Behavior Research Methods (2019) 51:2194–2208
https://doi.org/10.3758/s13428-019-01293-3

Survey-software implicit association tests: A methodological and empirical analysis

Thomas P. Carpenter[1] · Ruth Pogacar[2] · Chris Pullig[3] · Michal Kouril[4,5] · Stephen Aguilar[6] · Jordan LaBouff[7] · Naomi Isenberg[1] · Alek Chakroff[8]

Published online: 12 September 2019

Abstract

The implicit association test (IAT) is widely used in psychology. Unfortunately, the IAT cannot be run within online surveys, requiring researchers who conduct online surveys to rely on third-party tools. We introduce a novel method for constructing IATs using online survey software (Qualtrics); we then empirically assess its validity. Study 1 (student n = 239) revealed good psychometric properties, expected IAT effects, and expected correlations with explicit measures for survey-software IATs. Study 2 (MTurk n = 818) showed predicted IAT effects across four survey-software IATs (ds = 0.82 [Black–White IAT] to 2.13 [insect–flower IAT]). Study 3 (MTurk n = 270) compared survey-software IATs and IATs run via Inquisit, yielding nearly identical results and intercorrelations that would be expected for identical IATs. Survey-software IATs appear to be reliable and valid, offer numerous advantages, and make IATs accessible for researchers who use survey software to conduct online research. We present all the materials, links to tutorials, and an open-source tool that rapidly automates survey-software IAT construction and analysis.

Keywords Implicit association test · Online research · Implicit measures

The past two decades have seen widespread use of the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998). The IAT assesses mental associations via a stimulus-sorting task (e.g., valenced associations with race, gender, etc.) and predicts cognition, affect, and behavior (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). The IAT has had a major impact on psychology, spawning a vast literature and making contributions to areas such as prejudice (Fazio & Olson, 2003), preference (Friese, Wänke, & Plessner, 2006), politics (Nosek, Banaji, & Greenwald, 2002), morality (Marquardt & Hoeger, 2009), suicidality (Nock et al., 2010), and personality (Fatfouta & Schröder-Abé, 2018). Trends indicate that usage is likely to increase in the coming years (Greenwald & Banaji, 2017).

The past decade has also seen the proliferation of online data collection. Online studies allow easy access to large, high-powered samples (Buhrmester, Talaifar, & Gosling, 2018; Paolacci & Chandler, 2014). Unfortunately, online platforms favor self-report methods, offering limited options for researchers who wish to include IATs. We suggest, instead, that researchers create IATs directly within online surveys by modifying the survey elements to create a "survey-software" IAT. Survey-software IATs would avoid substantial costs for those with access to survey software, provide more control over design, simplify data management, and reduce attrition (since participants are not required to install the software). In the present work, we develop this method in detail, discuss necessary considerations, and provide procedures, materials, and an open-source tool for rapidly implementing our procedure. We then empirically assess the validity of survey-software IATs across three online studies.

✉ Thomas P. Carpenter
tcarpenter@spu.edu

[1] Seattle Pacific University, Seattle, WA, USA

[2] University of Calgary, Calgary, Alberta, Canada

[3] Baylor University, Waco, TX, USA

[4] Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[5] University of Cincinnati College of Medicine, Cincinnati, OH, USA

[6] University of Southern California, Los Angeles, CA, USA

[7] University of Maine, Orono, ME, USA

[8] Harvard University, Cambridge, MA, USA

Running IATs online

Online data collection is increasingly common in psychology. Modern survey software (e.g., Qualtrics [www.qualtrics.com]) is used to implement many of these studies and can administer a range of procedures, from questionnaires to randomized experiments (all referred to hereafter as "online surveys"). For example, Cohen, Wolf, Panter, and Insko (2011) used survey software to assign participants to incentive conditions and measure effects on deceptive responding. Instead of requiring in-person participation, these methods allow participants to be recruited from anywhere, including university and crowdsourcing pools. Online methods can also make research more efficient: data files are generated automatically, data management is streamlined, and direct replication is often as simple as re-posting a link to a recruitment platform. Perhaps the biggest advantage, however, is power. Online participant pools are large (Buhrmester et al., 2018), and given modest financial resources, high-powered samples are accessible in short amounts of time (Paolacci & Chandler, 2014).

Unfortunately, researchers who use online surveys to collect data cannot use them to run IATs. The IAT requires sensitive reaction-time measurement, a feature not included in online-survey tools, leaving researchers with few options. Researchers can hire Project Implicit (www.implicit.harvard.edu) to build and host IATs on their proprietary platform. However, this is costly and limits data collection options. Alternatively, researchers can purchase access to Millisecond, a tool that allows one to run reaction-time tasks online. This option is also costly, and participants must be able and willing to install software onto their computers. When combined with an online survey, this also requires routing participants back and forth between IAT and survey platforms via hyperlink. All of this results in attrition; for example, Wang-Jones, Alhassoon, Hattrup, Ferdman, and Lowman (2017) found that 15% of online participants were unable (or unwilling) to complete such an IAT. A third option is Gorilla (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2019), which also allows researchers to run reaction-time studies. Like Inquisit, researchers pay to use the tool (fixed fee per participant), and participants are directed to and from the platform via hyperlink (data merged in analysis). A fourth option, which is less expensive but requires greater technical expertise, is to develop and host custom JavaScript-based IATs from scratch. Options such as the jsPsych code library (de Leeuw, 2015) simplify this process; however, it still requires considerable technical expertise (for a review of code libraries, see Anwyl-Irvine et al., 2019). All of the above are time- and labor-intensive. For those researchers who primarily use survey software, these options pose numerous challenges.

Survey-software IATs Given the issues above, it would be advantageous for researchers who primarily use online survey tools to be able to construct IATs directly within those tools. Not only would this reduce cost and simplify study design, but it would allow greater design flexibility and allow all data to be stored in one file, simplifying management. We examined whether it is possible to build such an IAT within online survey software—focusing here on Qualtrics, commercial survey software that is commonly used in psychology.[1] Critically, Qualtrics allows users to add custom JavaScript and HTML to the survey elements—code that could be used to create reaction-time processing and interactivity for web-based studies (de Leeuw, 2015). Thus, it may be possible to build a functional IAT within Qualtrics by adding code to survey elements to make them interactive and capture reaction times. This would allow researchers who use Qualtrics to build and embed IATs within their online surveys and experiments. Because Qualtrics is currently the only mainstream survey tool to allow custom JavaScript, our approach is dependent for the moment on Qualtrics and would likely be most useful for researchers who already use it or have access to it.

With the goal of developing a survey-software IAT, we proceeded in two stages. First, we assessed viability. Although the premise is simple, designing a valid procedure to execute it is not. Online survey tools are complex, and several issues need to be addressed. How would such a survey be designed? What code should be added? What issues must be avoided to ensure validity? These methodological issues need to be solved before researchers can implement survey-software IATs. Second, we empirically assessed the validity of survey-software IATs. Empirical validation is especially necessary given the potential for widespread use of survey-software IATs on online platforms. Thus, a rigorous empirical test of survey-software IATs was our second objective.

The IAT procedure

The IAT assesses the degree to which target pairs (e.g., *insects* vs. *flowers*) and categories (e.g., *pleasant* vs. *unpleasant*) are mentally associated (see Lane, Banaji, Nosek, & Greenwald, 2007, for in-depth review). Participants place hands on the keyboard and complete seven blocks of stimuli sorting trials. In each trial, a word or image appears on the screen representing a category or target. The participant sorts the

[1] Although Qualtrics appears to be widely used, precise statistics are difficult to obtain, because few articles cite survey platforms. However, many study protocols are available on the Open Science Framework (www.osf.io). A search (July 16, 2019) returned 2,546 results referring to "Qualtrics." By contrast, "survey monkey" and "surveymonkey" returned only 57 results, and "Inquisit" returned 237 results. Similarly, a PsycINFO search returned 408 results for Qualtrics, 298 results for Survey Monkey, and 9 for Inquisit. Of the tools queried, Qualtrics was thus the most prevalent. Finally, a spokesperson for Qualtrics reported that as of July 16, 2019, "approximately 2,000 universities" (personal communication) have active research accounts, making it available for thousands of researchers to use.

stimulus by pressing a key with the designated hand (e.g., left for *insects* or *unpleasant*; right for *flowers* or *pleasant*). During the sorting, stimuli alternate between target trials (e.g., insects and flowers) and category trials (e.g., pleasant and unpleasant words; Greenwald et al., 1998). An interstimulus interval of 250 ms is also added between trials (Greenwald et al., 1998; Penke, Eichstaedt, & Asendorpf, 2006; used by default in our materials), although this value is seldom reported and varies across articles (e.g., 150–300 ms in Bluemke & Friese, 2006, 2008; Schnabel, Banse, & Asendorpf, 2006). When participants make errors, they either see an error message for 300 ms (Greenwald et al., 1998) or correct errors before proceeding (Greenwald, Nosek, & Banaji, 2003; Rudman, 2011; used by default in our materials).

The premise behind the IAT is that one can more rapidly sort stimuli when pairings are compatible with associations. For example, if one sees insects as less pleasant than flowers, one should respond faster when using the same hand for *insects* and *unpleasant* stimuli and the other hand for *flowers* and *pleasant* stimuli (known as a "compatible block"). Conversely, one should be slower when pairings are reversed (an "incompatible block"). Participants complete the procedure under both conditions; an association is indicated if people are faster in one condition relative to the other.

A full IAT takes approximately 5 min. Block 1 is a practice block (20 trials) of only targets (e.g., *insects*, *flowers*); Block 2 is a practice block (20 trials) of only categories (e.g., *pleasant*, *unpleasant*). Next is a combined block (e.g., "incompatible" block: *insects* + *pleasant*, *flowers* + *unpleasant*) using both targets and categories; the hand pairings are determined by the initial left/right assignments in the previous blocks (randomized). This is subdivided into 20 practice trials (Block 3) and 40 critical trials (Block 4; scoring uses data from B3 and B4). Following this is another practice block (Block 5), consisting of the categories with the sides reversed (e.g., *unpleasant*, *pleasant*). This helps wash out left–right associations learned in the early blocks; on the basis of empirical analysis, Nosek, Greenwald, and Banaji (2005) recommended 40 trials. Finally, participants repeat the combined block with the categories in their reversed positions (e.g., "compatible" block: *insects* + *unpleasant*, *flowers* + *pleasant*). As before, this is divided into 20 practice trials (Block 6) and 40 critical trials (Block 7).

Data in the combined blocks (B3 + B4 and B6 + B7) are then analyzed. A standardized difference score (*D* score) is calculated for each participant, indicating in which condition (compatible vs. incompatible) participants were faster. A *D* score of 0 indicates no difference in speeds; a positive score indicates that one was faster in the compatible block; and a negative score indicates that one was faster in the incompatible block. This procedure was validated by Greenwald et al. (2003, see Lane et al., 2007, for step-by-step directions), who found it maximized IAT reliability and yielded other advantages (see Back, Schmukle, & Egloff, 2005). *D* scores can then be used in statistical analyses.

Although the procedure described above has resulted in a fruitful literature, it should be noted that the IAT is not without limitations. Although it is more reliable than many "implicit" procedures, test–retest reliability for the IAT remains low by self-report standards ($r = .54$; Gawronski, Morrison, Phills, & Galdi, 2017). Furthermore, the IAT assesses only *relative* associations between targets. For example, it cannot reveal whether one associates *flowers* + *pleasant* in an absolute sense (cf. Lane et al., 2007). So long as these limitations are taken into account, however, the IAT remains an important tool for researchers (Greenwald & Banaji, 2017).

A survey-software IAT methodology

We next describe a procedure for running a survey-software IAT. The concept introduced here is novel and could be applied using materials other than those provided. We nevertheless provide all materials—templates, examples, editable code, and so forth—as well as an open-source tool that automates the procedure. Researchers who wish to run survey-software IATs can therefore do so freely and easily. Readers can also find all materials and supplements on the Open Science Framework (OSF) page for this project at https://osf.io/ntd97/, in addition to several tutorials and examples available for download. Although our tool is presently configured with respect to Qualtrics, our procedure could be adapted to any survey tool that allows for custom code.

Our general procedure is to add HTML and JavaScript code to survey questions such that their original functionality is "overridden" and replaced with components of an IAT. We build the IAT out of modified "text-entry" questions, which have the ability to store considerable data (in a hidden text field). Each text-entry question is converted into a block of IAT trials. This allows the entire block to execute without requiring data transfer *during* the task (which could impair response-time measurement).

The survey-software IAT consists of seven modified survey questions (one for each IAT block). Four permutations are built, counterbalancing left/right starting positions of targets and categories (Nosek et al., 2005), with one randomly assigned to each participant (in Qualtrics,[2] via a randomizer; Fig. 1). Researchers can change the position of the IAT by moving this randomizer or by placing it inside other randomizers (e.g., to counterbalance implicit/explicit measures) so

[2] In our materials, permutations are named (**RP** = Target A initially on right with positive; **RN** = Target A starts right with negative; **LP** = Target A starts left with positive; **LN** = Target A starts left with negative). Question names include this code (e.g., "RN2" for the second block in this permutation).

long as the randomizer's contents are left intact.[3] Any participant with a computer and keyboard (but not mobile users; this should be noted in recruitment materials) should be able to complete this IAT. A schematic table of each Qualtrics question is given in Table 1.

What code should be added? Both HTML and JavaScript code must be added to each question to create the IAT—this is done automatically via our tool or could be done manually using our materials. HTML creates the visual appearance, whereas JavaScript adds interactivity. In our procedure, HTML is added into the 'question text' and generates visible features of the IAT, such as the on-screen labels and instructions (example surveys containing code available on our OSF page). Elements that appear or disappear (e.g., instructions, fixation cross, etc.) are named; this allows the JavaScript to reference and assign interactivity to them.

The JavaScript makes the IAT functional (view/edit code in Qualtrics by clicking "Add JavaScript" under question preferences; examples available on our OSF page; our code is annotated for readers). First, it hides the text box and next button. Next, if present, images are downloaded (prior to the start of the block) and stored in memory. Following this, stimuli for the block are selected from the pools. Following Greenwald et al. (1998), stimuli are sampled randomly without replacement (resetting the pools when they are depleted). In other words, each stimulus is sampled before repeats are allowed, and each stimulus is sampled again before triplicates are allowed, and so forth. This ensures that the proportions of stimuli within blocks is consistent, keeping the contents of the IAT consistent across IAT blocks and participants. The resulting stimuli are placed into an array representing individual trials (target/category stimuli appear in an alternating format for combined blocks; Greenwald et al., 1998).[4]

The code then listens for the participant to press the space bar, which commences the IAT. Reaction times are captured following de Leeuw (2015), recording timestamps at the start and end of each trial and saving the differences in milliseconds. This procedure was further validated by Reimers and Stewart (2015). Across a variety of machines for reaction times ≈ 300–600 ms, they found latencies were largely consistent within machines (within-system *SD*s ≈ 10 ms) and were inflated by approximately 50–80 ms. Differences across web browsers were small (typically < 10 ms). Thus, they concluded, although a given latency may be inflated somewhat, the degree of inflation should be roughly constant for a given machine and therefore be largely irrelevant for within-subjects tests such as the IAT.

When the participant completes the last trial, the code writes results to the question's hidden text field, saving it with the survey data, and advances to the next "question" (IAT block). This automatic advancement allows all seven IAT blocks to flow seamlessly together. Because error handling varies in IAT research, we developed two versions of our code. Originally, IAT participants saw an error message for 300 ms when an error was made (Greenwald et al., 1998). In a newer variant, the message persists until the error is corrected (Greenwald et al., 2003). Code-savvy users can modify other aspects of our code (e.g., colors, interstimulus interval, etc.). Our automated tool includes several such options, described in detail in our supplementary materials.

How should the data be stored? Because each IAT block is built from a text-entry question, data are saved as strings of text. In our procedure, results from individual trials are stored as structured text strings separated by commas. Each trial result contains a number identifying the stimulus, whether the response was correct ("C") or incorrect ("X"), and the reaction time in ms. For example, Stimulus 12 answered correctly in 732 ms is 12C732.[5]

After data collection, the IAT must be scored, to produce a "*D* score" for each participant (Greenwald et al., 2003). This can be done in any program that can parse text data (e.g., Excel, R, etc.). We provide R scripts that work with our materials to automatically perform Greenwald et al.'s data cleaning and scoring algorithm[6] (instructions/example at github.com/iatgen/iatgen). Alternatively, our automated tool allows researchers to upload raw data from Qualtrics (must be in "Legacy Export" CSV format), which it then scores automatically, displaying diagnostics and making clean *D* scores available for download that can be easily merged with the source data.

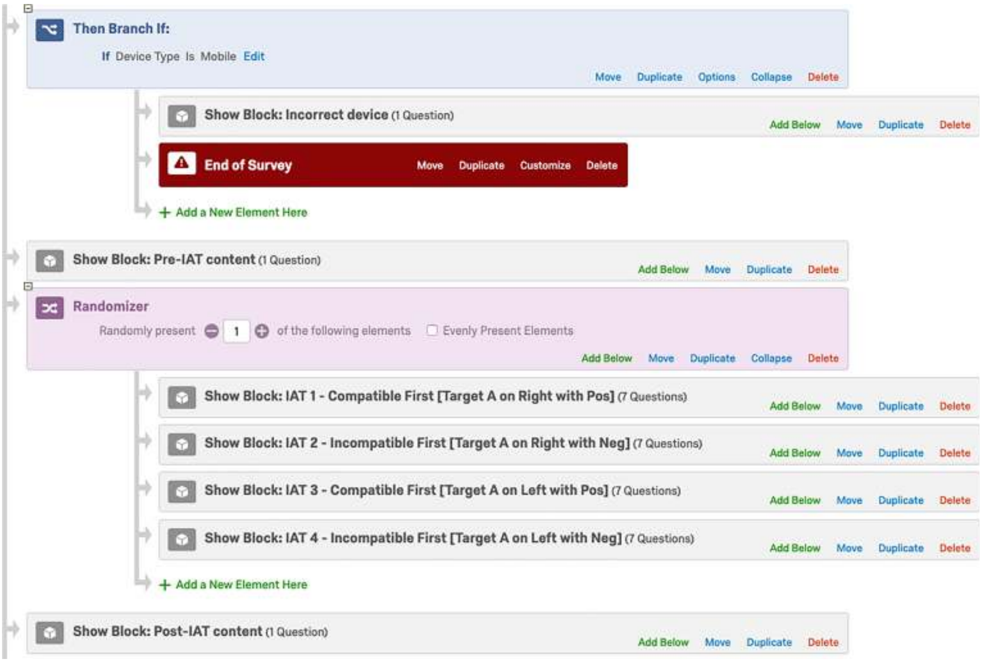
How are images handled? It is important that image download not interfere with reaction times; our solution was to

[3] For the IAT to function in Qualtrics, one must also not change the visual theme or use animated page transitions. Our code is optimized around a visual theme (*Minimal 2014*). Changing this may require modifying our code. Animated "page transitions" interfere with the IAT code.

[4] By default, our code randomizes the order of the selected stimuli within each block (after alternating target/category trials; Gawronski, 2002). Following comments from a reviewer, we also include an option to display stimuli in the order sampled—that is, without displaying repeats to the participant until all other stimuli from that category have appeared. This can be selected in our R package/automated tool, and raw code is available in our online materials.

[5] The meanings of stimulus numbers can be found in our JavaScript code, looking under the "STIMULI POOLS" section, and examining the table of "index" numbers.

[6] Per Greenwald et al. (2003; see also Lane et al., 2007), trials over 10,000 ms are deleted (scored as missing), as is as any IAT data from participants with > 10% of responses < 300 ms. If participants were not forced to correct errors, errors are scored as the participant's block mean plus a penalty (typically 600 ms); otherwise, no penalty is added for errors. Next, within-person difference scores are calculated using each participant's block means—once for practice combined blocks (Blocks 3 + 6) and once for the critical combined blocks (Blocks 4 + 7). These are divided by their inclusive standard deviations (i.e., *SD* of "Block 3 merged with 6" and *SD* of "Block 4 merged with 7"), generating two scores per participant. These are then averaged, creating a single *D* score.



Then Branch If:

If Device Type Is Mobile Edit

Show Block: Incorrect device (1 Question)

End of Survey

Show Block: Pre-IAT content (1 Question)

Randomizer

Randomly present 1 of the following elements □ Evenly Present Elements

Show Block: IAT 1 - Compatible First [Target A on Right with Pos] (7 Questions)

Show Block: IAT 2 - Incompatible First [Target A on Right with Neg] (7 Questions)

Show Block: IAT 3 - Compatible First [Target A on Left with Pos] (7 Questions)

Show Block: IAT 4 - Incompatible First [Target A on Left with Neg] (7 Questions)

Show Block: Post-IAT content (1 Question)

Fig. 1 Screenshot of the survey flow in Qualtrics containing an implicit association test (IAT). The IAT consists of the randomizer, containing four Qualtrics blocks; each block is one permutation of the IAT

download images *prior* to the start of each block and store them in a JavaScript object (in our code: called “images,” with individual stimuli referred to as images[0], images[1], images[2], etc.). To ensure access, we recommend that users host their own images (e.g., personal website or within Qualtrics). We recommend images be in the .PNG or .JPG format and be approximately 250 pixels high (tutorial on our OSF page).

How can researchers implement this procedure? Using our materials, researchers have two options to implement our procedure. First, researchers could edit raw code in our survey examples or code files provided on our OSF page. Only two portions of the code need to be modified: the on-screen labels in the HTML and the stimuli in the “STIMULI POOLS” section of the JavaScript. In both cases, this is simply a matter of changing out our insects/flowers example with the desired targets and categories.

However, to make the procedure more accessible, we created a supplementary tool that rapidly and automatically implements the procedure described in this article. This tool customizes all code and pastes it into a survey template. It is available both as a web-based tool (http://iatgen.org; tutorial on OSF page and loadable example available within the tool) and as the `iatgen` R package (version 1.2.4; tutorial, instructions, and source code available at github.com/iatgen/iatgen). The web-based tool allows IAT creation (and analysis) via on-screen menus, providing a downloadable ready-to-run Qualtrics IAT survey; the R package does the same but also can also generate JavaScript and HTML for manual editing if desired.

Empirical validation

It is an empirical question whether an IAT, constructed out of modified survey elements, would be both valid and accurate enough to compete with reaction-time software. We tested the validity of survey-software IATs across three online studies, examining whether survey software IATs could (1) detect IAT effects, (2) correlate as expected with explicit measures, (3) demonstrate sound psychometrics (e.g., error rates, internal consistency, etc.), and (4) produce results comparable to traditional IATs. Materials, data, and analysis code for all studies are included on our OSF page (https://osf.io/ntd97/).

Study 1

Study 1 examined a survey-based IAT testing associations between atheists (vs. theists) and distrust (vs. trust). We selected atheists as a target as a part of a separate, broader program of research on interreligious bias. We chose atheists because prejudice towards atheists is common and associated with belief in God (Gervais, Shariff, & Norenzayan, 2011; Gervais et al., 2017). We constructed an atheist/trust IAT and examined correlations with a battery of measures used in religion/prejudice research with a high-powered online sample. As an initial test of the survey-software IAT, we expected to see a large IAT effect, sound performance from the IAT, and correlations with explicit measures in the range expected for prejudice IATs ($r \approx .22$; Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005).

Table 1 Schematic diagram of a survey-software implicit association test (IAT) in Qualtrics (insect–flower example)

Qualtrics Block/IAT Permutation	Qualtrics Question	IAT Block	Left Hand	Right Hand
IAT 1: Compatible first [Flowers on Right + Pos]				
	Q1	B1	Insects	Flowers
	Q2	B2	Unpleasant	Pleasant
	Q3	B3	Insects + unpleasant	Flowers + pleasant
	Q4	B4	Insects + unpleasant	flowers + pleasant
	Q5	B5	Pleasant	Unpleasant
	Q6	B6	Insects + pleasant	Flowers + unpleasant
	Q7	B7	Insects + pleasant	Flowers + unpleasant
IAT 2: Incompatible first [Flowers on Right + Neg]				
	Q8	B1	Insects	Flowers
	Q9	B2	Pleasant	Unpleasant
	Q10	B3	Insects + pleasant	Flowers + unpleasant
	Q11	B4	Insects + pleasant	Flowers + unpleasant
	Q12	B5	Unpleasant	Pleasant
	Q13	B6	Insects + unpleasant	Flowers + pleasant
	Q14	B7	Insects + unpleasant	Flowers + pleasant
IAT 3: Compatible first [Flowers on Left + Pos]				
	Q15	B1	Flowers	Insects
	Q16	B2	Pleasant	Unpleasant
	Q17	B3	Flowers + pleasant	Insects + unpleasant
	Q18	B4	Flowers + pleasant	Insects + unpleasant
	Q19	B5	Unpleasant	Pleasant
	Q20	B6	Flowers + unpleasant	Insects + pleasant
	Q21	B7	Flowers + unpleasant	Insects + pleasant
IAT 4: Incompatible first [Flowers on Left + Neg]				
	Q22	B1	Flowers	Insects
	Q23	B2	Unpleasant	Pleasant
	Q24	B3	Flowers + unpleasant	Insects + pleasant
	Q25	B4	Flowers + unpleasant	Insects + pleasant
	Q26	B5	Pleasant	Unpleasant
	Q27	B6	Flowers + pleasant	Insects + unpleasant
	Q28	B7	Flowers + pleasant	Insects + unpleasant

B1–B7 = Blocks within the IAT, each represented by one survey question (Q1–Q28). Category = pleasant/unpleasant, Target = insect/flower. The four IAT permutations represent left–right counterbalancing and are built separately and randomly assigned. Preconfigured code and a code-configuration tool are available in our online materials.

Method

Participants A total of 239 students at a state university in New England (66% female; M_{age} = 19.20, SD_{age} = 2.75) participated for course credit. Participants were predominantly White (91%; 4% Black; 1% Asian, Hispanic, Native American, and Other). Most believed in God (60%); the plurality were not religiously affiliated (39%; 34% Catholic; 20% Protestant; 4% Other; 1% Buddhist, Jewish, Muslim). Data collection was terminated after one academic term, provided that we had enough participants to detect the average I/E correlation (ρ = .22; 80% power: $n \geq 157$).

Implicit measures Participants completed a survey-based IAT in Qualtrics comparing "Atheist" and "Theist" targets on a "Trust" versus "Distrust" dimension (the stimuli are given in the Appendix), with the goal of assessing relative trust for theists over atheists. Error feedback was provided by displaying an "X" for 300 ms (Greenwald et al., 1998).

Explicit measures Participants completed a battery of explicit measures of Atheist attitudes and known predictors of religious intergroup bias to investigate convergent validity of the IAT. Positive atheist attitudes were measured using a 14-item scale (e.g., "I like Atheists"; "I feel positively toward Atheists"; 1 = *strongly disagree*, 7 = *strongly agree*; M = 64.35, SD = 15.56; α = .95; Pittinsky, Rosenthal, & Montoya, 2011). This was supplemented with six semantic differentials scored as negative attitudes (from 1 = *desirable*, *pleasant*, *positive*, *agreeable*, *cooperative*, or *friendly*, to 7 = *undesirable*, *unpleasant*, *negative*, *disagreeable*, *noncooperative*, or *hostile*; M = 21.89, SD = 6.91; α = .92). The intergroup emotion was assessed with the statement, "When I interact with Atheists, I feel . . ." and providing ten semantic differentials (e.g., 1 = *irritable*, *fearful*, or *supportive*; 7 = *patient*, *confident*, or *resentful*; M = 46.66, SD = 9.55; α = .93; Capozza, Vezzali, Trifiletti, Falvo, & Favara, 2010).

Participants also completed affective measures, including five trust items (e.g., "most Atheists are trustworthy"; 1 = *strongly disagree*, 7 = *strongly agree*; M = 25.72, SD = 5.15; α = .96), and 12 items indicating anticipated anxiety (e.g., worried, anxious, nervous) when interacting with atheists (1 = *not at all*, 5 = *completely*; M = 25.85, SD = 10.16; α = .94; Stephan et al., 2002). Empathy for atheists was assessed following Kuchenbrandt, Eyssel, and Seidel (2013), embedding six empathy words (stirred, compassionate, affectionate, warmhearted, moved, sympathetic; 1 = *not at all*, 5 = *completely*; M = 16.85, SD = 3.99; α = .78) among filler items.

Beyond explicit measures of prejudice, we assessed known predictors of atheist prejudice (LaBouff & Ledoux, 2016). We assessed intergroup contact (Vezzali, Capozza, Giovannini, & Stathi, 2012). Participants were asked "with how many Atheists do you frequently associate?" (1 = *none*, 5 = *six or more*; M = 3.07, SD = 1.38) and "how many Atheists would you like to frequently associate with"? (M = 3.42, SD = 1.38). Participants also estimated the number of atheists they knew (M = 22.91, SD = 24.52) and anticipated knowing in 5 years (M = 38.14, SD = 29.51). Finally, participants were asked, "if you knew someone was an Atheist, how interested would you be in striking up a conversation with them?" (1 = *not at all*, 5 = *interested*; M = 3.43; SD = .86).

Given the role of religiosity in atheist prejudice (Gervais et al., 2011), participants completed single-item measures of personal religiosity (1 = *not at all religious*, 7 = *very much religious*; M = 2.64, SD = 1.68), interest in religion (1 = *not at all*, 9 = *extremely*; M = 4.05, SD = 2.19), and belief in God (dichotomous: yes/no; 138 = yes). Participants also completed a 12-item Religious Fundamentalism measure, assessing commitment to the absolute and inerrant nature of one's religious beliefs (e.g., "Whenever science and sacred scripture conflict, science is probably right"; 1 = *strongly disagree*, 7 = *strongly agree*; M = 34.57, SD = 14.34; α = .94; Smith & Winter, 2002). Finally, we administered the Authoritarianism–Conservatism–Traditionalism scale (ACT; Duckitt & Bizumic, 2013), which includes 12 authoritarianism items (e.g., "What our country really needs is a tough, harsh dose of law and order"; 1 = *strongly disagree*, 7 = *strongly agree*; M = 48.09, SD = 9.34; α = .83), 12 conservatism items (e.g., "Our leaders should be obeyed without question"; M = 45.76, SD = 11.12; α = .87), and 12 traditionalism items (e.g., "It is important that we preserve our traditional values and moral standards"; M = 37.29, SD = 11.74; α = .88).

Results

The IAT data were processed using the D-score data-cleaning and scoring algorithm (Greenwald et al., 2003; Lane et al., 2007). The dropped-trial (trials > 10 s) rate was low, at < 0.001% of trials. The rate of dropped participants due to excessively fast responses was low at 1% (three participants). Next, errors were replaced with participant block means of correct trials plus 600 ms (i.e., D_{600} procedure; Greenwald et al., 2003). The error rate was 10%. Finally, the IATs were D-scored, with positive scores indicating trust toward atheists. The internal consistency of the IAT was assessed via two popular methods, a split-half procedure (De Houwer & De Bruycker, 2007), estimate = .86, and a variant of Cronbach's alpha (Schnabel, Asendorpf, & Greenwald, 2008), α = .85.

IAT results We observed a strong effect in the survey software IAT: scores trended toward atheist distrust, $M_{D\text{ Score}} = -0.60$, SD = 0.40, d = 1.48, which significantly differed from zero, $t(235) = -22.70$, $p < .001$, 95% $\text{CI}_{\text{D Score}}$ [− 0.65, − 0.55]. The distribution is shown in Fig. 2. This score is interpreted as a relative measure of distrust of atheists as compared to distrust of theists. We compared this implicit distrust of atheists/trust of theists to explicit measures of intergroup attitudes and predictors of atheist prejudice.

Explicit measures The IAT D scores were associated with less negative explicit attitudes toward atheists, $r = -.29$, $p < .001$; more positive explicit attitudes, $r = .38$, $p < .001$; more positive overall emotion, $r = .24$, $p < .001$; less intergroup anxiety, $r = -.21$, $p = .001$; greater explicit trust of atheists, $r = .21$, $p = .001$; more frequent self-reported contact (actual and desired) on all measures, $|r|$s = .16 to .28, ps ≤ .01; lower personal religiosity, $r = -.29$, $p < .001$; less interest in religion, $r(234) = -.18$, $p = .004$; lower personal spirituality, $r = -.16$, $p = .01$; and lower religious fundamentalism, $r = -.20$, $p = .002$. In addition, higher IAT scores were associated with less authoritarianism, $r = -.19$, $p = .004$; conservatism, $r = -.22$, $p < .001$; and traditionalism, $r = -.21$, $p = .001$. Empathy toward atheists was not significantly correlated with the IAT, $r = .06$, $p = .35$. People who believed in God had a significantly stronger IAT effect, $M_{D\text{ Score}} = -0.69$, SD = 0.35, than did

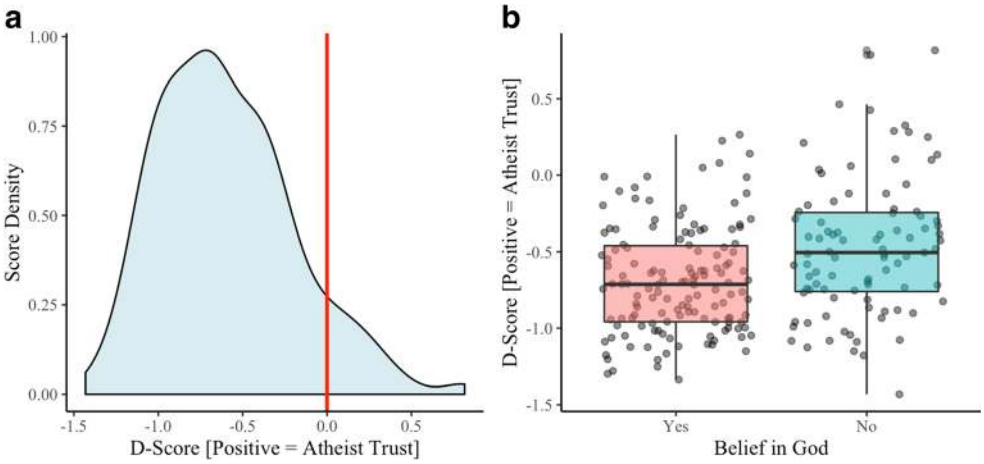


Fig. 2 (a) *D*-score distribution in Study 1. The majority of the sample (91%) had negative (< 0; i.e., distrust) associations. (b) Group differences in implicit association test (IAT) scores, showing significantly stronger responses among those reporting a belief in God

those who reported no belief, $M_{D\text{ Score}} = -0.47$, $SD = 0.45$ $t(227) = -4.11$, $p < .001$, 95% CI_{diff} [– 0.32, – 0.11], $d = -0.55$. This result is visualized in Fig. 2.

Discussion

Study 1 served as an initial demonstration of the validity of the survey-software IAT. Internal consistency was close to the meta-analytic average (.79; Hofmann et al., 2005). Error rates were reasonable; Rudman (2011) identified error rates at 5%–10% for most IATs. The survey-software IAT identified a strong IAT effect; this was also comparable with existing prejudice IATs (Lane et al., 2007). Finally, although the observed correlations between our IAT and self-report measures were weaker than some of the strongest observed relationships (e.g., $r = .60$; Gervais et al., 2011) in smaller samples with different stimuli, our results were consistent with implicit/explicit correlations in meta-analyses that reduce the variability found in a single study (average $r = .22$, Hofmann et al., 2005). Encouraged by these results, we next sought to examine established IAT effects using participants from MTurk, an online participant pool.

Study 2

In Study 2, participants completed one of four survey-software IATs: insect–flower attitudes (Greenwald et al., 1998), Black–White race attitudes (Nosek et al., 2002), gender–career stereotypes (Nosek et al., 2002), and Mexican–American attitudes. We used survey-software IATs with a mixture of stimulus modalities (image, text).

Participants The participants were 818 individuals (60% male, $M_{\text{age}} = 35.06$, $SD = 11.89$, age range: 18–84 years) recruited on MTurk.com from 50 states and Washington DC. Participants were predominantly White (72%; 10% Black, 10% Asian, 7% Hispanic, 4% Native, 1% Islander and Other), from the United States, had at least a 95% approval rating (Peer, Vosgerau, & Acquisti, 2014), and received $0.60 for participation. We sought $n = 200$ per IAT (total $n \sim 800$), giving 80% power per IAT to detect effects as small as $d = 0.20$, although IAT effects are typically much larger (Lane et al., 2007).

Procedure Participants were randomly assigned to one of four IATs (the stimuli were all drawn from published sources; see the Appendix). In the Black–White race IAT, participants sorted face images along with pleasant/unpleasant words (Nosek et al., 2002). In the insect–flower IAT, participants sorted insect/flower words and pleasant/unpleasant words (Greenwald et al., 1998). In the gender–career stereotypes IAT, participants sorted male/female words and family/career words (Nosek et al., 2002). In the Mexican–American IAT, participants compared "American Surname" and "Mexican Surname" categories; the names were the seven most common surnames in each country (omitting Garcia, which was also common in the US; Instituto Nacional Electoral, 2005; United States Census Bureau, 2010). For this IAT, groups were compared on the same pleasant/unpleasant dimension as in the race and insect IATs.

Results

The primary results are given in Table 2. Effect sizes ranged from $d = 0.82$ (gender–career IAT) to $d = 2.13$ (insect–flower IAT), with all *t*s > 10.99 and all *p*s < .001. Error rates were comparable to the student sample (8%–10%) and reliabilities were satisfactory (.68–.87), although the drop rate for overly fast responding was somewhat higher on MTurk than in the student sample (11%–18%). Although it was not a primary

Table 2 Implicit association test (IAT) results from Study 2

	n	Scored	*M*	*SD*	95% CI	*t*	*d*	Rel.	*α*	Err.	Drop
Black–White	200	172	.36	.43	[.30, .42]	10.99***	0.84	.86	.86	.08	.14
Insect–Flower	208	182	.72	.34	[.67, .77]	28.68***	2.13	.87	.86	.10	.12
Gender–Career	205	182	.32	.39	[.26, .38]	11.09***	0.82	.80	.77	.08	.11
Mexican–American	205	169	.48	.35	[.43, .53]	17.87***	1.37	.68	.73	.09	.18

*** $p < .001$. Rel. = split-half reliability; Err. = error proportion; Drop = proportion of participants dropped for > 10% of responses < 300 ms (Greenwald et al., 2003).

aim, we also examined racial differences in the two racial-attitude IATs.

On the Black–White IAT, we found a large effect of race, $t(139) = 5.30$, $p < .001$, $d = 1.41$, with exclusively White participants having more pro-White scores ($M = 0.42$, $SD = 0.40$) than exclusively Black participants ($M = – 0.14$, $SD = 0.43$). We observed a similarly large difference on the Mexican–American IAT, $t(134) = 3.48$, $p < .001$, $d = 1.14$, with non-Latinx White participants favoring American surnames ($M = 0.51$, $SD = 0.34$) more than non-White Latinx participants ($M = 0.13$, $SD = 0.33$).

Discussion

As with the college sample in Study 1, the survey-software IATs in Study 2 detected strong, significant implicit effects. Effect sizes were large and consistent with published research (e.g., Greenwald et al., 1998; Lane et al., 2007; Nosek et al., 2002). Despite the fact that the MTurk population has a higher proportion of White individuals than the rest of the United States (Buhrmester et al., 2018), racial differences in the IATs mirror those reported by Nosek et al. (2002). Thus, we observed IAT effects consistent with those known in the literature. One limitation, however, is that we did not directly compare it against a reaction-time software IAT in the same sample. We addressed this limitation in Study 3.

The MTurk sample also had slightly higher drop rates due to overly fast responding, with rates ranging from 11%–18%. Although the Mexican–American IAT had a higher drop rate than the others, a 2 (drop) × 4 (IAT) contingency table analysis revealed no significant differences in drop rates across IATs, $\chi^2(3) = 3.88$, $p = .27$. Nevertheless, drop rates were higher than in Study 1. This can happen when participants indiscriminately press buttons at high speed to skip past the IAT (aka, “button mashing”). Consistent with this, the error rate for all combined trials for all dropped participants was close to even: 48% for the Black–White IAT, 50% for the insect–flower IAT, 47% for the gender–career IAT, and 49% for the Mexican–American IAT. As we noted earlier, the Greenwald et al. (2003) scoring procedure removes these participants; error rates for the remaining participants were within the standard range for IAT research (cf. Rudman, 2011) and indicate that the vast majority of the participants did follow the instructions.

Study 3

We observed IAT effects consistent with past research in Studies 1 and 2. However, neither study directly compared a survey-software IAT with one run via reaction-time software. This was the goal of Study 3. Participants completed both survey-software and traditional (Inquisit) IATs, within subjects. In addition to assessing whether *D* scores were comparable across measures, we also examined whether survey-software and Inquisit IATs would correlate at levels expected for identical IATs. However, as the expected correlation is somewhat low ($r \approx .54$; Gawronski et al., 2017) and may vary by IAT, we included a second Inquisit IAT to help benchmark this value. Finally, we sought to expand our validation of the survey-software IAT by examining an IAT in the consumer domain, a common use for the IAT (Hofmann et al., 2005; Pogacar, Kouril, Carpenter, & Kellaris, 2018). We focused here on ice cream, a product that evokes strong explicit and implicit reactions (Carpenter, 2015). Furthermore, implicit/explicit (I/E) correlations tend to be higher in the consumer domain (estimated $\rho = .36$; Hofmann et al., 2005). We expected such correlations in Study 3.

Method

Participants The participants were 270 individuals (65% male; $M_{\text{age}} = 33.50$, $SD = 9.46$; age range = 19–70) recruited on MTurk.com from 43 states and Washington DC. The participants identified as White (66%; 21% Black, 9% Asian, 6% Hispanic, 2% Native, 1% Islander and Other), followed the same inclusion criteria from Study 2, and received $3.00.

Because we sought to compare *D* scores within subjects across IATs, we powered our study to detect a trivial-to-small within-subjects effect (Cohen's $f = .10$), requiring a minimum sample size for 80% power of $n = 150$ (calculations based on the meta-analytic mean IAT test–retest, $r = .54$; Gawronski et al., 2017). Because of the potential for missing data (i.e., participants not completing the software-based IAT), we collected the largest sample possible given the resources, which exceeded our minimum ($n = 270$).

Implicit measure The IAT was a consumer-preference IAT adapted from Carpenter (2015). The stimuli were photos of Ben & Jerry's and Haagen Dazs ice creams. Flavors were matched to create balanced stimulus sets (the materials are available on the OSF page). These were evaluated on a "good" versus "bad" dimension (see the Appendix).

Participants completed both a survey-software IAT and two traditional IATs run via third-party reaction-time software (Millisecond's Inquisit software; www.millisecond.com). Participants first completed a survey-software IAT and an Inquisit IAT, in counterbalanced order. An additional Inquisit IAT was completed last, providing a benchmark for the test–retest correlation for this IAT. For the Inquisit IAT, participants received a link to the Inquisit platform from within the survey; they were instructed to download and install Inquisit and then to open the IAT. At the end of the Inquisit IAT, participants were provided with a completion code and instructed to return to Qualtrics and enter this code. For consistency, participants completed the same permutation for all IATs. For example, if Ben & Jerry's appeared on the right and was initially paired with the positive category, it did so in all three IATs.

Explicit measures Participants also completed an explicit brand preference measure from Carpenter (2015). Participants imagined they were shopping and deciding between Ben & Jerry's or Haagen Dazs, considering preferences for 15 s and thinking through reasons for their preferences. Participants then indicated which brand they viewed more positively, which they perceived to be better, and which they preferred more. All items were scored on a 7-point scale such that a score of 4 represented no preference and higher scores indicate preference for Ben & Jerry's. Scales appeared horizontally on the screen, and left/right positioning was matched to the positioning in the IATs. The measure was reliable, $\alpha = .96$.

Results

IAT diagnostics All participants completed the survey-software IAT ($n = 270$). In contrast, given that the Inquisit IAT was completed via an external link and required a software download, fewer participants completed it (completed *n*s = 248 and 239). Most (71%) who skipped one Inquisit IAT skipped both. The drop rates due to excessive speed were similar across all three IATs: 60 participants (22% of completes) for the survey-software IAT, 58 participants (23% of completes) for the first Inquisit IAT, and 53 participants (22% of completes) for the final Inquisit IAT.[7] The dropped trial (trials > 10 s) rates were low and comparable for all IATs (< 1% of trials). The error rates were also within the expected range (8% for all three IATs). The survey-software and first Inquisit IAT were similarly reliable (survey software split-half estimate = .80, $\alpha = .83$; Inquisit split-half estimate = .72, $\alpha = .84$). The second Inquisit IAT (which was completed last for all participants) had somewhat lower consistency (split-half estimate = .55, $\alpha = .65$).

Comparing D scores Similar results were obtained for each IAT. Participants slightly favored Ben & Jerry's on the survey software IAT, $M_{D\ \text{Score}} = 0.10$, $SD = 0.40$, $d = 0.25$, $t(209) = 3.66$, $p < .001$, 95% $\text{CI}_{D\ \text{Score}}$ [0.05, 0.15]. Nearly identical results were observed on the first Inquisit IAT, $M_{D\ \text{Score}} = 0.09$, $SD = 0.43$, $d = 0.22$, $t(189) = 3.03$, $p = .003$, 95% $\text{CI}_{D\ \text{Score}}$ [0.03, 0.16]. The same was again observed on the second Inquisit IAT, $M_{D\ \text{Score}} = 0.10$, $SD = 0.35$, $d = 0.28$, $t(185) = 3.79$, $p < .001$, 95% $\text{CI}_{D\ \text{Score}}$ [0.05, 0.15]. The results are plotted in Fig. 3a. For interested readers, descriptive statistics for reaction times are given in Table 3.

We next compared these *D* scores via within-subjects analyses of variance. Although adequately powered,[8] this analysis detected no significant effect of IAT modality, with a near-zero estimated effect size, $F(2, 366) = 0.004$, $p = .99$, $\omega_p^2 < .0001$. Planned contrasts using within-subjects Fisher–Hayter post-hoc tests (Kirk, 2008) showed no difference between the survey-software IAT and either the first Inquisit IAT, $q(3, 366) = -0.01$, $p > .99$, or the second Inquisit IAT, $q(3, 366) = 0.11$, $p = .94$. The two Inquisit IATs did not differ from one another, $q(3, 366) = 0.11$, $p = .94$. Thus, the IATs appeared comparable.

To supplement the analysis above, Bayesian *t* tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009) can quantify support *for* the null hypothesis via Bayes factors, with BF_{01} representing the likelihood of the null hypothesis being true relative to the alternative (values > 10 considered "strong" evidence). We observed strong support for the null hypothesis comparing the survey software and first Inquisit IATs, $\text{BF}_{01} = 17.09$, and the second Inquisit IAT, $\text{BF}_{01} = 17.04$. These data

[7] As in Study 2, the error rates for these participants were near 50% (i.e., chance): 49% for the survey software, 48% for the first Inquisit, and 49% for the second Inquisit. Thus, these responses do not represent legitimately completed IATs and were filtered out by the IAT scoring algorithm automatically.

[8] This achieved $n = 184$; using the a priori power analysis specifications, this returned a minimum of 88% power for our smallest effect size of interest, $f = .10$.

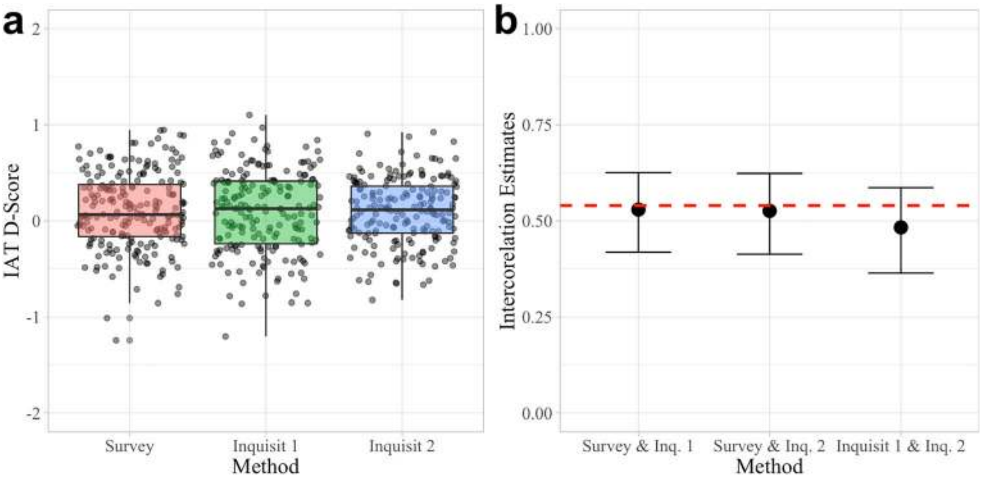


Fig. 3 (a) Boxplots and distributions of implicit association test (IAT) scores in Study 3, showing similarity across methods. (b) Correlations between IATs across methods, with the meta-analytic mean for identical IATs drawn

thus strongly indicate that the IATs produced identical or highly similar results.[9]

Correlations As expected, the survey-software IAT correlated with the first Inquisit IAT, $r(187) = .53, p < .001$, 95% CI [.42, .63], and the second Inquisit IAT, $r(183) = .53, p < .001$, 95% CI [.41, .62]. The two identical Inquisit IATs intercorrelated at similar levels: $r(182) = .48, p < .001$, 95% CI [.36, .59]. The survey-software IAT was also correlated with explicit preferences, $r(208) = .37, p < .001$, 95% CI [.24, .48]. Similar results were obtained for the first Inquisit IAT, $r(188) = .53, p < .001$, 95% CI [.42, .62], and the second Inquisit IAT, $r(184) = .40, p < .001$, 95% CI [.27, .51].[10]

Discussion

Study 3 provided evidentiary support for the validity of the survey-software IAT.

As in Studies 1–2, the survey-software IAT displayed acceptable psychometric properties (internal consistency, error rates, correlations with explicit measures, etc.). Study 3 further showed that these results were highly consistent with those produced by commercial reaction-time software. Internal consistency was similar across all IATs as well (with the survey-software IAT yielding the highest value), as were error rates, drop-trial rates, and drop-participant rates. Correlations with explicit measures were similar and consistent with meta-analytic benchmarks for consumer-preference IATs.

Most importantly, the IAT *D* scores were nearly identical. Our a priori question was whether meaningful differences would emerge between the survey-software IATs and those run via third-party reaction-time software. Despite a well-powered test to detect a trivial difference, they did not. Further supporting this finding, Bayesian analyses and equivalence tests revealed strong evidentiary support *for* the null hypothesis of no difference. Although we do not rule out the possibility that some trivial difference may exist, no support could be found for this in the present data. Despite the fact that reaction times as captured via JavaScript are somewhat slower than those captured via commercial reaction-time software (Reimers & Stewart, 2015), the IAT does not appear to be discernibly impacted. This can also be observed in our own data, with slower reaction times in the Qualtrics IAT yet highly similar IAT scores. Finally, the survey-software and Inquisit IATs were intercorrelated with each other at levels expected for identical IATs. The two identical Inquisit IATs were correlated at a similar level.

The data from Study 3 also revealed an advantage of survey-software IATs on platforms such as MTurk: asking participants to download and install the Inquisit IAT resulted in a number of participants simply skipping the Inquisit IAT, consistent with Wang-Jones et al. (2017). One might wonder whether these same individuals "skipped" the survey-software IAT by not taking it seriously, but this was not the case: skipping both Inquisit IATs was not associated with low-quality (i.e., overly fast) responses on the survey software IAT, ϕ =

[9] Priors are standard Cauchy per Rouder et al. (2009). Within a non-Bayesian framework, one can similarly test for equivalence via equivalence tests (Lakens, 2017), which returned similar results. Within thresholds of $D \pm .10$ as the smallest effect size of interest (within the "no effect" range via Project Implicit), the hypothesis of equivalence was supported for both the first Inquisit IAT, $t(183) = 3.42, p < .001$, and the second Inquisit IAT, $t(183) = 3.71, p < .001$.

[10] We also compared the implicit–explicit (I/E) correlations for the survey-software and first Inquisit IATs (which were presented in a random order). For the 189 participants who completed both IATs, the survey-software I/E correlation was $r(187) = .45, p < .001$, 95% CI [.32, .55], and the first Inquisit I/E correlation was $r(187) = .53, p < .001$, 95% CI [.42, .63]. These were compared using a z test for dependent overlapping correlations (Hittner, May, & Silver, 2003), and were not significantly different, $z = 1.41, p = .16$.

Table 3 Reaction times and error rates by block and implicit association test (IAT)

Block	Type	IAT	Q1	Median	Q3	*M* (log)	*SD* (log)	Error
B&J + Pos/Hag + Neg	Practice	Qualtrics	653.91	749.27	905.74	6.68	0.27	0.07
		Inquisit 1	543.83	607.83	757.45	6.50	0.28	0.06
		Inquisit 2	534.76	605.70	696.98	6.43	0.22	0.07
	Critical	Qualtrics	657.71	751.39	876.52	6.65	0.24	0.08
		Inquisit 1	549.05	613.33	753.33	6.47	0.24	0.07
		Inquisit 2	534.12	606.56	698.86	6.43	0.22	0.07
Hag + Pos/B&J + Neg	Practice	Qualtrics	680.14	764.05	893.85	6.70	0.26	0.08
		Inquisit 1	575.47	644.23	762.36	6.51	0.25	0.07
		Inquisit 2	555.69	622.92	723.60	6.46	0.21	0.08
	Critical	Qualtrics	695.06	788.10	893.30	6.69	0.21	0.09
		Inquisit 1	572.19	649.88	758.38	6.51	0.22	0.09
		Inquisit 2	555.32	629.35	720.49	6.47	0.21	0.08

B&J = Ben & Jerry's; Hag = Haagen Dazs; Q1 = 25th percentile; Q3 = 75th percentile; Error = error rate. *M* and *SD* calculations were performed using log-transformed reaction times due to skewness/extreme values (skewness = 1.8–2.5).

.004, $\chi(1) = .004$, $p = .95$. In other words, Inquisit-skippers were no less likely to take the survey-software IAT seriously. Consequently, moving the IAT into the survey itself should be expected to increase the amount of usable data available to researchers.

General conclusion

Survey-software IATs appear to be viable and valid. Using our procedure, it is possible to build a functional IAT without specialized reaction-time software, commandeering functions of mainstream survey software via added HTML and JavaScript. Although mainstream survey software is not intentionally designed to run IATs, we have shown that a creative survey design and added code make this a possibility. Although this procedure would be laborious if done entirely from scratch, we offer a full suite of materials (templates, editable examples, etc.), as well as a supplementary tool that rapidly automates the entire process. Across samples, domains, and variants, the survey-software IAT functioned reliably, consistently with expectations, and in accordance with meta-analytic trends, published research, and reaction-time software.

The ability to run IATs from within survey software has major implications for IAT research. As we noted earlier, online data collection—often via survey software—is increasingly mainstream in psychology. However, until now, researchers have lacked the ability to add IATs to online surveys, requiring links to external platforms that raise a number of issues. The results from Study 3 were consistent with those of Wang-Jones et al. (2017), showing that a small but consistent portion of online participants simply skip external IATs. In an unforeseen benefit of the survey-software IAT, by removing the burden of asking participants to install software and embedding the IAT within the survey, this issue was eliminated. Thus, embedding the IAT in the survey avoids substantial data loss to researchers.

Given the widespread use of online data collection in psychology, this method offers the potential to increase the already widespread use of the IAT. As we noted earlier, IAT studies are typically conducted using small, in-person samples; however, small samples are less informative, yielding imprecise parameter estimates (e.g., slopes, correlations) and have been criticized as a factor leading to issues with replicability (Szucs & Ioannidis, 2017). Although it is certainly possible to collect a small online sample (or a large in-person sample), large online samples are typically more feasible (Buhrmester et al., 2018). Given that the IAT performs well online (Nosek et al., 2002), this may help push online research beyond self-report and allow for large, adequately powered tests of researchers' hypotheses in relatively short amounts of time. It should be noted that there might be situations in which an in-person IAT is desirable. For example, researchers may wish to include correlations with observed behavior or may desire greater environmental control. However, researchers who wish to run the IAT in a laboratory can use the survey-software IAT in the lab.

One cautionary note can be drawn from comparing the results of Study 1 and Studies 2–3: although the IAT itself performed adequately with both student and crowdsource (i.e., MTurk) populations, a small proportion of MTurk participants took the IAT less seriously. Although low-quality data is a known issue in online research (Gosling & Mason, 2015), the scoring procedures for the IAT appeared to mitigate this issue. Specifically, participants who indiscriminately and rapidly press buttons to skip past the IAT are flagged by the *D*-score drop mechanism (Greenwald et al., 2003). This may be one benefit to using a tool such as the IAT, rather than an alternative tool (e.g., Nosek & Banaji, 2001) in an environment such as MTurk.

The survey-software IAT does have limitations. First, our materials are not easily modifiable, meaning that single-category IATs (Karpinski & Steinman, 2006), multidimensional IATs (e.g., Gattol, Sääksjärvi, & Carbon, 2011), and brief IATs (Sriram & Greenwald, 2009) are not possible without reworking code. If users wished to modify the procedure in some way—for example, to use images instead of text as the on-screen labels—this would require some reworking of our code. Relatedly, participation is restricted to individuals on computers with keyboards; mobile participants are presently unable to complete our survey-software IAT. Another limitation is that it is relatively easy to "mess up" the survey-software IAT (e.g., researchers could inadvertently move a survey component outside of a randomizer). However, we believe the advantages vastly outweigh the disadvantages. For example, it is also easy to inadvertently compromise an IAT run via external software (e.g., when editing an Inquisit script). One additional question for future research is participants' experiences of different IAT formats. Although we presume that participants prefer not to download software to complete a study, this is not currently known.

A final limitation is that our method is presently limited to Qualtrics, a commercial tool. Although many researchers currently use and have access to Qualtrics (e.g., through a university account), it would likely not be cost-effective for researchers to pay for a Qualtrics subscription solely to run IATs. Should any other mainstream platform allow custom HTML/JavaScript, we do intend to expand our approach to those platforms. Although we have no reason to think that our results would be any different in other platforms, our validation data speaks only to Qualtrics and thus future implementations should be empirically tested before they see widespread use. It is also possible that Qualtrics could update their technology in the future. Fortunately, our approach is relatively robust to changes, consisting primarily of generic HTML/JavaScript code that is copy/pasted into a Qualtrics survey template. Furthermore, it uses a Qualtrics-provided and supported platform designed to support such interactivity. Thus, our approach relies on features that should make it less vulnerable to updates.

Conclusion

Survey-software IATs appear to be a viable and valid way of conducting IAT research, avoiding costs to many researchers, allowing for large online sample recruitment, simplifying data management, making study completion considerably simpler for participants, and avoiding data loss to researchers. Using our provided materials (or automated tool), implementing the procedure described here will hopefully facilitate high-quality online IAT research.

Open practice statement The data and materials for all experiments are available at https://osf.io/jsezc/. No experiments were preregistered.

Appendix: IAT stimuli

Study 1: Atheist trust

Trust: Truthful, Credible, Dependable, Honest, Trustworthy

Distrust: Sneaky, Lying, Devious, Dishonest, Deceitful**Atheist:** Atheist, Non-believer

Theist: Believer, Religious

Study 2: Black–White photo

Pleasant: Joy, Happy, Laughter, Love, Friend, Pleasure, Peace, Wonderful

Unpleasant: Evil, Agony, Awful, Nasty, Terrible, Horrible, Failure, War

African American/Euro-American: morphed-face images (Nosek et al., 2007)

Study 2: Insect–flower

Pleasant/Unpleasant as above

Insects: Wasp, Flea, Roach, Centipede, Moth, Bedbug, Gnat

Flowers: Orchid, Tulip, Rose, Daffodil, Daisy, Lilac, Lily

Study 2: Mexican–US surname

Pleasant/Unpleasant as above

American Surname: Smith, Johnson, Williams, Brown, Jones, Miller, Davis

Mexican Surname: Rodríguez, Hernandez, García, Martínez, González, López, Pérez

Study 2: Gender–career

Career: Executive, Management, Professional, Corporation, Salary, Office, Business, Career

Family: Home, Parents, Children, Family, Cousins, Marriage, Wedding, Relatives

Male: John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill

Female: Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna

Study 3: Ice cream preference

Good: Good, Excellent, Superb, Best, Wonderful

Bad: Bad, Terrible, Awful, Worst, Horrible

Ben & Jerry's/Haagen Dazs: photos of ice cream containers (see the Method section)

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed J. K. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-019-01237-x

Back, M. D., Schmukle, S. C., & Egloff, B. (2005). Measuring task-switching ability in the Implicit Association Test. *Experimental Psychology*, *52*, 167–179. https://doi.org/10.1027/1618-3169.52.3.167

Bluemke, M., & Friese, M. (2006). Do features of stimuli IAT effects? *Journal of Experimental Social Psychology*, *42*, 163–176. https://doi.org/10.1016/j.jesp.2005.03.004

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology*, *38*, 977–997. https://doi.org/10.1002/ejsp.487

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*, 149–154. https://doi.org/10.1177/1745691617706516

Capozza, D., Vezzali, L., Trifiletti, E. T., Falvo, R., & Favara, I. (2010). Improving intergroup relationships within and outside the contact situation: The role of common in-group identity and emotions of empathy and anxiety. *Testing, Psychometrics, Methodology in Applied Psychology*, *17*, 17–36.

Carpenter, T. P. (2015). *Enjoy! A dual-process examination of product-pleasure associations and preferences* (Unpublished doctoral dissertation). Baylor University, Waco, TX.

Cohen, T. R., Wolf, S. T., Panter, A. T., & Insko, C. A. (2011). Introducing the GASP scale: A new measure of guilt and shame proneness. *Journal of Personality and Social Psychology*, *100*, 947. https://doi.org/10.1037/a0022641

De Houwer, J., & De Bruycker, E. (2007). The Implicit Association Test outperforms the extrinsic affective Simon task as an implicit measure of inter-individual differences in attitudes. *British Journal of Social Psychology*, *46*, 401–421. https://doi.org/10.1348/014466606X130346

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*, 1–12. https://doi.org/10.3758/s13428-014-0458-y

Duckitt, J., & Bizumic, B. (2013). Multidimensionality of right-wing authoritarian attitudes: Authoritarianism-Conservatism-Traditionalism. *Political Psychology*, *34*, 841–862.

Fatfouta, R., & Schröder-Abé, M. (2018). Agentic to the core? Facets of narcissism and positive implicit self-views in the agentic domain. *Journal of Research in Personality*, *74*, 78–82. https://doi.org/10.1016/j.jrp.2018.02.006

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225

Friese, M., Wänke, M., & Plessner, H. (2006). Implicit consumer preferences and their influence on product choice. *Psychology & Marketing*, *23*, 727–740. https://doi.org/10.1002/mar.20126

Gattol, V., Sääksjärvi, M., & Carbon, C.-C. (2011). Extending the Implicit Association Test (IAT): Assessing consumer attitudes based on multi-dimensional implicit associations. *PLoS ONE*, *6*, e15849. https://doi.org/10.1371/journal.pone.0015849

Gawronski, B. (2002). What does the implicit association test measure? A test of the convergent and discriminant validity of prejudice-related IATs. *Experimental Psychology*, *49*, 171–180. https://doi.org/10.1026/1618-3169.49.3.171

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*, 300–312. https://doi.org/10.1177/0146167216684131

Gervais, W. M., Shariff, A. F., & Norenzayan, A. (2011). Do you believe in atheists? Distrust is central to anti-atheist prejudice. *Journal of Personality and Social Psychology*, *101*, 1189–1206. https://doi.org/10.1037/a0025882

Gervais W. M., Xygalatas, D., McKay, R. T., van Elk, M., Buchtel, E. E., Aveyard, M., … Bulbulia, J. (2017). Global evidence of extreme intuitive moral prejudice against atheists. *Nature Human Behavior*, *1*, 0151.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist*, *72*, 861–871. https://doi.org/10.1037/amp0000238

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

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. https://doi.org/10.1037/0022-3514.85.2.197

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. https://doi.org/10.1037/a0015575

Hittner, J. B., May, K., & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *Journal of General Psychology*, *130*, 149–168. https://doi.org/10.1080/00221300309601282

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. https://doi.org/10.1177/0146167205275613

Instituto Nacional Electoral (2005). *Estadísticas Lista Nominal y Padrón Electoral*. Retrieved from http://portalanterior.ine.mx/archivos3/portal/historico/contenido/Estadisticas_Lista_Nominal_y_Padron_Electoral/

Karpinski, A., & Steinman, R. B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, *91*, 16–32. https://doi.org/10.1037/0022-3514.91.1.16

Kirk, R. E. (2008). *Statistics: An introduction* (5th ed.). Belmont, CA: Thomson Wadsworth.

Kuchenbrandt, D., Eyssel, F., & Seidel, S. K. (2013). Cooperation makes it happen: Imagined intergroup cooperation enhances the positive effects of imagined contact. *Group Processes and Intergroup Relations*, *16*, 635–647. https://doi.org/10.1177/1368430212470172

LaBouff, J. P., & Ledoux, A. M. (2016). Imagining atheists: Reducing fundamental distrust in atheist intergroup attitudes. *Psychology of Religion and Spirituality*, *8*, 330–340. https://doi.org/10.1037/rel0000066

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*, 355–362. https://doi.org/10.1177/1948550617697177

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: IV. What we know (so far) about the method. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 59–102). New York, NY: Guilford Press.

Marquardt, N., & Hoeger, R. (2009). The effect of implicit moral attitudes on managerial decision-making: An implicit social cognition approach. *Journal of Business Ethics*, *85*, 157–171. https://doi.org/10.1007/s10551-008-9754-8

Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind: implicit cognition predicts suicidal behavior. *Psychological Science*, *21*, 511–517. https://doi.org/10.1177/0956797610364762

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

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*, 101–115. https://doi.org/10.1037/1089-2699.6.1.101

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the implicit association test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31*, 166–180. https://doi.org/10.1177/0146167204271418

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36–88. https://doi.org/10.1080/10463280701489053.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184–188. https://doi.org/10.1177/0963721414531598

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*, 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Penke, L., Eichstaedt, J., & Asendorpf, J. B. (2006). Single-Attribute Implicit Association Tests (SA-IAT) for the assessment of unipolar constructs: The case of sociosexuality. *Experimental Psychology*, *53*, 283–291. https://doi.org/10.1027/1618-3169.53.4.283

Pittinsky, T. L., Rosenthal, S. A., & Montoya, R. M. (2011). Liking is not the opposite of disliking: The functional separability of positive and negative attitudes toward minority groups. *Cultural Diversity and Ethnic Minority Psychology*, *17*, 134–143. https://doi.org/10.1037/a0023806

Pogacar, R., Kouril, M., Carpenter, T. P., & Kellaris, J. J. (2018). Implicit and explicit preferences for brand name sounds. *Marketing Letters*, *29*, 241–259. https://doi.org/10.1007/s11002-018-9456-7

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*, 309–327. https://doi.org/10.3758/s13428-014-0471-1

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. https://doi.org/10.3758/PBR.16.2.225

Rudman, L. A. (2011). *Implicit measures for social and personality psychology.* Thousand Oaks, CA: Sage.

Schnabel, K., Asendorpf, J. B., & Greenwald, A. G. (2008). Using Implicit Association Tests for the Assessment of Implicit Personality Self-Concept. In *SAGE handbook of personality theory and assessment: Vol. 2. Personality measurement and testing* (pp. 508–528). https://doi.org/10.4135/9781849200479.n24

Schnabel, K., Banse, R., & Asendorpf, J. B. (2006). Assessment of implicit personality self-concept using the Implicit Association Test (IAT): Concurrent assessment of anxiousness and angriness. *British Journal of Social Psychology*, *45*, 373–396. https://doi.org/10.1348/014466605X49159

Smith, A. G., & Winter, D. G. (2002). Right-wing authoritarianism, party identification, and attitudes toward feminism in student evaluations of the Clinton–Lewinsky story. *Political Psychology*, *23*, 355–383. https://doi.org/10.1111/0162-895X.00285

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology*, *56*, 283–294.

Stephan, W. G., Boniecki, K. A., Ybarra, O., Bettencourt, A., Ervin, K. S., Jackson, L. A., … Renfro, C. L. (2002). The role of threats in the racial attitudes of Blacks and White. *Personality and Social Psychology Bulletin*, *28*, 1242–1254. https://doi.org/10.1177/01461672022812009

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*, e2000797. https://doi.org/10.1371/journal.pbio.2000797

United States Census Bureau. (2010). *Frequently occurring surnames in the 2010 census.* Retrieved from https://www.census.gov/library/visualizations/2016/comm/cb16-tps154_surnames_top15.html

Vezzali, L., Capozza, D., Giovannini, D., & Stathi, S. (2012). Improving implicit and explicit intergroup attitudes using imagined contact: An experimental intervention with elementary school children. *Group Processes & Intergroup Relations*, *15*, 203–212. https://doi.org/10.1177/1368430211424920

Wang-Jones, T. S., Alhassoon, O. M., Hattrup, K., Ferdman, B. M., & Lowman, R. L. (2017). Development of gender identity implicit association tests to assess attitudes toward transmen and transwomen. *Psychology of Sexual Orientation and Gender Diversity*, *4*, 169–183. https://doi.org/10.1037/sgd0000218