for answering specific research questions; for example, testing whether male and female children have equally strong identification with their sex or gender might be conducted by recoding the scores of male or female participants. In other cases, however, ingroup scoring would not even be possible, such as when assessing Black-White racial attitudes among Asian participants. Similarly, using the IAT to measure constructs that are not group-based – such as self-esteem (Cvencek, Greenwald, & Meltzoff, 2016) or fear (Field & Lawson, 2003) – precludes use of ingroup scoring. In these cases, researchers might consider using multiple regression techniques to parcel out the effects of group membership or alternative approaches that compute within-group and between-group correlations (as done in the Supplemental Material; Padhazur, 1982). While conceptually akin to computing results within each subgroup (e.g., Galdi et al., 2014), we recommend both approaches over conducting statistical tests in smaller subgroups (likely losing statistical power). We further suggest that whenever possible, researchers make their data publicly available, which will allow other researchers to recode the data however they want for their own questions, making it easier to compare data across papers and lab groups.

Reliability/variability of outcome variables. One additional factor that is critical for researchers to consider in conducting research about the predictive validity of the IAT is the

variance in one's behavioral measures. The meta-analysis using ingroup coding and excluding estimates from Study 5 indicated that the Race Attitude IAT had stronger predictive validity than the Gender Identity IAT. While IAT scores were similarly variable across domains, about half of White children exhibited an outgroup favoring behavior on the art contest (i.e., voting for a Black child to win) whereas no boys and only 2% of girls received an outgroup consistent clothing rating (i.e., a girl with a clothing score in the masculine range or a boy with a clothing score in the feminine range). However, the fact that we found less willingness for children to perform behavior favorable to/consistent with the outgroup in the case of gender (rather than race) does not mean that there is (necessarily) less variability in behavior to predict in the domain of gender (compared to race). Instead, our specific operationalizations of gender behavior may have had less variability than our specific operationalization of race behavior. Thus, we recommend that researchers conduct pilot testing to ensure that a prospective behavioral measure provides sufficient variability for correlational analyses with IAT measures. Our previous discussion about scoring procedures and accounting for between-group variability may also inform our understanding of when behavior measures should be more/less variable.

Secondary Findings

Across studies, we found little evidence that participant age predicted the magnitude of IAT scores and we never observed an association between age and stability of IAT scores. However, these results do not imply that children of different ages performed equivalently while completing the IAT. Indeed, we found that older children in our studies tended to have faster and less variable trial latencies, as well as fewer errors. However, a virtue of the IAT scoring algorithm is that the latency difference *between* critical blocks for each participant is adjusted by his or her variability of responses *within* the critical blocks. Thus, while younger children tend to

have longer response latencies (relative to older children), these differences were accompanied by increased variability in response latencies. In total, the IAT scoring algorithm significantly reduces the influence of individual difference factors – such as age, task switching, working memory - that correlate with response-speed (Greenwald et al., 2003; Klauer, Schmitz, Teige-Mocigemba & Voss, 2010), which likely explains why participant age had negligible effects in our IAT analyses.

Our results also revealed that relative explicit measures (Studies 2 through 4) – but not an absolute explicit measure (Study 1) - correlated with IAT scores. Thus, for some research questions (e.g., testing whether the IAT predicts behavior after parceling out the effects of explicit judgments), developmental researchers using the IAT may be wise to use a relative explicit measure. Moreover, continuing our discussion of the conditions in which the IAT is likely to more/less predictive of behavior, using behavioral measures that correspond to the relative nature of the IAT may also increase IAT-behavior correlations (for a discussion of enhanced predictive validity for "complementary" behaviors/judgments; see Greenwald et al., 2009).

Unexpectedly, we found that female participants consistently had higher Race Attitude IAT scores than male participants. While some previous developmental work using the Race Attitude IAT has not tested differences between male and female participants (e.g., Newheiser & Olson, 2012), others have found no significant differences (Dunham, Baron, & Banaji, 2006). However, as evidence from millions of participants suggests that adult males hold stronger implicit preferences than females (Nosek et al., 2007), we consider it important to further replicate this effect before drawing any strong conclusions from our results.

Limitations and Future Directions

Despite the contributions of the present research, we acknowledge several noteworthy limitations. First, it is critical to consider how sample characteristics may have affected our results. Indeed, except for participants in Study 5, most participants were residents of the Seattle metropolitan area that were tested either in a laboratory at the University of Washington or in nearby schools. Thus, insofar that there may be regional variability in IAT scores (Rae, Newheiser, & Olson, 2015) or norms governing cross-gender or cross-race behavior, results may have looked different had we obtained samples from other parts of the United States. Relatedly, while we elected to examine race attitudes using a Black-White racial contrast, Blacks comprise a relatively small proportion of the Seattle population – 7.9% ("Race and Ethnicity Quick Statistics", n.d.). Thus, results for our race studies might have been different had we included a larger racial minority group as a contrast category (e.g., Asians/Pacific Islanders -14.2% of residents). Finally, families that have the time and motivation to bring their child into the laboratory to participate in a research study may differ in many respects from families that cannot (e.g., socioeconomic status; Fernald, 2010). While this concern is partially alleviated by the fact that we also collected data in school settings, future work might go further by attempting to recruit children from schools located in low socioeconomic neighborhoods.

It is also important to consider how our results may have differed had we used different measures or chosen to examine the performance of the IAT in different domains. For example, research in adults has suggested that the IAT might be especially useful in predicting "spontaneous" (e.g., eye contact) rather than "deliberate" (e.g., voting) behavior (Friese, Bluemke, & Wänke, 2007; Steffens & Schulze König, 2006), which means that we may have found stronger predictive validity evidence had we elected to use measures tapping into more

spontaneous responses. Similarly, our results may have been different had we selected different constructs or domains. Indeed, while our internal meta-analysis of predictive validity estimates using ingroup scoring and excluding Study 5 estimates indicated that the IAT used here may be more predictive in the domain of race attitudes (compared to gender identity), the IATs used here varied on both *construct* (attitude vs. identity) and *social category* (race vs. gender). Because either one of these factors may explain why predictive validity estimates differed between these IATs, future work might disentangle these factors by using IATs measuring different constructs but with the same social categories (e.g., gender attitude vs. gender identity), or alternatively, comparing IATs assessing the same construct but towards different social categories (e.g., gender attitude vs. race attitude).

Conclusion

The current research set out to better understand how several factors (lag-time and domain) affect the extent to which children's IAT scores are stable over time and predictive of behavior. Unexpectedly, our results also enabled us to explore how additional factors (e.g., whether or not one accounts for between-group differences in a sample) may influence the magnitude of test-retest reliability and predictive validity estimates. With some caveats, the major take-home from this work is that the IAT used here indexes individual differences that are consistent across time (as much as a year) and predictive of behavior with elementary-aged children. We look forward to future work that will continue to understand when and where the IAT will be most or least useful with child populations.

References

References marked with an asterisk indicate reports that used the IAT (or close derivatives) with child samples.

Aron, A., & Aron, E. N. (2003). *Statistics for Psychology*. Prentice Hall/Pearson Education.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods, 46*, 668-688. doi:10.3758/s13428-013-0410-6

* Babcock, R. L., MaloneBeach, E. E., Hanighofer, J., & Woodworth-Hou, B. (2016). Development of a children's IAT to measure bias against the elderly. *Journal of Intergenerational Relationships, 14*, 167–178.

* Babcock, R. L., MaloneBeach, E. E., & Woodworth-Hou, B. (2016). Intergenerational intervention to mitigate children's bias against the elderly. *Journal of Intergenerational Relationships, 14*, 274-287.

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

* Bissell, K., & Hays, H. (2010). Understanding anti-fat bias in children: The role of media and appearance anxiety in third to sixth graders' implicit and explicit attitudes toward obesity. *Mass Communication and Society, 14*, 113-140. doi:10.1080/15205430903464592

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97-111.

Bosson, J. K., Swann Jr, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited?. *Journal of Personality and Social Psychology*, *79*, 631-643. doi:10.1037/0022-3514.79.4.631

* Bruni, C. (2007). *Using the Implicit Association Test to explore environmental preferences in children* (Unpublished doctoral dissertation). California State University -San Marcos.

* Bruni, C. M., & Schultz, P. W. (2010). Implicit beliefs about self and nature: Evidence from an IAT game. *Journal of Environmental Psychology*, *30*, 95-102. doi:10.1016/j.jenvp.2009.10.004

Buhrmester, M. D., Blanton, H., & Swann Jr, W. B. (2011). Implicit self-esteem: nature, measurement, and a new way forward. *Journal of Personality and Social Psychology*, *100*, 365-385. doi:10.1037/a0021341

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: Sage

* Cheetham, T. J., Turner-Cobb, J. M., & Gamble, T. (2015). Children's implicit understanding of the stress—illness link: Testing development of health cognitions. *British Journal of Health Psychology*. doi:10.1111/bjhp.12181

* Ćirović, I., Jošić, S., & Žeželj, I. (2011). Application and validation of an Implicit Association Test in the measurement of implicit prejudice among children. *Suvremena Psihologija*, *14*, 171-181.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

* Corenblum, B. (2016). Ethnic Identity Development among First Nation Children: An Individual Growth Approach. *The Canadian Journal of Native Studies*, *36*, 29-55.

* Corenblum, B., & Armstrong, H. D. (2012). Racial-ethnic identity development in children in a racial-ethnic minority group. *Canadian Journal of Behavioral Science*, *44*, 124-137. doi:10.1037/a0027154

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

* Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2016). Implicit measures for preschool children confirm self-esteem's role in maintaining a balanced identity. *Journal of Experimental Social Psychology*, *62*, 50-57. doi:10.1016/j.jesp.2015.09.015

* Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2011). Measuring implicit attitudes of 4-year-olds: The preschool implicit association test. *Journal of Experimental Child Psychology*, *109*, 187-200. doi:10.1016/j.jecp.2010.11.002

* Cvencek, D., Kapur, M., & Meltzoff, A. N. (2015). Math achievement, stereotypes, and math self-concepts among elementary-school students in Singapore. *Learning and Instruction*, *39*, 1-10. doi:10.1016/j.learninstruc.2015.04.002

* Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*, 766-779. doi:10.1111/j.1467-8624.2010.01529.x

* Cvencek, D., Meltzoff, A. N., & Kapur, M. (2014). Cognitive consistency and math–gender stereotypes in Singaporean children. *Journal of Experimental Child Psychology*, *117*, 73-91. doi:10.1016/j.jecp.2013.07.018

* Cvencek, D., Nasir, N. I. S., O'Connor, K., Wischnia, S., & Meltzoff, A. N. (2015). The development of math–race stereotypes: “They say Chinese people are the best at math”. *Journal of Research on Adolescence*, *25*, 630-637. doi:10.1111/jora.12151

* Degner, J., & Wentura, D. (2010). Automatic prejudice in childhood and early adolescence. *Journal of Personality and Social Psychology, 98*, 356-374. doi:10.1037/a0017993

* Diesendruck, G., & Menahem, R. (2015). Essentialism promotes children's inter-ethnic bias. *Frontiers in Psychology*, *6*, 1180. http://doi.org/10.3389/fpsyg.2015.01180

* Dunham, Y., Baron, A. S., & Banaji, M. R. (2006). From American city to Japanese village: A cross-cultural investigation of implicit race attitudes. *Child Development*, *77*, 1268-1281. doi:10.1111/j.1467-8624.2006.00933

* Dunham, Y., Baron, A. S., & Banaji, M. R. (2007). Children and social groups: A developmental analysis of implicit consistency in Hispanic Americans. *Self and Identity*, *6*, 238-255. doi: 10.1080/15298860601115344

* Dunham, Y., Baron, A. S., & Banaji, M. R. (2015). The development of implicit gender attitudes. *Developmental Science*, *19*, 781-789. doi:10.1111/desc.12321

* Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in children. *Child Development*, *82*, 793-811. doi:10.1111/j.1467-8624.2011.01577.x

Dunham, Y., & Emory, J. (2014). Of affect and ambiguity: The emergence of preference for arbitrary ingroups. *Journal of Social Issues*, *70*, 81-98. doi:10.1111/josi.12048

* Dunham, Y., Newheiser, A. K., Hoosain, L., Merrill, A., & Olson, K. R. (2014). From a different vantage: Intergroup attitudes among children from low-and intermediate-status racial groups. *Social Cognition*, *32*, 1-21. doi:10.1521/soco.2014.32.1.1

* Dunham, Y., Srinivasan, M., Dotsch, R., & Barner, D. (2014). Religion insulates ingroup evaluations: the development of intergroup attitudes in India. *Developmental Science*, *17*, 311-319. doi:10.1111/desc.12105

* Emeh, C. C., Mikami, A. Y., & Teachman, B. A. (2015). Explicit and implicit positive illusory bias in children with ADHD. *Journal of Attention Disorders, 18,* 456–465. doi:10.1177/1087054715612261

Fernald, A. (2010). Getting beyond the "convenience sample" in research on early cognitive development. *Behavioral and Brain Sciences, 33,* 91-92.

* Field, A. P., & Lawson, J. (2003). Fear information and the development of fears during childhood: Effects on implicit fear responses and behavioural avoidance. *Behaviour Research and Therapy, 41,* 1277-1293. doi:10.1016/S0005-7967(03)00034-2

* Field, A. P., Lawson, J., & Banerjee, R. (2008). The verbal threat information pathway to fear in children: the longitudinal effects on fear cognitions and the immediate effects on avoidance behavior. *Journal of Abnormal Psychology, 117,* 214-224. doi:10.1037/0021-843X.117.1.214

* Fioravanti-Bastos, A. C. M., Filgueiras, A., & Landeira-Fernandez, J. (2014). Using a visualized reaction-time task to assess implicit cognition in Brazilian and Japanese-descendant children. *International Journal of Psychological Studies, 6,* 80-87. doi:10.5539/ijps.v6n3p80

Friese, M., Bluemke, M., & Wänke, M. (2007). Predicting voting behavior with implicit attitude measures: The 2002 German parliamentary election. *Experimental Psychology, 54,* 247-255.

* Galdi, S., Cadinu, M., & Tomasetto, C. (2014). The roots of stereotype threat: when automatic associations disrupt girls' math performance. *Child Development, 85,* 250-263. doi:10.1111/cdev.12128

* George, M. (2015). *A cross-cultural investigation of minority and non-White majority children's implicit attitudes toward racial outgroups* (Unpublished doctoral dissertation). York University Toronto.

Geoge, J., & James, L. (1993). Personality, affect, and behavior in groups revisited: comment on aggregation, levels of analysis, and a recent application of within and between analysis. *Journal of Applied Psychology, 78*, 798-804.

* Glover, V. A. (2015). *Assessing the effect of race saliency in measures of children's implicit bias.* (Unpublished doctoral dissertation). University of Nevada, Las Vegas.

* Gonzalez, A. M. (2015). *Malleability of implicit intergroup bias across development.* (Unpublished doctoral dissertation). University of British Columbia.

* Gonzalez, A. M., Dunlop, W. L., & Baron, A. S. (2016). Malleability of implicit associations across development. *Developmental Science*.

* Gonzalez, A. M., Steele, J. R., & Baron, A. S. (2016). Reducing children's implicit racial bias through exposure to positive out-group exemplars. *Child Development, 88,* 123 – 130. doi:10.1111/cdev.12582

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464-1480. doi:10.1037/0022-3514.74.6.1464

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie*, *48*, 85-93. doi:10.1026/0949-3946.48.2.85

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. doi:10.1037/h0087889

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17-41. doi:10.1037/a0015575

* Grumm, M., Hein, S., & Fingerle, M. (2011). Predicting aggressive behavior in children with the help of measures of implicit and explicit aggression. *International Journal of Behavioral Development, 35*, 352-357. doi:10.1177/0165025411405955

* Hauser, J. C. (2010). *Understanding explicit and implicit anti-fat attitudes and their relations to other prejudiced attitudes, controllability beliefs and social desirability in children, adolescents, and young adults* (Unpublished doctoral dissertation). Bowling Green State University.

* Heiphetz, L. A. (2013). *The influence of beliefs on children's and adults' cognition and social preferences*. (Unpublished doctoral dissertation). Harvard University.

* Heiphetz, L., Spelke, E. S., & Banaji, M. R. (2013). Patterns of implicit and explicit attitudes in children and adults: Tests in the domain of religion. *Journal of Experimental Psychology: General*, *142*, 864-879. doi:10.1037/a0029714

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*, 1369-1385. doi:10.1177/0146167205275613

* Huijding, J., Field, A. P., De Houwer, J., Vandenbosch, K., Rinck, M., & Van Oeveren, M. (2009). A behavioral route to dysfunctional representations: The effects of training approach or avoidance tendencies towards novel animals in children. *Behaviour Research and Therapy*, *47*, 471-477. doi:10.1016/j.brat.2009.02.011

Hummert, M. L., Garstka, T. A., O'Brien, L. T., Greenwald, A. G., & Mellott, D. S. (2002). Using the Implicit Association Test to measure age differences in implicit social cognitions. *Psychology and Aging*, *17*, 482- 495. doi:10.1037/0882-7974.17.3.482

* Hutchison, S. M. (2015). *Explicit and implicit measures of weight-related attitudes in young children: Associations with perspective taking and executive function* (Unpublished doctoral dissertation). University of Victoria.

* Ibáñez, A., Gleichgerrcht, E., Hurtado, E., González, R., Haye, A., & Manes, F. F. (2010). Early neural markers of implicit attitudes: N170 modulated by intergroup and evaluative contexts in IAT. *Frontiers in Human Neuroscience*, *4*, 188. doi:10.3389/fnhum.2010.00188

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm*, *65*, 2276-84. doi:10.2146/ajhp070364

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. *The Quarterly Journal of Experimental Psychology*, *63*, 595-619.

* Kurman, J., Rothschild-Yakar, L., Angel, R., & Katz, M. (2015). How good am I? Implicit and explicit self-esteem as a function of perceived parenting styles among children with ADHD. *Journal of Attention Disorders*. Advance online publication. doi:1087054715569599.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4:863. doi: 10.3389/fpsyg.2013.00863.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the implicit association test: IV. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 59–102). New York, NY: Guilford Press.

* Leeuwis, F. H., Koot, H. M., Creemers, D. H., & van Lier, P. A. (2015). Implicit and explicit self-esteem discrepancies, victimization and the development of late childhood internalizing problems. *Journal of Abnormal Child Psychology, 43,* 909-919. doi:10.1007/s10802-014-9959-5

* Lemmer, G., Gollwitzer, M., & Banse, R. (2014). On the psychometric properties of the aggressiveness-IAT for children and adolescents. *Aggressive Behavior, 41,* 84-95 doi:10.1002/AB.21575

* Lochbuehler, K., Sargent, J. D., Scholte, R. H., Pieters, S., & Engels, R. C. (2012). Influence of smoking cues in movies on children's beliefs about smoking. *Pediatrics, 130,* 221-227. doi:10.1542/peds.2011-1792

* Mandalaywala, T.M. & Rhodes, M. (in press). Racial essentialism is associated with prejudice towards Blacks in 5- and 6-year old White children. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37,* 435-442. doi:10.1006/jesp.2000.1470

* McQuade, J. D., Mendoza, S. A., Larsen, K. L., & Breaux, R. P. (2016). The nature of social positive illusory bias: Reflection of social impairment, self-protective motivation, or poor

executive functioning?. *Journal of Abnormal Child Psychology*, 1-12. doi:10.1007/s10802-016-0172-6

* Meyer, M., & Gelman, S. A. (2016). Gender Essentialism in Children and Parents: Implications for the Development of Gender Stereotyping and Gender-Typed Preferences. *Sex Roles*, 1-13. doi:10.1007/s11199-016-0646-6

* Newheiser, A. K., Dunham, Y., Merrill, A., Hoosain, L., & Olson, K. R. (2014). Preference for high status predicts implicit outgroup bias among children from low-status groups. *Developmental Psychology*, *50*, 1081-1090. doi:10.1037/a0035054

* Newheiser, A. K., & Olson, K. R. (2012). White and Black American children's implicit intergroup bias. *Journal of Experimental Social Psychology*, *48*, 264-270. doi:10.1016/j.jesp.2011.08.011

* Noel, J. G., & Thomson, N. R. (2012). Children's alcohol cognitions prior to drinking onset: Discrepant patterns from implicit and explicit measures. *Psychology of Addictive Behaviors*, *26*, 451-459. doi:10.1037/a0025531

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2006). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). New York: Psychology Press.

Nunnally, J. (1978). *Psychometric theory* (2nd ed). New York: McGraw Hill.

* O'Connor, R. M., Lopez-Vergara, H. I., & Colder, C. R. (2012). Implicit cognition and substance use: the role of controlled and automatic processes in children. *Journal of Studies on Alcohol and Drugs*, *73*, 134-143. doi:10.15288/jsad.2012.73.134

* O'Driscoll, C., Heary, C., Hennessy, E., & McKeague, L. (2012). Explicit and implicit stigma towards peers with mental health problems in childhood and adolescence. *Journal of Child Psychology and Psychiatry, 53,* 1054-1062. doi:10.1111/j.1469-7610.2012.02580.x

Olson, K. R., & Dunham, Y. (2010). The development of implicit social cognition. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 241-254). New York, NY: Guilford Press

* Olson, K. R., Key, A. C., & Eaton, N. R. (2015). Gender cognition in transgender children. *Psychological Science, 26,* 467-474. doi:10.1177/0956797614568156

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105,* 171-192. doi:10.1037/a0032734

Ozer, D. J. (1999). Four principles for personality assessment. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 671-686). New York: Guilford.

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 1–15). New York, NY: Guilford Press

* Passolunghi, M. C., Ferreira, T. I. R., & Tomasetto, C. (2014). Math–gender stereotypes and math-related beliefs in childhood and early adolescence. *Learning and Individual Differences, 34,* 70-76. doi:10.1016/j.lindif.2014.05.005

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction*. Chicago, IL: Holt, Rinehart & Winston

* Pieters, S., van der Vorst, H., Engels, R. C., & Wiers, R. W. (2010). Implicit and explicit cognitions related to alcohol use in children. *Addictive Behaviors*, *35*, 471-478. doi:10.1016/j.addbeh.2009.12.022

* Qian, M. K., Heyman, G. D., Quinn, P. C., Messi, F. A., Fu, G., & Lee, K. (2016). Implicit racial biases in preschool children and adults from Asia and Africa. *Child Development*, *87*, 285-296. doi:10.1111/cdev.12442

Raabe, T., & Beelmann, A. (2011). Development of ethnic, racial, and national prejudice in childhood and adolescence: A multinational meta-analysis of age differences. *Child Development*, *82*, 1715-1737.

"Race and Ethnicity Quick Statistics" (n.d.). Retrieved from http://www.seattle.gov/dpd/cityplanning/populationdemographics/aboutseattle/raceethnicity/default.htm

Rae, J. R., Newheiser, A. K., & Olson, K. R. (2015). Exposure to racial out-groups and implicit race bias in the United States. *Social Psychological and Personality Science, 6*, 535-543. doi: 10.1177/1948550614567357

Richetin, J., & Richardson, D. S. (2008). Automatic processes and individual differences in aggressive behavior. *Aggression and Violent Behavior*, *13*, 423-430. doi:10.1016/j.avb.2008.06.005

* Roddy, S., & Stewart, I. (2012). Children's implicit and explicit weight-related attitudes. *The Irish Journal of Psychology*, *33*, 166-180. doi:10.1080/03033910.2012.677996

* Rosen, P. J., Milich, R., & Harris, M. J. (2007). Victims of their own cognitions: Implicit social cognitions, emotional distress, and peer victimization. *Journal of Applied Developmental Psychology*, *28*, 211-226. doi:10.1016/j.appdev.2007.02.001

* Rutland, A., Cameron, L., Milne, A., & McGeorge, P. (2005). Social norms and self-presentation: Children's implicit and explicit intergroup attitudes. *Child Development*, *76*, 451-466. doi:10.1111/j.1467-8624.2005.00856.x

Schmukle, S. C., & Egloff, B. (2004). Does the Implicit Association Test for assessing anxiety measure trait and state variance?. *European Journal of Personality*, *18*, 483-494.

* Sinclair, S., Dunn, E., & Lowery, B. (2005). The relationship between parental racial attitudes and children's implicit prejudice. *Journal of Experimental Social Psychology*, *41*, 283-289. doi:10.1016/j.jesp.2004.06.003

* Skowronski, J. J., & Lawrence, M. A. (2001). A comparative study of the implicit and explicit gender attitudes of children and college students. *Psychology of Women Quarterly*, *25*, 155-165. doi:10.1111/1471-6402.00017

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

* Solbes, I., & Enesco, I. (2010). Explicit and implicit anti-fat attitudes in children and their relationships with their body images. *Obesity Facts*, *3*, 23-32. doi:10.1159/000280417

Steffens, M. C., & Schulze König, S. (2006). Predicting spontaneous big five behavior with implicit association tests. *European Journal of Psychological Assessment*, *22*, 13-20.

* Thomas, S. R., Burton Smith, R., & Ball, P. J. (2007). Implicit attitudes in very young children: An adaptation of the IAT. *Current Research in Social Psychology*, *13*, 75-85.

* Tomasetto, C., Galdi, S., & Cadinu, M. (2012). When the implicit precedes the explicit: Gender stereotypes about math in 6-year-old girls and boys. *Psicologia Sociale*, *7*, 169-186. doi:10.1482/3769

* Turner, R. N., Hewstone, M., & Voci, A. (2007). Reducing explicit and implicit outgroup prejudice via direct and extended contact: The mediating role of self-disclosure and intergroup anxiety. *Journal of Personality and Social Psychology, 93*, 369-388. doi:10.1037/0022-3514.93.3.369

* van Goethem, A. A., Scholte, R. H., & Wiers, R. W. (2010). Explicit-and implicit bullying attitudes in relation to bullying behavior. *Journal of Abnormal Child Psychology, 38*, 829-842. doi:10.1007/s10802-010-9405-2

*Vander Heyden, K. M., van Atteveldt, N. M., Huizinga, M., & Jolles, J. (2016). Implicit and explicit gender beliefs in spatial ability: stronger stereotyping in boys than girls. *Frontiers in Psychology, 7*. doi:10.3389/fpsyg.2016.01114

* Vezzali, L., Giovannini, D., & Capozza, D. (2012). Social antecedents of children's implicit prejudice: Direct contact, extended contact, explicit and implicit teachers' prejudice. *European Journal of Developmental Psychology, 9*, 569-581. doi:10.1080/17405629.2011.631298

* Williams, A., & Steele, J. R. (2016). The reliability of child-friendly race-attitude Implicit Association Tests. *Frontiers in Psychology*, 7.

* Xi, J., Zuo, Z., & Sang, B. (2011). Perceived social competence of resilient children. *Acta Psychologica Sinica, 43*, 1026-1037. doi:10.3724/SP.J.1041.2011.01026

* Žeželj, I., Jakšić, I., & Jošić, S. (2015). How contact shapes implicit and explicit preferences: attitudes toward Roma children in inclusive and non-inclusive environment. *Journal of Applied Social Psychology, 45*, 263-273. doi:10.1111/jasp.12293