Implicit Cross-Community Biases Revisited: Evidence for Ingroup Favoritism in the Absence of Outgroup Derogation in Northern Ireland

Sean Hughes and Dermot Barnes-Holmes

Ghent University

Sinead Smyth

Dublin College University

Author Note

Sean Hughes and Dermot Barnes-Holmes, Department of Experimental Clinical and Health Psychology, Ghent University, Belgium. Sinead Smyth, School of Nursing and Human Sciences, Dublin City University. The second author was supported by an Odysseus (Type 1) award from the Flanders Science Foundation (FWO) during preparation of this article. Electronic mail should be sent to sean.hughes@ugent.be. This paper is dedicated the memory of Ed Cairns, who inspired and facilitated this particular line of research.

Abstract

Despite their application in virtually every area of psychological science indirect procedures have rarely been used to study how Catholic and Protestants automatically respond to one another in Northern Ireland. What little evidence that does exist suggests that automatic ingroup favoritism occurs alongside outgroup derogation. That is, Catholics and Protestants automatically evaluate ingroup members more positively than outgroup members and also evaluate outgroup members more negatively than ingroup members. The current study addresses a methodological limitation in this early work and provides the first (non-relativistic) assessment of intergroup relational responding in a post-conflict setting using the Implicit Relational Assessment Procedure (IRAP). Contrary to earlier findings, participants displayed evidence of ingroup favoritism in the absence of outgroup derogation.

Keywords: implicit cognition, Northern Ireland, IRAP

Implicit Cross-Community Attitudes Revisited: Evidence for Ingroup Favoritism in the Absence of Outgroup Derogation in Northern Ireland

Over the course of three decades (1969–1999), approximately 3,700 people lost their lives and a further 35,000 were injured as a result of sectarian violence in Northern Ireland (McKittrick, Kelters, Feeney, Thornton, & McVea, 2007). This conflict (known as "The Troubles") stemmed from a complex interplay of historical, political, social/ethnic, economic, and psychological factors that largely played out across religious lines (see Cairns & Darby, 1998; McAlister, Scraton, & Haydon, 2009). Nevertheless, and despite its turbulent past, Northern Ireland has recently witnessed an end to large scale sectarian violence along with positive developments on economic and social fronts.

However, the conflict's legacy lingers. Catholics and Protestants still live in largely segregated residential areas (Shirlow & Murtagh, 2006), have limited social or personal contact (McAlister et al., 2009), are typically enrolled in non-integrated schools (NICIE, 2014), and partake in cultural events that are difficult to access by outgroup members. Researchers interested in studying segregation's consequences on cross-community attitudes (Hughes, Campbell, Lolliot, Hewstone, & Gallagher, 2013), intergroup contact (Turner, Tam, Hewstone, Kenworthy, & Cairns, 2013), as well as forgiveness, trust, and reconciliation (Hewstone et al., 2014) have overwhelmingly relied upon questionnaires and interviews. Although these *direct* procedures provide insight into the behavior of interest, they are deployed under the assumption that people not only have introspective access, but also the opportunity and motivation to accurately report on their psychological attributes or content. Yet research shows time and again that this assumption is often violated in socially sensitive situations, demand prone domains, or

instances in which the individual lacks introspective accessibility to the content under investigation (e.g., Gawronski & Payne, 2010; Nisbett & Wilson, 1977).

Implicit Cognition

These methodological shortcomings have recently spurred the development and use of *indirect* procedures (see Gawronski & De Houwer, 2014). At their core, these tasks attempt to circumvent a person's ability to strategically control his or her behavior as well as capture "automatic" responses that are emitted quickly, without intention and/or awareness. Although these responses unfold in the blink of an eye, they influence people's social perception, judgment, and actions, from their likelihood of attempting suicide (Nock et al., 2010), or breaking up with their romantic partner (Lee, Rogge, & Reis, 2010), to the quality and quantity of their interactions with other racial (McConnell & Leibold, 2001) or ethnic group members (Rooth, 2010).

Despite their application in many areas of psychological science, only a single study has used indirect procedures to examine "automatically" emitted intergroup responses in Northern Ireland. In their study, Tam et al. (2008) asked a group of Catholic and Protestant students to report how positively or negatively they felt towards members of their own or the other community using a feeling thermometer. Students were then asked to complete two Implicit Association Tests (IAT; Greenwald, McGhee, & Schwartz, 1998) wherein valenced adjectives had to be categorized with Catholic and Protestant names during one task (Name IAT) and images of Catholic or Protestant sectarian groups during a second task (Sectarian IAT). The authors found that, during the Name IAT, Catholics were quicker to relate Catholic (compared to Protestant) names and positive adjectives whereas Protestants were quicker to relate Protestant (compared to Catholic) names and positive adjectives. Although Catholic and Protestant students

also produced a similar set of response biases on the sectarian IAT, Catholic students showed even larger IAT effects favoring their (sectarian) ingroup than their Protestant counterparts (they also showed larger self-reported effects favoring their ingroup than Protestant students).

Critically, however, the IAT—by its very design—cannot disentangle the various ways in which the members of different groups "automatically" relationally respond to one another. For instance, an increased probability of automatically categorizing one's ingroup with positive adjectives does not imply that the same person will also categorize outgroup exemplars with negative adjectives with equal speed. Likewise, the degree to which people automatically categorize ingroup exemplars and negative adjectives using the same key may differ from the extent to which they categorize outgroup members and positive stimuli using the same key.

To circumvent this issue, Tam et al. asked the same participants to complete a go/no-go-association task (GNAT; Nosek & Banaji, 2001). During the GNAT, participants encountered four different types of trials (or "trial types") that either presented Catholic or Protestant names along with positive or negative adjectives (*Catholics-Good*; *Catholics-Bad*; *Protestants-Good*; *Protestants-Bad*). In each case, they were instructed to respond quickly to items that fell into one of the two stimulus categories presented onscreen (i.e., to press the space bar whenever they saw exemplars from one of the two categories and press nothing whenever they saw exemplars from the other two categories). The authors found that Catholics produced a larger GNAT effect than Protestants on the *Catholics-Good* trial type, whereas Protestants produced a larger GNAT effect than Catholics on the *Protestants-Good* trial type (a pattern of findings they referred to as ingroup favoritism). At the same time, Protestants also produced a larger effect than Catholics on the *Catholics-Bad* trial type, and Catholics produced a larger effect than Protestants on the *Protestants-Bad* trial type (a pattern of findings they referred to as outgroup derogation). Based

on these findings, the authors concluded that both groups showed evidence of ingroup favoritism *and* outgroup derogation at the implicit level.

On the surface, these findings appear to clearly distinguish between four different patterns of intergroup relational responding. Yet upon closer inspection, we believe that this may not be the case. This is due to the fact that the GNAT is non-relativistic as a procedure but relativistic as an effect. When we state that the GNAT is non-relativistic as a procedure, we are highlighting the fact that the procedure allows participants to respond to one class of stimuli (Catholics) independently of another class of stimuli (Protestants). Specifically, although label stimuli representing both classes are presented onscreen during each IAT trial, this is not the case with the GNAT, where only a Catholic *or* Protestant label stimulus is presented onscreen at any one time (along with another valenced label stimulus). Thus on each trial, participants have to press the space bar whenever a stimulus that appears in the middle of the screen (e.g., "Protestants," "Catholics," "bad," or "good") belongs to one of the two stimulus categories that appear at the top of the screen (e.g., "Catholics" or "good"). In effect, the response that is emitted on one trial (e.g., categorizing the word Catholics with the label "Catholics" or positively valenced words with the label "good") is independent from responses that are emitted on another trial (e.g., categorizing the word Protestants with the label "Protestants" or positive words with the label "good").

Although the GNAT allows participants to respond to Catholic and positive terms independently of Protestant and positive (or negative) terms, the scores obtained from the task are *relativistic* in nature—just like the IAT. This is because the typical GNAT (d') effect involves (amongst other things) subtracting reaction times from responses in the presence of target stimuli (e.g., "Catholics" or "good") from reaction times from responses in the presence of distractor

stimuli (e.g., “Protestants” or “bad”). As such, the outcome reflects how quickly people respond to “Catholics” and “good” relative to “Protestants” and “bad” and therefore does not speak to how people independently evaluate these two groups. In short, while the GNAT is non-relativistic as a procedure (allowing for participants to respond to “Catholics” independently of “Protestants”), it is relativistic as an effect (insofar as the outcome is generated by comparing responses to one group against another). This latter issue compromises its ability to disentangle one pattern of “automatic” relational responding from another. Indeed, if we are to acquire a more sophisticated appreciation for intergroup verbal relations in a post-conflict setting (such as Northern Ireland), then a task is needed that is non-relativistic at the procedural and effect levels (e.g., one that can help us learn to what extent Catholics automatically evaluative themselves as positive or negative *separately* from how they evaluative Protestants and vice-versa). The Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010) represents one such task.

The Current Research

Unlike the IAT or GNAT, the IRAP does not require participants to simply categorize stimuli with one another. Instead, it was designed to assess the speed and accuracy with which pre-existing or experimentally induced (arbitrarily applicable) relational responses (of varying complexity) are quickly emitted. It does so by placing an individual’s learning history into competition with a response contingency deemed inconsistent with that history of responding. During each trial, one stimulus is presented at the top of the screen (e.g., “Catholics” or “Protestants”), along with a second stimulus in the middle of the screen (e.g., positive or negative trait descriptions) and two response options at the bottom of the screen (“True” or “False”). By presenting specific combinations of stimuli together on each trial, and by requiring a

particular response to be emitted quickly and accurately, the IRAP can capture how people "automatically" respond when presented with four different stimulus relations (e.g., *Catholics-Good*; *Catholics-Bad*; *Protestants-Good*; *Protestants-Bad*). These trials are blocked, and just like the IAT, there are two different (repeatedly presented) types of test blocks. During one type of test block, participants are required to select "True" on trials that present "Catholics" and positive adjectives or "Protestants" and negative adjectives onscreen, and "False" on trials that present "Catholics" and negative adjectives or "Protestants" and positive adjectives. During a second type of block, participants are required to respond in the opposite way. The difference in time taken to affirm a stimulus relation in one block versus reject it in another block (defined as an IRAP effect) indicates (broadly speaking) the strength or probability with which those stimuli are related by participants on average.

Critically, a separate IRAP effect is calculated for each of the four stimulus relations (or trial types) so that a non-relative index of the four relational responses can be obtained. Thus, while the IRAP and GNAT are both comprised of four different types of trials, only the former provides an independently interpretable score for those trials. Put another way, it can determine whether participants are quicker to respond to Catholic exemplars as positive (or negative) separately from their responses to Protestant exemplars as positive (or negative). This in turn allows researchers to disentangle these four different patterns of relational responding, and in so doing, provide a more nuanced perspective on how two groups in a post-conflict setting automatically respond to one another (along an evaluative dimension)[1].

[1] Let us be clear from the outset. Stating that IRAP effects are *non-relative* is not the same as saying that they are *a-contextual*. Non-relative denotes that the effect itself is calculated in a way that is independent from other trial types (i.e., what we say about the *Catholics-Good* trial type is inferred from the speed with which people affirm versus reject the *Catholic-Good* relation and does not depend on how quickly they responded to "Catholics" and negative terms or "Protestants" and positive or negative terms). Nevertheless, this does not mean non-relativistic trial-type effects are a-contextual. In other words, the relational response on any given trial, and thus the effects calculated from those responses, could be moderated by contextual variables that are part of the IRAP or the wider context in

Drawing on both the IRAP and aforementioned work, we set out to provide the first non-relativistic account of automatic intergroup (relational) responding in a post-conflict setting (Northern Ireland). If Tam et al. (2008) are correct and students do show evidence of automatic ingroup favoritism as well as outgroup derogation, then we would expect two distinct response patterns to emerge from the IRAP. On the one hand, they should be quicker to select "True" than "False" when presented with an ingroup exemplar (e.g., the word "Catholics" for Catholic students) and positive adjectives. They should also select "False" more quickly than "True" when presented with an ingroup exemplar and negative adjectives. We label such a pattern of responding "ingroup favoritism." On the other hand, students should select "True" more quickly than "False" whenever they are presented with an outgroup exemplar and negative adjectives and select "False" more quickly than "True" when they encounter outgroup exemplars and positive adjectives. We label this latter response pattern as "outgroup derogation." Finally, we would expect Catholic students to explicitly evaluate themselves more positively (and less negatively) than Protestants while an opposite pattern of responding should hold for the latter group[2].

Method

Participants

Sixty-nine students (47 women) from a university in Northern Ireland participated in the current study in exchange for course credit. Forty-three identified themselves as belonging to the Protestant community, 24 identified themselves as belonging to the Catholic community, and two did not identify themselves as belonging to either community. The group ranged from 18 to 47

---

which the IRAP is embedded (for more on non-relative vs. a-contextual, see Hussey et al., 2016). We will return to this issue in greater detail later on in the General Discussion.

[2] The terms "ingroup favoritism" and "outgroup derogation" are often used in the social psychological literature to refer to a set of mental concepts and processes (see Dasgupta, 2004). Although we will also use those same terms in this paper, we make no appeals to, or assumptions about, those mental mechanisms. Instead we simply use these terms to orient the reader towards specific patterns of behavior (relational responses) that are emitted by members of different groups (Catholics and Protestants) in the presence of certain stimuli (ingroup vs. outgroup exemplars and trait descriptions).

years (M = 21.1, SD = 6.6) in age, and the vast majority (97%) reported that they had spent their entire lives in Northern Ireland.

Measures

Self-report measures. Students were administered a series of feeling thermometers and asked to indicate the degree to which they felt "cold" or "warm" towards Catholics or Protestants on a scale from 0° (*Negative Feelings*) to 100° (*Positive Feelings*), with 50° as a neutral point. In order for a pattern of behavior to be labeled as "explicit ingroup favoritism," ingroup ratings had to meet two criteria: They had to significantly differ from the 50° neutral point (in a positive direction) and significantly differ from ratings of the outgroup. Likewise, in order for a pattern of behavior to be labeled as "outgroup derogation,", outgroup ratings had to significantly differ from the 50° neutral point (in a negative direction) and significantly differ from ratings of the ingroup.

IRAP. "Automatic" relational responding was indexed using an IRAP. The task consisted of a minimum of two and a maximum of six practice blocks followed by a fixed set of (six) test blocks. Each block consisted of 24 trials that presented one of two label stimuli ("Catholics*'*" or "Protestants*'*") in the presence of one of six positive (*good, honest, nice, peaceful, friendly* and *safe*) or negative (*bad, dishonest, nasty, violent, aggressive* and *hostile*) target stimuli and required participants to emit one of two relational responses ("True" or "False"'). In this way, the IRAP was comprised of four different trial types: *Catholics-Good; Catholics-Bad; Protestants-Good; Protestants-Bad* (see Figure 1). The location of the response options was randomized across trials and the trials themselves were presented in a quasi-random order so that each of the four trial types appeared six times within a given block in a non-sequential manner.

Prior to the IRAP, participants were informed that they would complete a word categorization task that would present two different words on the screen (either the word "Catholics" or "Protestants" with a positive or negative adjective) and that they would have to respond to those words as using the "True" and "False" response options. Their task was to respond as quickly and accurately as possible during each block of trials. Visual illustrations of the four IRAP trial types were then presented, any remaining questions answered, and the practice phase initiated.

The practice phase consisted of two types of blocks: "pro-Catholic" and "pro-Protestant." During a "pro-Catholic" practice block, participants had to affirm the relation between "Catholics" and "good" or "Protestants" and "bad" as well as reject the relation between "Catholics" and "bad" or "Protestants" and "good." Stated more precisely, a correct response required participants to select "True" when "Catholics" appeared with a positive stimulus (e.g., "nice") or when "Protestants" appeared with a negative stimulus (e.g., "nasty"). At the same time, participants were also required to choose "False" when "Catholics" appeared with a negative word or when "Protestants" appeared with a positive word. Precisely the opposite pattern of responding was required during a "pro-Protestant" block. In order to clarify that the programmed contingencies would now reverse, onscreen instructions highlighted this to the participant after each block of trials (i.e., "During the next phase, the previously correct and wrong answers are reversed. This is part of the experiment. Please try to make as few errors as possible—in other words, avoid the red "X"). In both types of blocks, selecting the response option deemed correct removed all stimuli from the screen for a 400-ms inter-trial interval, after which the next trial was presented. If an incorrect response was emitted, a large red "X" appeared on screen directly below the target stimulus. The red "X" and all other stimuli remained

on the screen until a correct response was emitted, after which the screen cleared and the program progressed to the inter-trial interval.

The IRAP commenced with a pair of practice blocks that acquainted participants with the general task requirements. Progression from the practice to the test phase was made contingent upon highly accurate (at least 85% accuracy) and quick responding (median latency of less than 2000 ms) on a successive pair of practice blocks. In order to make these dual requirements clear to the participant, a feedback screen was presented whenever a person failed to achieve one or both mastery criteria during two consecutive practice blocks. This screen stated the criteria needed to complete the practice phase and presented accuracy and latency scores for the previous two blocks. Participants were then re-exposed to another pair of practice blocks until they either achieved the mastery criteria or a maximum of three pairs of practice blocks were completed. Failure to meet these criteria resulted in the participants being thanked, debriefed, and dismissed. Once the criteria were met, a fixed set of three pairs of test blocks were administered. The test blocks were similar to the practice blocks (i.e., a red "X" appeared when an error was made and participants were informed about their speed and accuracy between blocks) with two exceptions. First, and unlike during the practice phase, there was no performance criteria for progression from block to block. When they failed to maintain the mastery criteria, no feedback screen appeared between blocks requiring them to repeat the previous pair of blocks. Second, a new message appeared before each test block informing participants that "This is a test—go fast, making a few errors is okay".

Procedure

Upon arriving at the laboratory, participants were welcomed by the researcher, asked to sign statements of informed consent, and seated in front of a computer from which they received

all instructions. They were then informed that they would complete a questionnaire as well as computer based task and—given the sensitive nature of the study—that they would be randomly assigned an identification number in order to preserve their confidentiality and anonymity. Thereafter, participants completed an IRAP followed by the self-report task. Overall, the experiment lasted approximately 40 minutes.

Results

Data Preparation

Participant exclusion. The data of six participants were removed for the following reasons: two participants did not self-identify as members of the Catholic or Protestant communities and four individuals failed to achieve or maintain the IRAP mastery criteria across two or more pairs of IRAP blocks. This left a final sample of 63 participants (22 Catholics and 41 Protestants). Note that in-line with previous work (e.g., Nicholson & Barnes-Holmes, 2012), whenever one of these participants failed to maintain accuracy (79%) or latency (2000 ms) criteria on one of the six test blocks, all the data from that test block pair were excluded and analyses were conducted on the remaining two test block pairs. This was the case for eight participants.

IRAP. The primary data obtained from the IRAP was response latency, defined as the time in milliseconds (ms) that elapsed from the onset of each IRAP trial to the first correct response emitted by the participant. Responses latencies were included from trials on which a correct or incorrect response was emitted. To minimize contamination by individual differences associated with age, motor skills, and/or cognitive ability, response latencies were transformed into D-*IRAP* scores using an adaptation of Greenwald, Nosek, and Banaji's (2003) D algorithm (for more on this specific transformation see Appendix B). Four D-*IRAP* scores were calculated

for each participant, one for each of the IRAP trial types (i.e., *Catholics-Good*; *Catholics-Bad*; *Protestants-Good*; *Protestants-Bad*) and each score could theoretically range from -2 to +2. The scores from the two Protestant trial types were reverse scored (multiplied by -1) in order to facilitate interpretation of the data (for more on trial-type inversion and its rationale see Hussey, Thompson, McEnteggart, Barnes-Holmes, & Barnes-Holmes, 2015). Positive values indicate that participants were quicker to affirm the relation between group exemplars and positive stimuli or reject the relation between those same exemplars and negative stimuli. Negative scores indicate the opposite response pattern.

Hypothesis Testing

IRAP. Submitting IRAP scores to a 2 (*Community Background*) × 4 (*Trial Type*) mixed-model ANOVA revealed a main effect for Trial Type, $F(3, 61) = 19.55$, $p < .001$, $\eta^2 = 0.23$, as well as a two-way interaction between Trial Type and Community Background, $F(3, 61) = 2.77$, $p < .05$, $\eta^2 = .04$. With respect to the *Catholics-Good* trial type, Catholic students ($M = 0.46$, $SD = 0.33$) were significantly quicker to respond "True" than "False" in the presence of "Catholics" and positive adjectives, $t(21) = 6.54$, $p < .001$, $d = 1.39$. This was also the case for their Protestant counterparts ($M = 0.24$, $SD = 0.39$), $t(40) = 3.94$, $p < .001$, $d = 0.62$. It should be noted that Catholic students also affirmed the *Catholics-Good* relation to a significantly greater extent than their Protestant counterparts, $F(1, 62) = 4.98$, $p = .03$, $\eta^2 = .08$. With respect to the *Protestants-Good* trial type, Catholic students ($M = 0.27$, $SD = 0.39$) were significantly quicker to endorse the relation between Protestants and positive terms than reject it, $t(21) = 3.26$, $p = .004$, $d = 0.69$. This was also the case for Protestant students ($M = 0.42$, $SD = 0.37$), $t(40) = 7.38$, $p < .001$, $d = 1.1$. Interestingly, however, Protestants did not affirm this relation to a greater extent than their Catholic counterparts, $F(1, 62) = 2.44$, $p = .12$, $\eta^2 = .04$. No IRAP effects

emerged on the *Catholics-Bad* or *Protestants-Bad* trial types, suggesting that both groups affirmed and negated the relation between Catholics/Protestants and negative terms with equal ease (all *ps* > .45). Finally, although Catholics displayed an automatic positivity bias towards Catholics and Protestants, the effect on the *Catholics-Good* trial type was (marginally) larger than that on the *Protestants-Good* trial type, $t(21) = 2.01$, $p = .06$, $d = 0.43$. Likewise, although Protestants displayed an automatic positivity bias towards Catholics and Protestants, the effect on the *Protestants-Good* trial type was significantly larger than that on the *Catholics-Good* trial type, $t(40) = 2.34$, $p < .03$, $d = .36$ (see Figure 2).

Self-report measures. Eight (Protestant) participants opted not to provide explicit group ratings. Submitting the remaining data from the feeling thermometers to a 2 (*Community Background*) × 2 (*Evaluation*; Catholics vs. Protestants) repeated measures ANOVA revealed no main or interaction effects for either variable (all *ps* > .25). One-sample *t*-tests were used to compare ratings to the (neutral) mid-point of the scale. It appears that Catholic students responded (marginally) positively when asked to evaluate Catholics ($M = 59.52$, $SD = 22.02$), $t(20) = 1.98$, $p = .06$, and Protestants ($M = 56.82$, $SD = 17.29$), $t(21) = 1.85$, $p = .08$. Their Protestant counterparts responded positively towards Protestants ($M = 61.21$, $SD = 22.88$), $t(32) = 2.82$, $p < .01$, and neutrally towards Catholics ($M = 54.84$, $SD = 18.42$), $t(30) = 1.46$, $p = .15$. Ingroup ratings did not significantly differ from outgroup ratings for either Catholic, $t(20) = 0.56$, $p = .58$, or Protestant students, $t(29) = 1.13$, $p = .27$ (see Figure 3).

Discussion

The current paper sought to provide the first non-relativistic assessment of automatic intergroup (relational) responding in a post-conflict setting (Northern Ireland). Specifically, we were interested in the extent to which Catholics and Protestants automatically endorsed the

relation between ingroup members and positive stimuli or rejected the relation between ingroup members and negative stimuli (a pattern of behavior referred to as "ingroup favoritism"). We were also interested in whether those same groups would endorse the relation between outgroup members and negative stimuli or reject the relation between outgroup members and positive stimuli (a pattern of behavior referred to as "outgroup derogation"). Only a single study has examined "automatic" intergroup relational responding in Northern Ireland, and the authors reported that Catholics and Protestants showed both patterns of behavior (Tam et al., 2008). For reasons we discussed in the Introduction, this claim may have been premature given that the procedures they employed were unable to clearly distinguish between different relational responses from one another. When we used an IRAP to circumvent these issues, a different picture emerged. Consistent with Tam et al.'s original idea of automatic ingroup favoritism, Catholics and Protestants both produced significant IRAP effects on the *Catholics-Good* and *Protestants-Good* trial types, and each group produced larger effects on their ingroup compared to outgroup trial types (i.e., Catholics produced a larger effect on the *Catholics-Good* than the *Protestants-Good* trial type, whereas the opposite was true for Protestants). Critically, however, both groups showed no effects on the *Catholics-Bad* or *Protestants-Bad* trial types. Thus, unlike Tam et al., we found no evidence supporting the idea of outgroup derogation on either explicit or implicit measures.

Open Questions

Several possible explanations for the above findings present themselves. One explanation is that perhaps students exerted ("strategic control") over their IRAP performance in ways that they could not achieve on either the IAT or GNAT (i.e., their performance on the IRAP was not solely the product of responses to stimuli on the screen but was also influenced by verbal rules

such as "I should present myself as a tolerant non-prejudiced person"). This would explain the absence of IRAP effects on the *Catholics-Bad* and *Protestants-Bad* trial types. We believe this explanation is problematic for several reasons. First, both groups still produced larger IRAP effects whenever they had to evaluate their ingroup compared to their outgroup. If their IRAP performances were under the control of verbal rules like those mentioned above, then such an effect should not have emerged (presumably they would have produced similar effects on both trial types). Second, evidence indicates that although people can strategically influence their IRAP performances when motivated to do so, their ability to do so is rather unsophisticated. For instance, Hughes et al. (2016) found that participants could only manipulate individual trial-type performances when they were given explicit and repeated instructions on how to do so before each block of trials. When simply asked to fake their performances (in the absence of a faking strategy), participants could not influence their IRAP effects. When given some limited information on how to fake (i.e., by paying attention to their response speeds during the task) they were partially successful in reversing their effects (also see Drake, Seymour, & Habib, 2016). Thus, it seems reasonable to assume that the IRAP did indeed capture automatic relational responses towards members of the participants' in- and outgroups in a way that was uncontaminated by verbal rules related to self-presentation. Nevertheless, given the preliminary nature of this study, and the highly sensitive domain under investigation, we cannot rule out a "strategic control" explanation entirely. For instance, it could be that the participants exclusively attempted to influence their performances on those trial types that would lead them to appear prejudiced (e.g., the negativity trial types). With this in mind, future work could replicate the current study while including similar faking manipulations to those recently implemented by Hughes et al. (2016) or Drake et al. (2016).

Another possible explanation is that the vast majority of students who participated in this study grew up and were enculturated in "post-conflict" Northern Ireland. Although educational, residential, and social segregation still represents the norm, it may be that the current pattern of relational responses are in part reflective of the extraordinary changes in policing, governance, and shifting identity that has reshaped life in Northern Ireland over the last two decades. It may be that when students have to discuss (cross-community) issues that they would prefer to avoid or "put behind them," they self-report neutrality towards Catholics and Protestants. Yet when they are forced to respond under time pressure, evidence of automatic ingroup favoritism emerges in the absence of outgroup derogation.

However, we should be cautious in generalizing these findings, and, in particular, concluding that certain patterns of relational responding (automatic outgroup derogation) are entirely absent in a Northern Irish context. Indeed, before such a position can be adopted, our initial work needs to be both replicated and extended. For instance, directly comparing the IRAP effect with other implicit measures of beliefs (e.g., the Relational Response Task; De Houwer, Heider, Spruyt, Roets, & Hughes, 2015) may help us to determine if the current findings are due to specific properties of the task used or constitute changes in automatic relational responses across time in Northern Ireland. Testing a larger representative cross-section of the Northern Irish population would also allow us to generalize our findings from a university context to other parts of that society. Also recall that in Tam et al.'s (2008) study, participants completed two different IATs: one assessing automatic evaluations of Catholic and Protestant names and a second assessing responses to Catholic and Protestant sectarian groups. It may be that performance on the sectarian IAT negatively primed or influenced how students responded during the name IAT, especially given that political, social, and national identities are

interwoven in Northern Ireland. Researchers could systematically investigate what impact being primed with political, social, cultural and religious exemplars has on self-reported and automatic intergroup relational responses.

At a more general level, it seems important to acknowledge that our understanding of the precise behavioral processes or dynamics involved in IRAP performances remains unknown (see Barnes-Holmes, Barnes-Holmes, Hussey, & Luciano, 2016, for a recent discussion). Indeed, recent evidence suggests that procedural variables, such as the type of instructions provided before each block of trials (Finn, Barnes-Holmes, Hussey, & Graddy, 2016; O'Shea, Watson, & Brown, 2016) or the type of response options employed within the IRAP (Maloney & Barnes-Holmes, 2016), may impact the size and/or direction of IRAP effects. Recent research has also indicated that although the IRAP effect may be considered non-relative, it should not be considered a-contextual.

As we outlined in the Introduction, the term "non-relative" denotes that the IRAP effect is calculated such that each trial type is measured independently from other trial types (i.e., what we say about the *Protestant-Bad* trial type is inferred from the speed with which people affirm versus reject the *Protestant-Bad* relation only and does not depend on how quickly they respond to "Protestants" and positive terms or "Catholics" and positive or negative terms). However, the relational response on any given trial, and thus the effects calculated from those responses, *could* be moderated by contextual variables that are part of the IRAP or the wider context in which the IRAP is embedded. This is an empirical possibility that has several implications for the interpretation of our findings. First, contextual variables may have influenced the trial-type effects reported here. However, we still know relatively little about how or when contextual variables influence IRAP performances. For instance, in a recent study, Hussey et al. (2016)

administered two IRAPs to participants: one in which women and men were related as being either human or objects and another in which women and inanimate objects were related as being either human or objects. They found that the contrast category (men or objects) influenced performance on one women trial type but not another. In other words, the above study shows how one contextual variable (contrast category in the IRAP) can (but does not always) influence trial type performances. Whether this was also the case in our study is certainly worth considering. For instance, future work could test this by altering the nature of the contrast category (e.g., by comparing "Catholics" or "Protestants" to some neutral stimulus class [like nonsense words or random objects] or some other unrelated social group such as Hindus or Aboriginals). However, this approach runs into the immediate problem that by doing so, we capture a fundamentally different class of behaviors than those we set out to capture (i.e., how Catholics/Protestants feel about themselves in the context of Hindus is an entirely different question as to how Catholics feel about themselves in the context of Protestants). Indeed, we cannot see how one can measure ingroup favoritism and outgroup derogation in Northern Ireland without comparing one group to another, and in this context, there are only two logical groups to compare.

Second, the extent to which the influence of contextual variables may be seen as a potential strength or weakness of the IRAP appears to depend upon assumptions concerning exactly what the procedure is designed to capture. If one starts from the position that the procedure should be able to capture "absolute" or "pure" stimulus relations that are uncontaminated by contextual variables, then they are adopting an approach that deviates from the contextual behavioral science perspective we adopt here. If, however, one instead acknowledges that relational responses are always going to be subject to some form of contextual

control, then one can go about identifying such factors and either augment or diminish their control over the behavior of interest (see Bast, Barnes-Holmes, & Barnes-Holmes, 2015).

Future Directions

The current work draws attention to a relatively untapped research area that has yet to be seriously mined. Given the wealth of evidence indicating that performance on indirect procedures often predicts behavior better than that obtained from direct procedures (Greenwald, Poehlman, Uhlmann, & Banaji, 2009), it follows that the use of the former tasks may also unlock a more sophisticated understanding of intergroup relational responding in conflict and post-conflict societies such as Northern Ireland, Israel, Sierra Leone, and South Africa. When conducting this work, several points should be noted. First, recall that procedures such as the IAT or GNAT were designed to simply assess whether one set of concepts is categorized with a second set of concepts without regard to the way in which those stimuli are related. The IRAP, however, can assess the subtle ways in which people automatically relate stimuli with one another at increasing levels of complexity (e.g., "I want to believe Catholics are good"). This ability to disentangle the ways in which people automatically relate stimuli may further increase the predictive power of implicit measures and allow us to arrive at a more nuanced understanding of intergroup behaviors. For instance, it may be that the extent to which people automatically endorse relational responses concerning trust ("I do trust the other community" versus "I should trust the other community") predicts the frequency, duration and quality of their outgroup contact, such as their seating arrangements during class or their number of cross-group friends. Likewise, the extent to which participants automatically experience intergroup anxiety ("I'm afraid of being alone with Catholics"), empathy ("I'd help a Protestant if they were in trouble"), and forgiveness ("I can forgive the other community") towards outgroup members is

also unexplored territory. The IRAP may therefore represent a new tool for identifying the specific (automatic) relational responses that contribute to conflict and its resolution.

Second, the study of implicit intergroup relations in Northern Ireland has so far been restricted to a relatively small sample of university students. Generalizing these findings will require that researchers move outside of the laboratory and into residential, school, industrial, and governmental settings. Doing so may provide deeper insight into a range of issues, from the development of automatic intergroup relational responses to the impact of integrated educational strategies on automatic stereotypes and prejudice. For instance, intergroup contact is often thought to represent a key driving force for the reduction of stereotyping and prejudice. If so, then we would expect students attending schools with members from both communities to demonstrate reduced relational response biases compared to their peers in segregated schools. RFT researchers have also weighed in on this issue, pointing to the unique history of derived stimulus relating in Northern Ireland, how it may sustain old stereotypes and prejudices (Watt, Keenan, Barnes, & Cairns, 1991) and how that same behavioral process (arbitrarily applicable relational responding; AARR) could be used to "change the very context in which verbal relations are formed, and in this way to loosen up rigid verbal relations that characterize those stereotypes and prejudices" (Dixon, Dymond, Rehfeldt, Roche, & Zlomke, 2003, p.138). Third, future work could also examine whether the automatic intergroup relational responses captured by the IRAP subtly *influence* daily behavior, from the real-world hiring decisions of organizations to the voting intention of undecided voters in local elections. In other words, the inclusion of indirect procedures could help predict meaningful real-world behaviors and also account for the discrepancy that often emerges between self-reported attitudes and other social behaviors in this context.

Finally, although we focused on intergroup relational responding in a post-conflict setting, the IRAP could also help inform our understanding of automatic prejudice and stereotyping in other domains. Although previous IRAP work has documented distinct patterns of relational responding in the context of racial (Drake et al., 2015), sexual (Timmins, Barnes-Holmes, & Cullen, 2016), and gender issues (Farrell, Cochrane, & McHugh, 2015), it has rarely examined if those same relational responses predict real-world prejudiced or discriminatory behavior, such as biased hiring decisions and payment practices, inappropriate physical or verbal behavior towards outgroup members, or increased likelihood to act aggressively towards a particular racial or ethnic group. Future work could therefore examine if such behavior-behavior relations actually exist, and if so, attempt to influence them by identifying the environmental factors that give rise to "'automatic" relational responses in the first instance.

Compliance with Ethical Standards

Funding: This study was funded by a postgraduate scholarship to the first author from the Irish Research Council for Science, Engineering, and Technology (IRCSET).

Conflict of Interest: The authors (SH, DBH, and SS) declare that they have no conflicts of interest.

Ethical approval: All procedures performed were in accordance with the ethical standards of the host institution and in-line with the 1964 Helsinki declaration and its later amendments.

Ethical approval: This article does not contain any studies with animals performed by any of the authors.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

Barnes-Holmes, D., Barnes-Holmes, Y., Hussey, I., & Luciano, C. (2016). Relational frame theory: Finding its historical and philosophical roots and reflecting upon its future development: An introduction to part II. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), *The Wiley handbook of contextual behavioral science* (pp. 117-128), West Sussex, UK: Wiley-Blackwell.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record, 60,* 527-542.

Bast, D., Barnes-Holmes, Y., & Barnes-Holmes, D. (2015). Developing an individualized Implicit Relational Assessment Procedure (IRAP) as a potential measure of self-forgiveness related to negative and positive behavior. *The Psychological Record, 65,* 717–730.

Cairns, E., & Darby, J. (1998). The conflict in Northern Ireland: Causes, consequences, and controls. *American Psychologist, 53(7),* 754-760.

Dasgupta, N. (2004). Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations. *Social Justice Research, 17,* 143-169.

De Houwer, J., Heider, N., Spruyt, A., Roets, A., & Hughes, S. (2015). The relational responding task: Toward a new implicit measure of beliefs. *Frontiers in Psychology*, 6, 319. doi: 10.3389/fpsyg.2015.00319

Dixon, M. R., Dymond, S., Rehfeldt, R. A., Roche, B., & Zlomke, K. R. (2003). Terrorism and relational frame theory. *Behavior and Social Issues, 12,* 129-147.

Drake, C. E., Kramer, S., Sain, T., Swiatek, R., Kohn, K., & Murphy, M. (2015). Exploring the reliability and convergent validity of implicit propositional evaluations of race. *Behavior and Social Issues, 24*, 68-87.

Drake, C. E., Seymour, K. H., & Habib, R. (2016). Testing the IRAP: Exploring the reliability and fakability of an idiographic approach to interpersonal attitudes. *The Psychological Record, 66*, 153-163.

Farrell, L., Cochrane, A., & McHugh, L. (2015). Exploring attitudes towards gender and science: The advantages of an IRAP approach versus the IAT. *Journal of Contextual Behavioral Science, 4*, 121-128.

Finn, F., Barnes-Holmes, D., Hussey, I., & Graddy, J. (2016). Exploring the behavioral dynamics of the implicit relational assessment procedure: The impact of three types of introductory rules. *The Psychological Record 66,* 309-321.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd edition, pp. 283-310). New York, NY: Cambridge University Press.

Gawronski, B., & Payne, B. K. (Eds.). (2010). *Handbook of implicit social cognition: Measurement, theory, and applications*. New York: Guilford Press.

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.

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.

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.

Hewstone, M., Lolliot, S., Swart, H., Myers, E., Voci, A., Al Ramiah, A., & Cairns, E. (2014). Intergroup contact and intergroup conflict. *Peace and Conflict: Journal of Peace Psychology, 20*, 39-53.

Hughes, J., Campbell, A., Lolliot, S., Hewstone, M., & Gallagher, T. (2013). Inter-group contact at school and social attitudes: Evidence from Northern Ireland. *Oxford Review of Education, 39*, 761-779.

Hughes, S., Hussey, I., Corrigan, B., Jolie, K., Murphy, C., & Barnes-Holmes, D. (2016). Faking revisited: Exerting strategic control over performance on the Implicit Relational Assessment Procedure. *European Journal of Social Psychology, 46,* 632–648.

Hussey, I., Mhaoileoin, D. N., Barnes-Holmes, D., Ohtsuki, T., Kishita, N., Hughes, S., & Murphy, C. (2016). The IRAP is nonrelative but not a-contextual: Changes to the contrast category influence men's dehumanization of women. *The Psychological Record, 66,* 291-299.

Hussey, I., Thompson, M., McEnteggart, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2015). Interpreting and inverting with less cursing: A guide to interpreting IRAP data. *Journal of Contextual Behavioral Science, 4*, 157-162.

Lee, S., Rogge, R. D., & Reis, H. T. (2010). Assessing the seeds of relationship decay using implicit evaluations to detect the early stages of disillusionment. *Psychological Science, 21*, 857-864.

Maloney, E., & Barnes-Holmes, D. (2016). Exploring the behavioral dynamics of the Implicit Relational Assessment Procedure: The role of relational contextual cues versus relational coherence indicators as response options. *The Psychological Record, 66,* 395–403.

McAlister, S., Scraton, P., & Haydon, D. (2009). *Childhood in Transition. Experiencing Marginalisation and Conflict in Northern Ireland*. Queen's University Belfast, Save the Children, The Prince's Trust: Belfast.

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.

McKittrick, D., Kelters, S., Feeney, B., Thornton, C., & McVea, D. (2007). *Lost Lives: The Stories of the Men, Women and Children who Died as a Result of the Northern Ireland Troubles*. Edinburgh: Mainstream Publishing.

Nicholson, E., & Barnes-Holmes, D. (2012). Developing an implicit measure of disgust propensity and disgust sensitivity: Examining the role of implicit disgust propensity and sensitivity in obsessive-compulsive tendencies. *Journal of Behavior Therapy and Experimental Psychiatry, 43*, 922-930.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

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.

Northern Ireland Council for Integrated Education. (2014). Annual report. Retrieved January 3, 2014, from http://www.nicie.org/wp-content/uploads/2014/11/NICIE-Annual-Report-13-14-web.pdf

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

O'Shea, B., Watson, D. G., & Brown, G. D. A. (2016) Measuring implicit attitudes: A positive framing bias flaw in the Implicit Relational Assessment Procedure (IRAP). *Psychological Assessment, 28, 158-170.*

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17,* 523-534.

Shirlow, P., & Murtagh, B. (2006). *Belfast: Segregation, violence and the city*. London: Pluto Press.

Tam, T., Hewstone, M., Kenworthy, J. B., Cairns, E., Marinetti, C., Geddes, L., & Parkinson, B. (2008). Postconflict reconciliation: Intergroup forgiveness and implicit biases in Northern Ireland. *Journal of Social Issues, 64,* 303-320.

Timmins, L., Barnes-Holmes, D., & Cullen, C. (2016). Measuring implicit sexual response biases to nude male and female pictures in androphilic and gynephilic men. *Archives of sexual behavior, 45*, 829-841.

Turner, R. N., Tam, T., Hewstone, M., Kenworthy, J., & Cairns, E. (2013). Contact between Catholic and Protestant schoolchildren in Northern Ireland. *Journal of Applied Social Psychology, 43,* 216-228.

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the implicit relational assessment procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry, 48,* 59-65. doi:10.1016/j.jbtep.2015.01.004

Watt, A., Keenan, M., Barnes, D., & Cairns, E. (1991). Social categorization and stimulus equivalence. *The Psychological Record, 41,* 33-50.

Appendix A

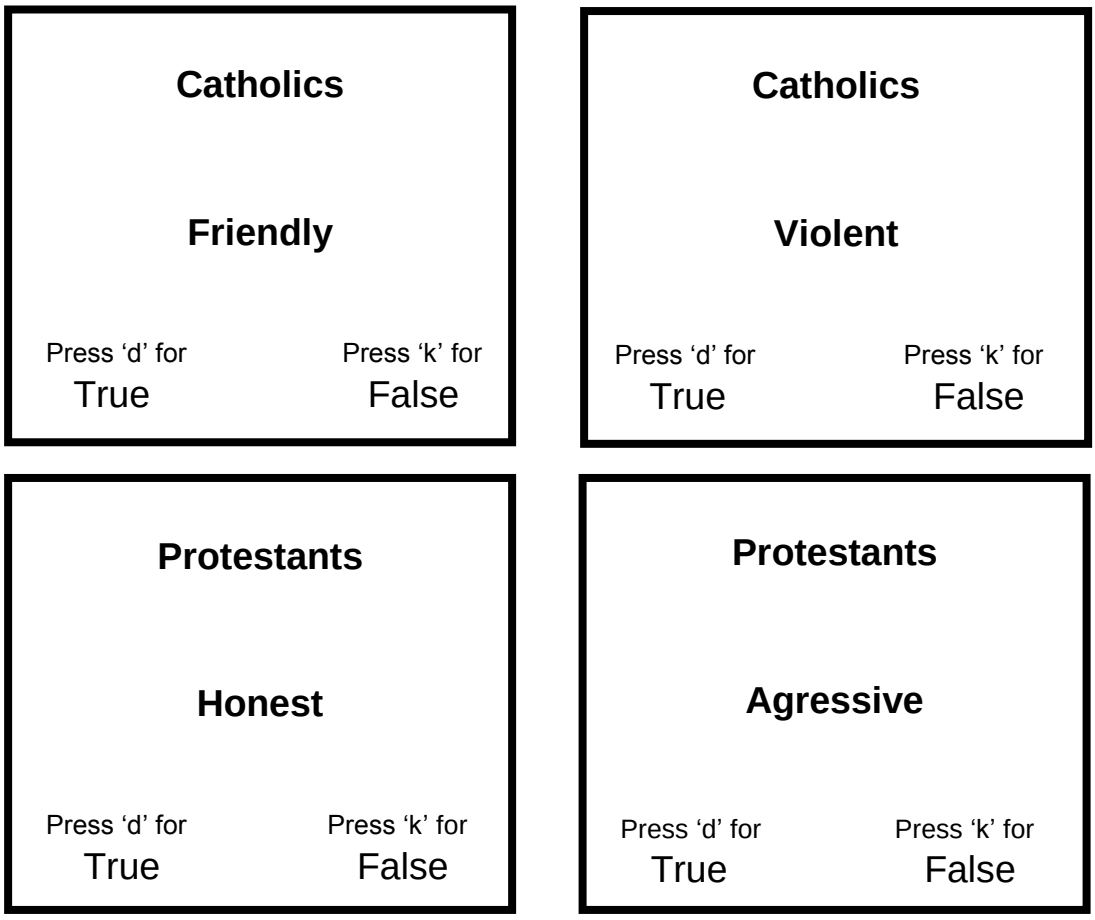

Figure 1. Examples of the four trial types used in the IRAP. The label stimuli (e.g., "*Catholics"* and "*Protestants"*), target stimuli (e.g., "friendly" and "violent") and relational response options ("True" and "False") are indicated.

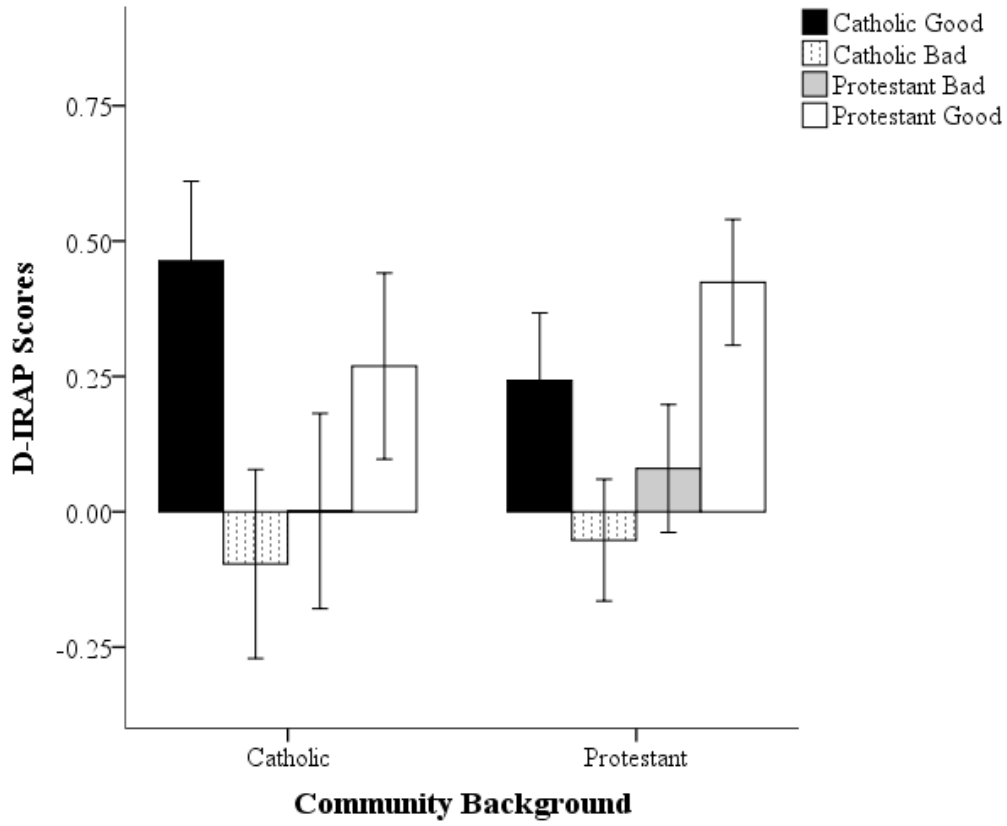

Figure 2. Mean D-*IRAP* scores as a function of Community Background (Catholic vs. Protestant). A positive value indicates a positivity bias while a negative score indicates a negativity bias towards a given group. Error bars indicate 95% confidence intervals.

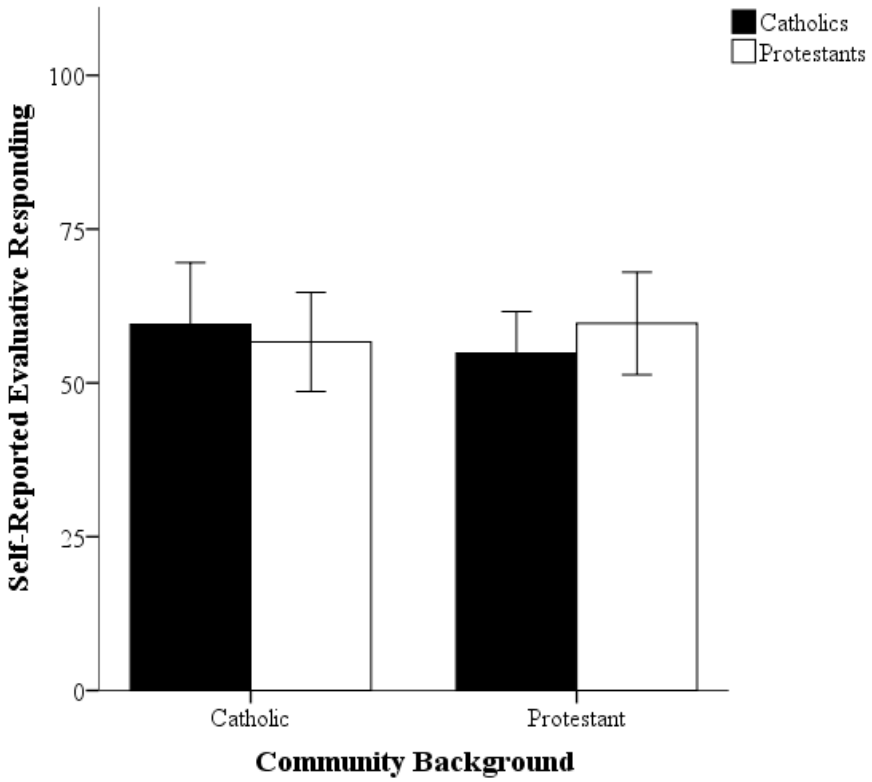

Figure 3. Mean self-reported evaluative responses as a function of Community Background (Catholic vs. Protestant). Error bars indicate 95% confidence intervals.

Appendix B

D-IRAP scores can be calculated in the following way: (1) discard response-latency data from practice blocks and only use test blocks data; (2) eliminate latencies above 10,000 ms from the data set; (3) remove all data for a participant if he or she produces more than 10% of test-block trials with latencies less than 300 ms; (4) compute 12 standard deviations for the four trial types: four from the response latencies from Test Blocks 1 and 2, four from the latencies from Test Blocks 3 and 4, and four from Test Blocks 5 and 6; (5) calculate the mean latencies for the four trial types in each test block (resulting in 24 mean latencies in total); (6) calculate difference scores for each of the four trial types for each pair of test blocks by subtracting the mean latency of the Rule A block from the mean latency of the corresponding Rule B block; (7) divide each difference score by its corresponding standard deviation (see step 4). This yields 12 *D*-IRAP scores, one score for each trial type for each pair of test blocks. Finally, (8) calculate four overall trial type scores by averaging the scores for each trial type across the three pairs of test blocks. Note that these four trial-type scores can be collapsed into an overall *D*-IRAP score if the researcher so chooses.