Faking a race IRAP effect in the context of single versus multiple label stimuli

Ciara Dunne[1], Ciara McEnteggart[2], Colin Harte[2], Dermot Barnes-Holmes[2], and Yvonne Barnes-Holmes[2]

[1] Department of Psychology, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

[2] Department of Experimental, Clinical and Health Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

Ciara Dunne: behaviourinfo@gmail.com

Ciara McEnteggart: ciara.mcenteggart@ugent.be

Colin Harte: colin.harte@ugent.be

Dermot Barnes-Holmes: dermot.barnes-holmes@ugent.be

Yvonne Barnes-Holmes: Yvonne.barnesholmes@ugent.be

Corresponding author:

Colin Harte

Department of Experimental, Clinical, and Health Psychology

Ghent University

Henri Dunantlaan 2

9000 Ghent

Belgium

Email: colin.harte@ugent.be

Authors' Note This article was prepared with the support of an Odysseus Group 1 grant awarded to the third author by the Flanders Science Foundation (FWO).

Abstract

In the current study, white participants were exposed to a single-label or multiple-label racial bias IRAP before and after a faking instruction (i.e., two exposures to the IRAP). The faking instruction involved asking all participants to imagine that they were a black person when completing the second IRAP. The results indicated that participants produced evidence of pro-white and anti-black biases both before and after receiving the faking instruction. Analyses of variance revealed no main or interaction effects for the single- versus multiple-label variable, and trial-type specific paired *t*-tests yielded no significant differences between the pre- and post-faking instruction IRAPs. The results were consistent with previous racial bias findings using the IRAP and supported the conclusion that faking only occurs when participants are provided with specific information about the task parameters. Implications for faking research, and the impact of instructions generally, on the IRAP are discussed.

Key words: Implicit Relational Assessment Procedure, IRAP, single and multiple labels, faking, racial bias

Novelty and Significance

What is already known about the topic?

- The IRAP has demonstrated relatively robust effects for white participants in the domain of racial bias.
- In continuing to explore the IRAP as a measure of racial bias, it seems important to examine potential moderating variables.

What this paper adds?

- The impact of using single- and multiple-labels have never before been analysed within a single IRAP study.
- The impact of faking instructions on racial bias in the IRAP have never before been analysed whereby the faking instructions did not focus on the specific parameters of the task.
- Implications for faking research, and the impact of instructions generally, on the IRAP are discussed.

Relational frame theory (RFT: Hayes, Barnes-Holmes, & Roche, 2001) is a functional-analytic account of human language and cognition. The theory stemmed from the study of verbal behaviour as derived relational responding, which focuses on the emergence of novel behaviours that have not been directly trained or reinforced (see Hayes, et al.). In an attempt to further develop methodologies for assessing relational responding, researchers began to explore relations that were likely to conflict with those already established in participants' pre-experimental histories (e.g., Barnes-Holmes, Barnes-Holmes, Stewart, & Boles 2010). One such approach, based directly on RFT, is the Implicit Relational Assessment Procedure (IRAP). The IRAP is a computer-based task that requires participants to respond quickly and accurately in ways that are either consistent or inconsistent with their pre-existing verbal histories, and assumes that individuals will respond more quickly to relations that are consistent rather than inconsistent with those histories. The difference in response latencies between consistent and inconsistent relational responding generates what is known as the IRAP effect. To date, the IRAP has been used to examine relational responding across a wide range of social and clinical domains, with a recent meta-analysis demonstrating robust effects and relatively high predictive validity (see Vahey, Nicholson, & Barnes-Holmes, 2015).

In an early study, Barnes-Holmes, Murphy, Barnes-Holmes, and Stewart (2010) used the IRAP to assess the racial biases of white Irish individuals toward black individuals in Ireland. Across half of the IRAP trials, participants were required to confirm that pictures of white men holding guns were "Safe" and pictures of black men holding guns were "Dangerous", whilst across the other half of the trials, they were required to confirm that pictures of black men holding guns were "Safe" and pictures of white men holding guns were "Dangerous". Results demonstrated an in-group (pro-white/anti-black) bias on the white-positive and black-negative trial-types, where participants were able to confirm that white people were positive and black people were negative more quickly than they were able to deny these relations.

In other studies that also examined black/white racial bias using the IRAP, broadly similar effects have been found (Drake et al., 2010; 2015; Power, Harte, Barnes-Holmes, & Barnes-Holmes, 2017). Although some differences did emerge in the results across studies, it is difficult to isolate key variables because the studies did differ in multiple ways (e.g., samples employed, inclusion criteria, stimuli presented, etc.). Overall however, the racial bias IRAP effect appears to be relatively robust for white participants, in that they typically show pro-white and anti-black effects; black participants, however, do not (see Power et al.).

In continuing to explore the IRAP as a measure of racial bias, it seems important to examine how specific variables may or may not moderate IRAP effects. Two such variables targeted in the current study were the impact of faking instructions and the use of single- versus multiple-labels as stimuli. The impact of faking instructions on racial bias in the IRAP has been examined in only one previous study (Hughes et al., 2016), and the use of single- versus multiple-labels has not been examined systematically in any study using the IRAP (see below for details).

In Experiment 3 in the 'faking' study reported by Hughes et al. (2016), the IRAP contrasted four positive labels ("safe", "friendly", "polite", and "kind") with four negative labels ("dangerous", "aggressive", "rude", and "violent"), and eight colour images of black individuals (four men and four

women), with eight colour images of white individuals (four men and four women); the words "True" and "False" were presented as response options. Participants were presented with faking instructions between a first and second IRAP. That is, participants were exposed to a baseline IRAP without any faking instructions. They were then provided with instructions that oriented attention toward the core parameters of the task, which emphasized speed and accuracy, and asked participants to try to trick the computer into thinking that they liked black people and disliked white people. Half of the participants received this instruction alone, whereas the remaining participants were also instructed before each block whether they should respond slowly or quickly to trick the computer as instructed previously. The results showed a reduction in the racial response bias on the IRAP for those participants who were simply oriented toward the core parameters of the task and asked to trick the computer. Participants who were told exactly how to do this showed a reversal in the response bias. It thus appears that clear evidence of faking only emerged when participants were explicitly instructed when to respond quickly and when to respond slowly on the IRAP, and were reminded to do so before each block. Faking was also demonstrated by Drake et al. (2016), but again, only with the use of highly specific task instructions (and the stimuli were not relevant to racial bias). At the current time, therefore, there has been no study of the ability of participants to fake an IRAP effect when the faking instruction does not orient them toward the core parameters of the task, and explicitly asks them to trick the computer. Critically, other studies that have provided faking instructions that do not focus on the parameters of the task have failed to show any impact on IRAP performances (McKenna et al., 2007; Hughes et al. Experiment 1), but none of these have been conducted with stimuli that have attempted to assess racial bias. The current study attempts to fill this gap.

At the time of writing, the impact of instructions on IRAP performances, including those that targeted faking, had been analysed across a number of studies. As noted above, another potentially important moderating variable, which has not been explored in any published IRAP research, is the effect of the use of single versus multiple labels. The term label, as used in the context of the IRAP, refers to the stimulus that appears at the top of the screen (the stimulus that appears in the middle of the screen is typically referred to as the 'target'). The original IRAP software only permitted researchers to present one of two labels on each trial, but one of up to 12 different targets. This type of IRAP is referred to as the single label-IRAP (SL-IRAP) because only one label is used to define each of the relevant categories (the word "Safe" versus "Dangerous"). Subsequent development of the IRAP software in 2008 allowed researchers to present multiple labels for each category in what was referred to as the multiple label-IRAP (ML-IRAP). The impact of this parameter, however, has never been analysed within a single study. In an early example of a single-label IRAP, Barnes-Holmes, Murphy, et al. (2010) presented the labels "Safe" and "Dangerous" with various pictures of black and white men holding guns. If, however, a multiple-label IRAP was used, the label "Safe" might be presented on some trials, with semantically similar words such as "Protector" and "Guardian" presented on other trials. Similarly, "Dangerous" could be presented with "Gangster" on some trials and with "Criminal" on other trials. Note that Hughes et al. (2016, Experiment 3) did employ a multiple-label IRAP with racial stimuli, but did not compare it directly with a single-label version.

In the current study, participants were exposed to one of two IRAPs: a single-label racial bias IRAP (SL-IRAP) or a multiple-label racial bias IRAP (ML-IRAP). Having completed one exposure to the IRAP, all participants were provided with a faking instruction that asked them to pretend that they were a black person living in a predominantly white country, before completing the same IRAP (either SL or ML) a second time. Thus, the experiment had three core aims: (1) to attempt to replicate the racial bias effect; (2) to determine if any evidence of racial bias on the IRAP during the first exposure would be eliminated following the faking instruction[1]; and (3) to determine whether there would be any interaction effects between the SL- and ML-IRAPs and faking. While these were the core aims of the current work, it should be noted that the data were collected before the faking study on racial bias was conducted. Hence, we refrained from making specific predictions concerning the impact of single versus multiple labels and their interaction with faking instructions in the context of racial bias.

Method

Participants

Forty-four participants, 20 male and 24 female, aged 18 to 30 years (M = 24 years), completed the experiment in a quiet cubicle in an experimental psychology lab. All participants were white Irish nationals drawn from a convenience sample of undergraduate students in an Irish university. Participants were randomly assigned to one of two conditions: the SL-IRAP or the ML-IRAP.

Setting

Each participant completed all stages of the experiment on an individual basis. The experimenter remained outside the room and was only present during instructional and debriefing stages.

Apparatus and Materials

All participants completed the IRAP on a standard personal computer. The IRAP software (2009 version) presented the stimuli and recorded participant responses.

SL-IRAP. The SL-IRAP presented two label stimuli (*Safe* or *Dangerous*), with one of 6 target stimuli, consisting of the 3 pictures of white men holding a gun and 3 pictures of black men holding a gun, as well as two response options (*True* and *False*). All six men pictured were wearing plain white t-shirts and were standing in front of the same red-brick background. The same stimuli had been employed in a virtual reality study of racial prejudice by Greenwald, Oakes, and Hoffman (2003). Based on the various sample-target combinations, the IRAP comprised four trial-types; *Positive/White*, *Positive/Black*, *Negative/White*, and *Negative/Black* (see Figure 1). The IRAP software recorded all response data, including accuracy, and latency.

[1] The faking instruction employed in the current study did not specify the core parameters of the task itself nor instruct the participant when to respond quickly or slowly. Instead, the instruction could be seen as asking the participant to take the perspective of a black person during the task. Although the data for the current study was collected over five years ago, recent research using the IRAP has indicated that it may be sensitive to perspective-taking (Barbero-Rubio, Lopez-Lopez, Luciano, Eisenbeck, 2016; Kavanagh, Barnes-Holmes, Barnes-Holmes, McEnteggart, & Finn, 2018).

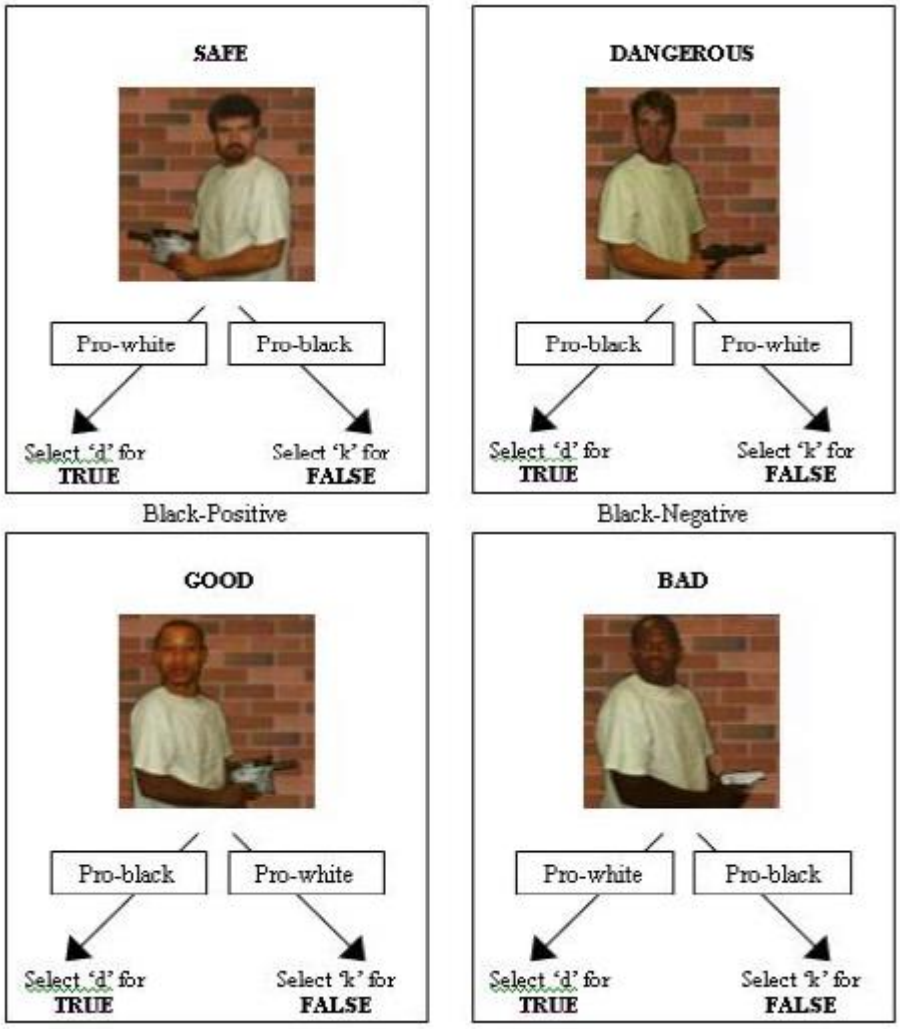

Figure 1. Examples of the IRAP trial-types. The superimposed arrows with text boxes indicate the responses deemed Pro-white or Pro-black, but these boxes and arrows did not appear on-screen during the experiment.

ML-IRAP. The ML-IRAP was similar to the SL-IRAP, except that six words denoted safety (*safe, good, hero, guardian, protector,* and *police*) and six words denoted danger (*dangerous, bad, villain, criminal, robber,* and *gangster*) as label stimuli. The same target stimuli and response options as the SL-IRAP were again presented.

Procedure

Participants were first allocated to one of the two conditions: the SL-IRAP or the ML-IRAP. For both conditions, the experiment consisted of four stages. In Stage 1, participants completed a range of questionnaires that aimed to assess racial bias: Discrimination and Diversity Scales (DS and DV; Wittenbrink, Judd, & Park, 1997); the Modern Racism Scale (MRS; McConahay, 1986); and Likert Scales. The details of these measures and their results are not reported here because correlational analyses between the IRAPs (at pre- and post-faking instruction) and the measures failed to yield any statistically

significant effects. Stage 2 involved exposure to the pre-faking instruction, baseline IRAP. Stage 3 involved the delivery of the faking instruction. Finally, Stage 4 involved a second exposure to the IRAP that was identical to Stage 2, but with a reduced latency criterion to control for practice effects (see Figure 2 for an illustration of the experimental sequence).

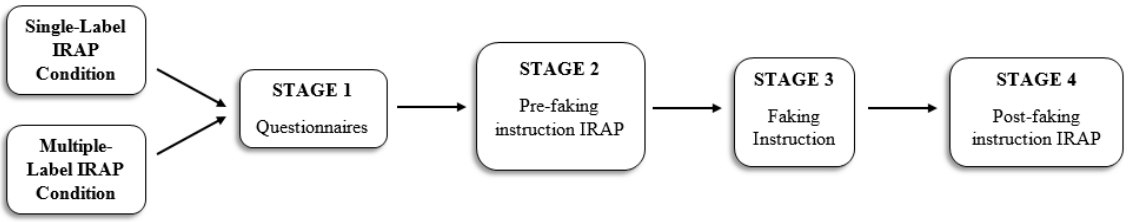

Figure 2. An illustration of the experimental sequence.

Stage 1: Questionnaires. Each participant completed the questionnaires in the following order: (1) DS and DV Scales; (2) the MRS; and (3) the Likert Scales.

Stage 2: Pre-faking instruction IRAP. Prior to the first practice block, participants were verbally instructed on how to complete the IRAP, as is standard practice in IRAP research. They were advised that each trial would present a word on top of the screen, with a picture in the center of the screen, and that their task was to respond with *True* or *False*, as appropriate (see Figure 1). Participants were informed that the pattern of responding would switch to an opposite pattern across each block. These instructions also highlighted the criterion for accurate (=/>80%) and fast (=/<2,000 ms.) responding.

The IRAP consisted of blocks of 24 trials, with each of the four trial-types presented 6 times within each block. On each trial, a label (e.g., *Safe*) appeared at the top, a target (e.g., picture of a white man holding a gun) in the middle, and both response options (*True and False)* on the bottom left- and right-hand corners. Participants selected a response by pressing *D* (for the left option) or *K* (for the right). If a participant emitted a correct response, the screen cleared, and the next trial appeared. If a participant responded incorrectly, a red X appeared until a correct response was emitted.

The feedback contingencies for the IRAP alternated across blocks in one of two patterns. One pattern was defined as a pro-white/anti-black pattern, the other as a pro-black/anti-white pattern. The pro-white/anti-black pattern required that participants respond in the following way: Safe-White/True; Safe-Black/False; Dangerous-White/False; Dangerous-Black/True. The pro-black/anti-white pattern required the opposite: Safe-White/False; Safe-Black/True; Dangerous-White/True; Dangerous-Black/False. Hence, correct responding involved switching between each pattern from block to block. The order in which the two types of blocks were presented was counterbalanced across participants.

The IRAP commenced with a minimum of two practice blocks. If participants failed to achieve both accuracy and latency criteria across a pair of blocks, they received automated feedback, and practice blocks continued to a maximum of four pairs of blocks. Failing to meet the criteria after four pairs of practice blocks terminated participation and these data were discarded. When the criteria were reached on

a pair of practice blocks, participants proceeded automatically to three pairs of test blocks. No performance criteria were employed for participants to progress through test blocks, but performance feedback was presented at the end of each block to encourage participants to maintain the criteria. The program automatically recorded response accuracy (based on the first response emitted on each trial) and response latency (time in ms. between trial onset and emission of correct response) on each trial.

Stage 3: Faking instruction. Stage 3 involved a printed faking instruction that asked participants to imagine during the next IRAP (Stage 4) that they were a black person living in a predominantly white country. The purpose of this instruction was to determine if participants could deliberately change the pattern and/or size of the IRAP effects from their baseline performances. The instructions were as follows:

> You are currently completing a measure of racial prejudice on a computer. Having completed one exposure to the computer task, I would like you complete a second exposure. However, this time I would like you to do your very best to imagine that you are a black person while you complete the task. That is, imagine that you are a black person who lives in a predominately white country. Please write below in your own words what you have just read above.

Stage 4: Post-faking instruction IRAP. Stage 4 involved a second exposure to the IRAP. That is, participants exposed to a baseline SL-IRAP were again exposed to an SL-IRAP, while participants exposed initially to an ML-IRAP were again exposed to an ML-IRAP. The only difference between the IRAPs here and in Stage 2 is that the second IRAP involved a latency criterion that was now reduced to 1,750 ms. to control for practice effects. Upon completion, participants were thanked and debriefed.

Results

The primary datum was response latency, defined as time in ms. between trial onset and a correct response. In accordance with previous IRAP studies, response latency data were transformed into *D*-IRAP scores for each participant (see Nicholson & Barnes-Holmes, 2012). The foregoing data transformation yielded positive *D*-IRAP scores for positive biases and negative *D*-IRAP scores for negative biases (i.e., the *D*-IRAP scores for the two black trial-types were inverted, see Hussey, Thompson, McEnteggart, Barnes-Holmes, & Barnes-Holmes, 2015). The mean *D*-IRAP scores for each of the four trial-types for each IRAP, both pre- and post-faking, are presented in Figure 3. In general, the pattern of results showed positive biases for both IRAPs, at both pre- and post-faking, across the same three trial-types (*White-Positive*, *Black-Positive*, *White-Negative*). Negative biases were produced for the *Black-Negative* trial-type on both IRAPs at both pre- and post-faking.

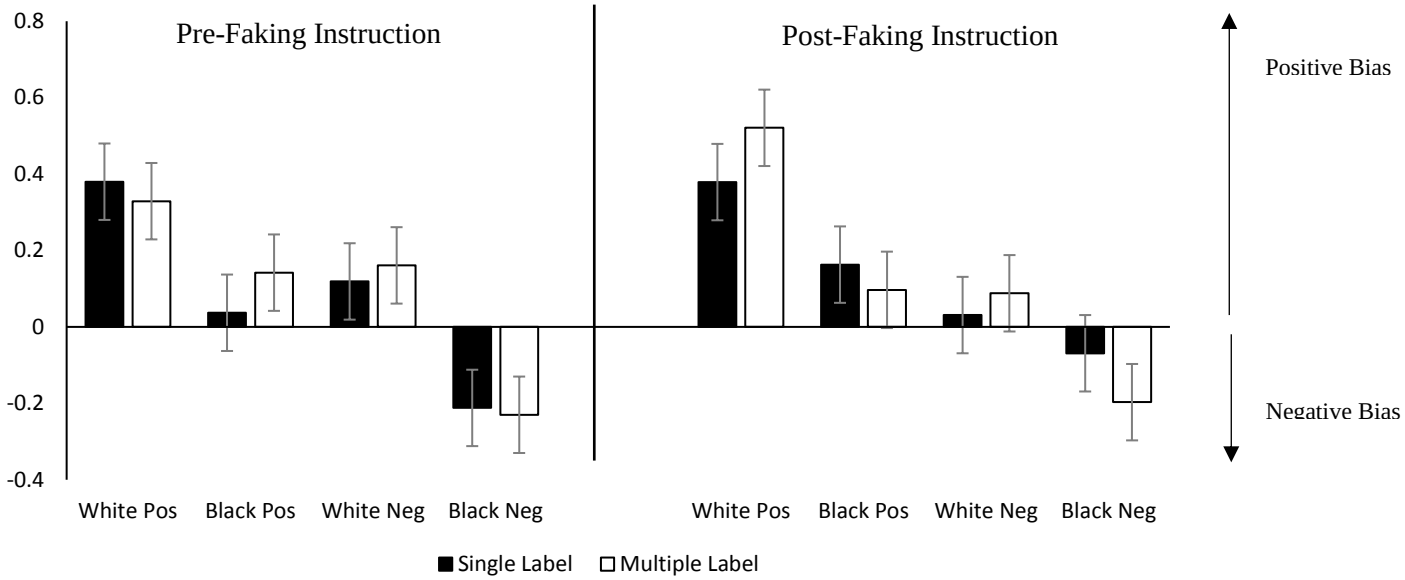

Figure 3. Mean *D*-IRAP scores for the four IRAP trial-types for SL- and ML-IRAPs pre- and post-faking instruction. The values for the four Black trial-type *D*-IRAP scores were inverted such that a positive *D*-IRAP score indicates a positive bias whereas a negative score indicates a negative bias.

An exploratory 2x2x4 mixed repeated measures analysis of variance (ANOVA) was conducted on the *D*-IRAP scores, with condition (i.e., SL and ML) as the between-participant variable and trial-type and IRAP exposure (pre- and post-faking instruction) as within-participant variables. The analysis revealed only a significant effect for trial-type, [$F(3, 126) = 21.25$, $p < .0001$, $\eta p^2 = .33$], but no other main or interaction effects (all *ps* > .27). Given the lack of effect for condition, the data were collapsed across IRAP type (i.e., SL and ML). Given that previous studies had reported significant effects for faking, albeit under specific conditions, we conducted follow-up tests described below to determine any suggestive trends in our findings that may indicate a faking effect.

Scheffe post-hoc tests indicated that the effects for each trial-type, collapsed across the SL- and ML-IRAPS, differed significantly from each other (all *p*s < .05), except for the comparison between *Black-Positive* and *White-Negative* ($p > .99$). Four paired *t*-tests confirmed that each of the four IRAP effects did not differ significantly from pre- to post-faking instruction (*p*s > .17). Eight one-sample *t*-tests indicated that three effects in the pre-faking instruction IRAP were significant (*p*s < .05), except for *Black-Positive* ($p = .18$). In the post-faking instruction IRAP, three effects were significant (*p*s < .05), except for *White-Negative* ($p = .26$).

Overall, therefore, the general pattern of results did not differ significantly following the faking instruction, with participants maintaining a significant negative bias on the *Black-Negative* trial-type across both IRAP exposures. The positive bias on the *Black-Positive* trial-type became significant, whereas the positive bias on *White-Negative* became non-significant following the faking instruction. Although these changes could be seen as reflecting the impact of the faking instruction, it is important to

note that there was little evidence of significant change when comparing these trial-types directly (using paired *t*-tests) across exposures. Furthermore, it should be noted that the largest change across IRAP exposures was for the *White-Positive* trial-type (*Mean Diff*. = -.1), which counter-intuitively showed a shift towards an increasingly positive bias toward white people following the faking instruction.

Discussion

The current study sought to investigate the potential impact of using single versus multiple labels in an IRAP with the use of a faking instruction in the context of racial bias, and to investigate any potential interaction among these variables. A main effect was obtained for IRAP trial-type, but we found little evidence that this was moderated by either the use of single versus multiple labels or the faking instruction. Indeed, the general pattern of effects for race previously reported by Barnes-Holmes, Murphy et al. (2010) and by Power et al. (2017) were found again in the current study. Specifically, a negative racial bias was found on the *Black-Negative* trial-type, and this was observed at both pre- and post-faking instruction. Negative racial biases have also been found in other IRAP studies (Drake et al., 2015; 2010), although direct comparisons with the current research are difficult because there were many methodological differences (see Power et al. for a more detailed discussion).

As noted in the Introduction, the only study that has shown the significant impact of a faking instruction on a racial bias IRAP is one that oriented participants toward the core parameters of the task and explicitly asked them to "trick the computer". Indeed, evidence for a full faking effect (i.e., a complete reversal in the relevant IRAP effect from pre- to post-instruction) was only obtained when participants were explicitly instructed to respond slowly on some trials and quickly on others, and were reminded to do so before each block of test trials. In contrast, the faking instructions presented in the current study were broadly similar to those employed in previous studies that have reported the absence of a faking effect, in that they were presented only once at the beginning of the IRAP, did not highlight the core parameters of the task, nor explicitly ask participants to slow down or speed up on certain trials. It is reassuring, therefore, that the lack of a faking effect observed in the current study is consistent with other published studies that have used broadly similar faking instructions (e.g., see Hughes et al., 2016, Experiment 1). At this point, it appears that IRAP effects can be faked, but only under very specific forms of instruction in which the parameters of the task are made apparent to participants. It remains to be determined, however, the extent to which such faking may be moderated by the domain targeted within an IRAP. For example, it may be that faking is more or less readily observed with an IRAP that targets race, rather than a clinically-relevant domain.

The current study also sought to determine the potential impact of using single versus multiple labels in the IRAP, and no main or interaction effects were found for this variable. Such a result may be reassuring for previous IRAP research, some of which has employed single labels, whilst other research employed multiple labels (e.g., Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008; Cagney, Harte, Barnes-Holmes, Barnes-Holmes, & McEnteggart, 2017; McKenna et al., 2007; Nicholson & Barnes-Holmes, 2012). On balance, it is important to bear in mind that this variable may be less important when the labels are relatively simple (e.g., single words or pictures) versus more complex (e.g., full statements or complex pictorial stimuli). Indeed, this point was highlighted recently in a study by Drake,

Timko, and Luoma (2016) that presented participants with an IRAP with just two labels ("I am willing to have" and "I try to get rid of") and six targets (anxiety, fear, worry, contentment, happy, and relaxation). Given the time constraints to respond in under 2000ms, it is possible that at least some participants responded to just the first two words of each label to discriminate successfully between them. As a result, participants may have read, for example, "I am anxiety" from the label and target combination "I am willing to have" and "anxiety". Interestingly, the correlations between the IRAP data and participant scores on the Drexel Defusion Scale and the Acceptance and Action Questionnaire were consistent with this interpretation (see Kavanagh, Hussey, McEnteggart, Barnes-Holmes, & Barnes-Holmes, 2016). Thus, although the current study suggests that the use of single versus multiple labels may have limited impact on IRAP effects, it would be unwise to conclude that this remains the case in all IRAP research contexts.

Although the current study failed to find a significant effect for the faking instructions, it is important to acknowledge that a number of recent IRAP studies have reported significant effects for instructions generally. For example, Finn, Barnes-Holmes, Hussey, and Grady (2016) reported that the specific pattern of trial-type effects found on an IRAP may be moderated by the extent to which instructions on how to respond on the IRAP are specific or general. Interestingly, however, these instruction-based effects were found when the relevant instructions were presented repeatedly before each test block. In contrast, Finn, Barnes-Holmes, and McEnteggart (2018) failed to find a clear effect for instructions that were presented only once at the beginning of the IRAP. Again, therefore, it appears that instruction effects on the IRAP, at a generic level, are most impactful when they are presented repeatedly and before each test block. At the present time, it remains unclear why such instructional effects occur. For example, when instructions are presented only once at the beginning of an IRAP, do they fail to impact upon performance because participants: (i) "forget" to follow them, unless reminded before each test block; or (ii) "remember" them but are not sufficiently motivated to follow them without reminders? Future research could certainly address this issue.

A related issue concerns the ecological validity of studies that involve examining instructional effects, particularly those related to faking. Given that the IRAP is increasingly used in clinically-relevant research, it seems important to better understand the role of instructions in terms of when and how they have their impact on IRAP performances. On the one hand, the potential impact of faking could be seen as largely irrelevant when the IRAP is employed in the standard way, given that such effects are only observed when highly detailed instructions with regard to the task parameters are repeatedly presented before each block of trials. On the other hand, some caution may be required in interpreting IRAP effects when they are obtained from samples of participants who have been exposed to the IRAP or other latency-based measures across many previous studies (see Finn et al., 2018). Indeed, as argued by Finn et al., such previous exposures could function in a similar manner to the presentation of specific instructions, particularly when participants are fully debriefed after each study. Furthermore, the same general point could be made with respect to research conducted with many, if not all, latency-based measures. That is, it may be important for researchers to record how many latency-based measures participants have completed prior to the study presently being reported.

In reflecting upon the findings of the current study, it seems important to consider why the two IRAPs (SL versus ML) and the faking manipulation failed to produce any statistically significant differences in performance. In the case of the two IRAPs, the simplest explanation would be that the relational/verbal functions of "safe" and "dangerous" used in the SL-IRAP overlapped considerably with the functions of the other words presented in the ML-IRAP. In RFT terms, functionally there was little, if any, difference between the SL- and ML-IRAPs in terms of the stimulus control provided by the label stimuli (i.e., because in these contexts, the label stimuli used across the IRAPs participated in the same frames of coordination; frames containing safety words versus danger words). In terms of the faking manipulation, perhaps a future study might attempt to increase the extent to which participants were encouraged to take the perspective of a black person while completing the post-faking IRAP. For example, participants might be asked to view a short video clip designed to evoke a strong sense of empathy with black people who have experienced discrimination or prejudice while living in a predominantly white country (e.g. Finlay & Stephan, 2000). Indeed, such research could be important in developing techniques for reducing racial bias if the empathy manipulation was found to impact on the IRAP performances, which appeared to be absent in the current study.

In summary, the current work may have the following implications for IRAP research and work on derived stimulus relations generally. First, the current work further demonstrates the utility and robust nature of the IRAP as a measure of racial bias. Second, the lack of a faking effect observed is consistent with other related research in the literature. Future research should perhaps consider the implications of this variable in clinically-relevant domains because the impact of this variable has not yet been determined. Third, while the lack of an effect found for the single- versus multiple-label manipulation is promising for IRAP research conducted to date, it should nonetheless still be a variable taken into consideration when designing and conducting future research using the IRAP because this may be moderated by context and domain. Finally, the current research has implications for broader research using rules and instructional effects, and perhaps research on other moderating variables such as procedures designed to increase empathy with an outgroup. Given the increasing use of the IRAP in clinically relevant domains, and the growing literature on the impact of instructions on performance, the use of any sort of instructions (faking or otherwise) has serious implications for not only future IRAP research, but also for the use of latency-based measures generally.

Compliance with Ethical Standards

Declaration of Interest: This article was prepared with the support of an Odysseus Group 1 grant awarded to the fourth author by the Flanders Science Foundation (FWO).

Ethical Approval: All procedures performed in the studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

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

References

Barber-Rubio, A., Lopez-Lopez, J., Luciano, C., & Eisenbeck, N. (2016). Perspective-taking measured by implicit relational assessment procedure (IRAP). *The Psychological Record, 66*(2), 243-252. doi: 10.1007/s40732-016-0166-3

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.

Barnes-Holmes, D., Hayden, E., Barnes-Holmes, Y., & Stewart, I. (2008). The Implicit Relational Assessment Procedure (IRAP) as a response-time and event-related potentials methodology for testing natural verbal relations. *The Psychological Record, 58*, 497-516.

Barnes, D., Lawlor, H., Smeets, P.M., & Roche, B. (1996). Stimulus equivalence and academic self-concept among mildly mentally handicapped and nonhandicapped children. *The Psychological Record, 46,* 87-107.

Barnes-Holmes, D., Murphy, A., Barnes-Holmes, Y., & Stewart, I. (2010). The Implicit Relational Assessment Procedure (IRAP): Exploring the impact of private versus public contexts and the response latency criterion on pro-white and anti-black stereotyping among white Irish individuals. *The Psychological Record, 60*, 57-66.

Cagney, S., Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C. (2017). Response biases on the IRAP for adults and adolescents with respect to smokers and nonsmokers: The impact of parental smoking status. *The Psychological Record. 67*(4), 473-483. doi: 10.1007/s40732-017-0249-9

Cairns, E. (1984). Social identity in Northern Ireland. *Human Relations, 37*, 1095-1102.

Dixon, M., Rehfeldt, R. A., Zlomke, K.M., & Robinson, A. (2006). Exploring the development and dismantling of equivalence classes involving terrorist stimuli. *The Psychological Record, 56*, 83-103.

Drake, C.E., Kellum, K.K., Wilson, K.G., Luoma, J.B., Weinstein, J.H., & Adams, C.H. (2010). Examining the Implicit Relational Assessment Procedure: Four preliminary studies. *The Psychological Record, 60,* 81-86.

Drake, C.E., Kramer, S., Sain, T., Swiatek, R., Kohn, K., & Murphy, M. (2015). Exploring the reliability and convergent validity of implicit racial evaluations. *Behavior and Social Issues, 24,* 68-87. doi: 10.5210/bsi.v.24i0.5496

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. doi: 10.1007/s40732-015-0160-1.

Dymond, S. & Barnes, D. (1995). A transformation of self-discrimination response functions in accordance with the arbitrarily applicable relations of sameness, more-than, and less-than. *The Journal of the Experimental Analysis of Behavior, 64*, 163-184.

Dymond, S. & Barnes, D. (1996). A transformation of self discrimination response functions in accordance with the arbitrarily applicable relations of sameness, more than, less than: Erratum. *The Journal of the Experimental Analysis of Behavior, 66*, 348-360.

Finlay, K.A. & Stephan, W.G. (2000). Improving intergroup relations: The effects of empathy on racial attitudes. *Journal of Applied Social Psychologcy, 30*(8), 1720-1737.

Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2018). Exploring the single-trial-type-dominance-effect in the IRAP: Developing a Differential Arbitrarily Applicable Relational Responding Effects (DAARE) model. *The Psychological Record, 68*(1), 11-25. doi: 10.1007/s40732-017-0262-z

Finn, M., Barnes-Holmes, D., Hussey, I., & Grady, J. (2016). Exploring the behavioural dynamics of the Implicit Relational Assessment Procedure: The impact of three types of introductory rules. *The Psychological Record, 66*(2), 309-321. doi: 10.1007/s40732-016-0173-4

Hayes, S.C., Barnes-Holmes, D., & Roche, B. (2001). *Relational Frame Theory: A post Skinnerian account of human language and cognition.* New York, NY: Plenum.

Hughes, S., Hussey, I., Corrigan, B., Jolie, K., Murphy, C., and Barnes-Holmes, D. (2016). Faking revisited: Exerting strategic control over performance on the Implicit Relational Assessment Procedure. *European Journal of Social Psychology.* doi: 10.1002/ejsp.2207.

Gil, E., Luciano, C., Ruiz, F.J., Valdivia-Salas, S. (2012). A preliminary demonstration of transformation of functions through hierarchical relations. *International Journal of Psychology and Psychological Therapy, 12*(1), 1-19.

Green, G., Stromer, R., & Mackay, H.A. (1993). Relational learning in stimulus sequences. *The Psychological Record, 43*, 599-616.

Greenwald, A. G., Oakes, M. A., & Hoffman, H. G. (2003). Targets of discrimination: Effects of race on responses to weapons holders. *Journal of Experimental Psychology, 39*, 399-405.

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*(3), 157-162. doi: 10.1016/j.jcbs.2015.05.001

Kavanagh, D., Barnes-Holmes, Y., Barnes-Holmes, D., McEnteggart, C., & Finn, M. (2018). Exploring differential trial-type effects and the impact of a read-aloud procedure on deictic relational responding on the IRAP. *The Psychological Record, 68*(2), 163-176. doi: 10.1007/s40732-018-0276-1

Leslie, J.C., Tierney, K.J., Robinson, C.P., Keenan, M., Watt., A., & Barnes, D. (1993). Differences between clinically anxious and non-anxious subjects in a stimulus equivalence training task involving threat words. *The Psychological Record, 43,* 153-161.

McConahay, J. B. (1986). Modern racism, ambivalence, and modern racism scale. In J.F. Dovidio & S.L. Gaertner (Eds.), *Prejudice, discrimination and racism* (pp. 91-125). Orlando, FL: Academic Press.

McHugh, L., Barnes-Holmes, Y., & Barnes-Holmes, D. (2004). Perspective-taking as relational responding: A developmental profile. *The Psychological Record, 54*(1), 115-144.

McKenna, I. M., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2007). Testing the fake-ability of the Implicit Relational Assessment Procedure (IRAP): The first study. *International Journal of Psychology and Psychological Therapy, 7*, 253-268.

Merwin, I.M., & Wilson, K.G. (2005). Preliminary findings on the effects of self-referring and evaluative stimuli on stimulus equivalence class formation. *The Psychological Record, 55*, 561-575.

Nicholson, E., & Barnes-Holmes, D. (2012). The Implicit Relational Assessment Procedure (IRAP) as a measure of spider fear. *The Psychological Record, 62*, 263-277.

Power, P.M., Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y. (2017). Exploring racial bias in a country with a recent history of immigration of black Africans. *The Psychological Record*. doi: 10.1007/s40732-017-0223-6.

Roche, B. & Barnes, D. (1996). Arbitrarily applicable relational responding and sexual categorization: A critical test of the derived difference relation. *The Psychological Record, 46*, 451-475.

Steele, D.L. & Hayes, S.C. (1991). Stimulus equivalence and arbitrarily applicable relational responding. *Journal of the Experimental Analysis of Behavior, 56*, 519-555.

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.

Watt, A.W., Keenan, M., Barnes, D., & Cairns, E. (1991). Social categorization and stimulus equivalence. *The Psychological Record, 41,* 371-388.