Faking Revisited: Exerting Strategic Control over Performance on the Implicit Relational Assessment Procedure

Sean Hughes[2], Ian Hussey[2], Bethany Corrigan[1], Katie Jolie[1], Carol Murphy[1] and Dermot Barnes-Holmes[2]

[1]*National University of Ireland Maynooth*

[2]*Ghent University*

Author Note

Preparation of this paper was supported by a Government of Ireland Research Fellowship to SH and Grant BOF16/MET_V/002 from Ghent University to Jan De Houwer. SH, IH, and DBH, Department of Experimental Clinical and Health Psychology, Ghent University, Belgium. CM, BC, and KJ, Department of Psychology, Maynooth University, Ireland. Electronic mail should be sent to sean.hughes@ugent.be.

Abstract

Across four studies we demonstrate that effects obtained from the Implicit Relational Assessment Procedure (IRAP), like those obtained from other indirect procedures, are not impervious to strategic manipulation. In Experiment 1, we found that merely informing participants to 'fake' their performance without providing a concrete strategy to do so did not eliminate, reverse, or in any way alter the obtained outcomes. However, when those same instructions orientated attention towards the core parameters of the task, participants spontaneously derived a strategy that allowed them to eliminate their effects (Experiment 2). When participants were provided with a viable response strategy they successfully reversed the direction of their overall IRAP effect (Experiment 3). By refining the nature of those instructions we managed to target and alter individual trial-type effects in isolation with some success (Experiment 4).

Keywords: IRAP, Faking, Race, Sexual, Attitudes, Disgust

Faking Revisited: Exerting Strategic Control over Performance on the Implicit Relational Assessment Procedure

Self-report questionnaires constitute some of the most widely-used and versatile tools in the modern psychologist's armamentarium. Researchers from nearly every corner of the discipline draw upon these "direct" procedures in order to capture people's verbally-reported thoughts, feelings, and actions. However, if these tasks are to provide a valid index of the psychological phenomenon of interest then two basic pre-conditions must be met. First, people need to possess reliable introspective access to the behavior under investigation. Second, the outcomes obtained from those procedures should not be contaminated by strategic attempts to manipulate task performance. Unfortunately it appears that these two conditions are frequently violated – especially in socially sensitive domains. The fallibility of introspection coupled with the capacity to strategically distort task performance spurred the development of a new class of "indirect" procedures that many hoped would circumvent these methodological shortcomings. Examples include the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) as well as semantic and evaluative priming (Wittenbrink, Judd, & Park, 1997). These procedures are assumed to reduce one's ability to exert control over their behavior and to capture thoughts and feelings under the various conditions of automaticity (see Gawronski & De Houwer, 2014).

Over the past decade, however, researchers have begun to question the above assumptions. Although indirect procedures are certainly *less* sensitive to self-presentational biases compared to their direct counterparts they are far from immune to strategic manipulation or "faking". For example, making the IAT's procedural parameters apparent to the participant by providing them with simple or detailed instructions (Cvencek, Greenwald, Brown, Gray, &

Snowdon, 2010; Kim, 2003), prior experience (Steffens, 2004), or some combination of the two (Fiedler & Bluemke, 2005) increases their ability to exert control over the direction and magnitude of the IAT effect (Röhner, Schröder-Abé, & Schütz, 2011). Similar findings have also been obtained for variants of the IAT (Stieger, Göritz, Hergovich, & Voracek, 2011; Verschuere, Prati, & De Houwer, 2009) as well as other indirect procedures, such as evaluative priming (Teige-Mocigemba & Klauer, 2013), the AMP (Teige-Mocigemba, Penzl, Becker, Henn, & Klauer, 2015) and the Approach-Avoidance Task (AAT; Langer et al., 2010). Although researchers have recommended statistical indices to detect and correct for faking attempts (Agosta, Ghirardi, Zogmaister, Castiello, & Sartori, 2011; Cvencek et al., 2010) these algorithms are only partially successful, and have yet to be applied to other indirect procedures (Röhner, Schröder-Abé, & Schütz, 2013). Given the susceptibility of the IAT and priming to strategic manipulation, it seems important to determine if other indirect procedures are also sensitive to those same factors. If so, then initial assumptions about this class of measures need to be revised and greater steps taken to protect against self-presentation and impression management. If not, and a subset of indirect procedures are relatively more impervious to the aforementioned influences, then they could be deployed in domains where social-desirability concerns are especially problematic.

The Implicit Relational Assessment Procedure (IRAP)

Our goal in the current paper is to examine whether the outcomes obtained from an indirect procedure known as the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Barnes-Holmes, Bowles, & Stewart, 2010) are also susceptible to strategic manipulation. The IRAP sets out to examine the speed and accuracy with which people automatically *relate* rather than simply *categorize* stimuli. To illustrate, imagine that you are

interested in implicit self-esteem and decide to administer an IAT to a group of depressed and non-depressed participants. During a first test phase, participants categorize self-related items (e.g., their name) and positive words (e.g., HAPPY) using one response key and other-related items (e.g., another person's name) and negative words (e.g., INCOMPETENT) using a second response key. During another test phase, response mappings are reversed so that self-related items and negative words are assigned to the first key whereas other-related items and positive words are assigned to the second key. The difference in how well someone performs during the first relative to the second phase is considered to provide an overall measure of how readily they categorize self-related words with positive or negatively valenced adjectives. Critically, however, such an effect does not reveal how a person *relates* those concepts. For non-depressed individuals, it may be that the IAT effect reflects the extent to which they believe that they *are* good (actual self-esteem) whereas for their depressed counterparts the same score reflects how much they *want to be* good (ideal self-esteem). An indirect procedure that merely categorizes different classes of stimuli would not be able to distinguish between these two beliefs and instead would show evidence for positive self-evaluations in both cases (e.g., De Raedt, Schacht, Franck, & De Houwer, 2006).

In contrast, the IRAP allows one to capture beliefs that are emitted under the various conditions of automaticity. During each trial, the computer presents a label stimulus at the top of the screen (e.g., "*I am*" versus "*I am not*"), a target stimulus in the middle of the screen (e.g., a positive or negative adjective) and two response options at the bottom of the screen (e.g., "True" and "False"). By presenting specific combinations of label and target stimuli together, and by requiring a certain response to be emitted quickly and accurately, the task repeatedly exposes participants to four different types of trials (e.g., *I Am-Good*; *I Am-Bad*; *I Am Not-Good*; *I Am*

Not-Bad). These trials are grouped together into blocks. During one type of block participants are required to endorse the beliefs "*I am-Good*" and "*I am not-Bad*" while rejecting the beliefs "*I am-Bad*" and "*I am not-Good*". During a second type of block participants are required to do precisely the opposite, endorsing the latter beliefs ("*I am-Bad*" and "*I am not-Good*") while rejecting the former ("*I am-Good*" and "*I am not-Bad*"). The difference in time taken to respond to stimuli during these different blocks of trials – defined as the IRAP effect – indicates the strength or probability with which those stimuli are related. Although data from all trials can be combined to create an overall IRAP effect - indicating how quickly people tend to endorse one set of beliefs relative to another - most researchers tend to compute a separate IRAP effect for each of the four trial-types (see Barnes-Holmes et al. 2010). This enables them to independently examine, and subsequently compare, the speed and accuracy with which people endorse different beliefs, such as *I Am-Good, I Am-Bad, I Am not-Good* and *I Am not-Bad* (for more on the IRAP and its relationship to other indirect procedures see Gawronski & De Houwer, 2015; Nosek, Hawkins, & Frazier, 2011)[1].

At the time of writing, there were forty-one empirical publications about the IRAP distributed across fourteen peer-review journals (see Hughes & Barnes-Holmes, 2013 for a review of this work). Several points are worth noting at this juncture. First, the IRAP is similar to other indirect procedures insofar as it captures behaviors that are not typically accounted for by

[1] Note that a procedure is often described as being "direct" whenever participants are asked to self-assess the to-be-measured construct (e.g., confirm or reject a specific belief or attitude) and "indirect" whenever the construct is assessed indirectly on the basis of other behaviour (e.g., when the attitude or belief is inferred from reaction time performance in a speeded categorization task) (De Houwer, 2006). From this perspective the IRAP is relatively more direct than other tasks such as the IAT or AMP (given that participants are required to endorse or reject the relation between different stimuli). Yet the belief or attitude is still inferred indirectly by assessing its effect on task performance. In other words, it is not the participants' expressed attitudes or beliefs that are measured, but rather their average response latencies to tasks that require them to endorse or reject a relevant belief. Therefore it seems that, just like automaticity, the directness/indirectness of a procedure is not an "all-or-nothing" property. Rather procedures can be arranged along a continuum from relatively more indirect to relatively more direct depending on their specific properties. It seems that tasks like the IAT fall closer to the indirect end of the continuum, traditional self-report questionnaires fall closer to the direct end, and the IRAP falls somewhere in between (for a more detailed treatment of this topic see Barnes-Holmes et al., 2006).

self-report methodologies, especially in socially-sensitive domains (e.g., Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010; Roddy, Stewart, & Barnes-Holmes, 2011). Second, the behaviors captured by the IRAP often predict future behaviors and differentiate between groups in ways that self-reports do not. For instance, a recent meta-analysis of the predictive validity of IRAP effects found that it predicts similar outcomes to that of the IAT, and that both measures did so in ways that alluded self-report procedures (Vahey, Nicholson, & Barnes-Holmes, 2015).

Third, it is worth repeating that unlike virtually all other indirect procedures, the IRAP was designed to capture the extent to which stimuli are automatically related rather than simply categorized with one another. Although the speed and accuracy with which these relational responses are emitted could be explained by many different mental models, they are consistent with the idea that once propositional beliefs are acquired they can be activated automatically. Although many researchers take the position that human cognition is carved into two conceptually distinct mental systems (i.e., associative linking vs. propositional reasoning) that function under different operating conditions (i.e., automatic vs. controlled) (e.g., Gawronski & Bodenhausen, 2011), findings from the IRAP and other recently developed indirect procedures (e.g., De Houwer, Heider, Spruyt, Roets, & Hughes, 2015) offer another possibility: that humans can automatically relate stimuli in a wide number of ways and that the manner in which stimuli are related matters. Indeed, several researchers have found that relating the same stimuli in different ways (e.g., “I am good” vs. “I want to be good”; “I want unhealthy foods” vs. “Unhealthy foods make me hungry”) can lead to very different outcomes on the IRAP – outcomes that are often absent from self-report procedures, which predict future behavior, and differentiate between known groups (e.g., Carpenter, Martinez, Vadhan, Barnes-Holmes, &

Nunes, 2012; McKenna, Hughes, Barnes-Holmes, Yoder, & O'Shea, in press; Remue, Hughes, De Houwer, & De Raedt, 2014; Rönspies et al., 2015). In this sense the IRAP effect may represent one measure of implicit propositional knowledge (i.e., it captures the activation of propositions concerning how stimuli are related to one another under certain conditions of automaticity) (for a more detailed treatment of this topic see Hughes, Barnes-Holmes, & De Houwer, 2011; De Houwer, 2014; De Houwer et al., 2015)[2].

Faking the IRAP

Despite its growing popularity in social (Drake et al., 2015), cognitive (Remue et al., 2014) and clinical psychology (McKenna et al., in press), only a single published study has examined whether participants can strategically modify their performance on this task (McKenna, Barnes-Holmes, Barnes-Holmes, & Stewart, 2007). In their study, McKenna and colleagues exposed participants to a baseline IRAP in order to assess their automatic evaluations of pleasant and unpleasant words. Thereafter participants were divided into three groups and administered instructions that either: (a) described a strategy for reversing their previous IRAP effect, (b) asked participants to reverse their effect but provided no information on how to do so, or (c) provided an overview of the task (control). When participants repeated the IRAP for a second time the authors found that none of the above instructions reversed, or even attenuated effects.

[2] As we mentioned above, automatic (or implicit) is not an all-or-nothing concept but rather an umbrella term for a collection of operating conditions under which an assumed mental process operates (De Houwer et al., 2009). For instance, a measure may be implicit in the sense that the outcome is obtained even when participants have to respond quickly, without intention, awareness, or control. Different measures (IAT, AMP) can therefore vary in the extent to which they qualify as implicit (e.g., some may require speed and intention, while others are based on a lack of awareness or control). Although no systematic program of research has sought to determine to what extent the IRAP is implicit along each of these dimension, several studies show that it is implicit in the sense of being highly dependent on speeded performance (e.g., Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010), and that effects emerge in the absence of control (e.g., Dawson, Barnes-Holmes, Gresswell, Hart, & Gore, 2009; Carpenter et al., 2012). The extent to which IRAP effects also depend on intention and awareness are topics worthy of consideration.

While these findings are certainly encouraging, and support the notion that the IRAP is a procedure that is not readily amenable to faking, they were obtained with a relatively small sample, in a single, non-socially sensitive domain, where motivation to strategically modify one's performance was likely low. It is also important to note that the IRAP itself has undergone many changes over the past decade and that more recent iterations of the task significantly differ from those used by McKenna and colleagues in their faking study (see Hughes & Barnes-Holmes, 2013). It therefore seems prudent to re-evaluate their claims about the IRAPs resistance to faking with a version of the task that is currently used today.

Overview of the Current Research

In what follows we report a series of experiments that investigated whether instructions to reverse the direction and magnitude of IRAP effects would enable participants to manipulate how they respond towards members of other racial groups and sexual orientations, as well as towards clinically (disgust), and socially relevant (pleasant/unpleasant) stimuli. In each study, participants were exposed to a baseline IRAP, followed by a set of faking or control instructions. Thereafter they were administered the task again so that the impact of these instructions could be ascertained. We systematically varied the nature of the instructions across studies, from basic requests to alter task performance in the absence of a recommended strategy (*Experiment 1*), to those that allowed participants to derive such a strategy for themselves (*Experiment 2*). We also provided detailed instructions on how to reverse effects on all four IRAP trial-types (*Experiment 3*) or how to reverse a single trial-type effect while leaving the other three untouched (*Experiment 4*). By adopting this approach, we sought to determine what level of instructions (if any) are necessary to successfully eliminate or reverse IRAP effects.

EXPERIMENT 1

Experiment 1 sought to replicate the findings of McKenna et al. (2007). In particular, we were interested in whether participants could strategically alter their automatic evaluations of pleasant and unpleasant words in the absence of a recommended strategy for doing so.

Method

Participants

Sixty students at Ghent University (45 women), ranging in age from 18 to 59 years (M = 23.2, SD = 5.9) completed the study in exchange for €5 or course credit. Allocation to the faking and control conditions was counterbalanced across participants. Students reported that they had either complete no, or a single, IRAP prior to the study.

Materials

IRAP stimuli. Six positively (*happy, friendship, joy, peace, love, pleasure*) and six negatively valenced items (*Hitler, pedophile, cancer, incest, murder, suicide*) served as label stimuli during the IRAP (positive: M = 6.42, SD = 0.76; negative: M = 1.24, SD = 0.46). These items were selected from a large pool of Dutch words whose valence had previously been normed using a scale ranging from 1 (very negative) to 7 (very positive) (Moors et al., 2013). Six positive (*good, pleasant, fun, positive, fantastic, excellent*) and six negative adjectives (*bad, unpleasant, nasty, negative, horrible, terrible*) were used as target stimuli while 'Same' and 'Opposite' served as the two response options.

Procedure

Upon arriving at the laboratory participants were welcomed by the researcher, seated in front of a computer, and provided with a brief description of the procedures they would subsequently encounter. Once they had provided their informed consent, they were exposed to a

pre-instructions IRAP, a set of control or faking instructions, followed by a post-instructions IRAP. A similar experimental sequence was adopted in Experiments 1-4.

Pre-Instructions IRAP. The IRAP consisted of a minimum of one and a maximum of three pairs of practice blocks followed by a fixed set of three pairs of test blocks. Each block consisted of twenty four trials that presented either a positive or negative label stimulus at the top of the screen, a positive or negative target stimulus in the middle of the screen and two relational response options ('Same' or 'Opposite') at the bottom of the screen. In this way the IRAP was comprised of four different trial-types: *Positive-Positive*; *Negative-Negative*, *Positive-Negative* and *Negative-Positive* (see Figure 1). The presentation of trials was varied in a quasi-random order such that each trial-type appeared an equal number of times within every block. The allocation of the two response options to the left or right side of the screen was fixed across successive trials.

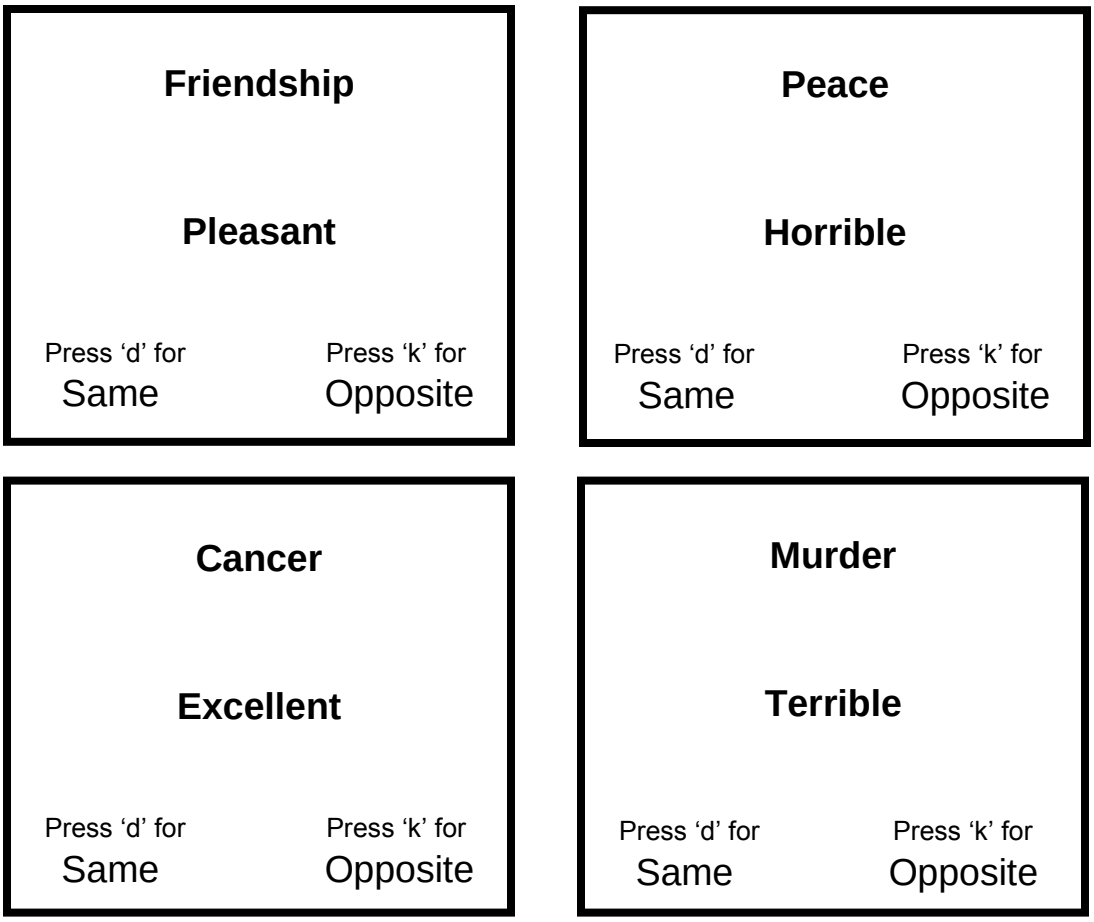

Figure 1. Examples of the four IRAP trial-types used in Experiment 1. A valenced label stimulus appeared at the top of the screen (e.g., '*Friendship*' or '*Murder*'), along with a valenced target stimulus in the middle of the screen (e.g., '*Pleasant*' or '*Terrible*') and two relational response options ('*Same*' and '*Opposite*') at the bottom of the screen.

Prior to the task participants were informed that a number of words would appear onscreen and that they would have to relate those words based on one of two responses rules: either Rule A ("*Please act as if good words are good and bad words are bad*") or Rule B ("*Please act as if good words are bad and bad words are good*"). Participants were required to respond in a manner that was consistent with the response rule that applied to that given block of trials. These two rules for responding were alternated across successive blocks, resulting in three pairs of IRAP test blocks (Block 1-Rule A; Block 2-Rule B; Block 3-Rule A; Block 4-Rule B; Block 5-Rule A; Block 6-Rule B).

The IRAP commenced with a pair of practice blocks. Participants progressed from the practice to the test blocks whenever they responded with a pre-defined accuracy (at least 80% accuracy) and speed (median latency of less than 2000ms) on a successive pair of practice blocks. Failure to meet these criteria resulted in re-exposure to another pair of practice blocks until participants either achieved those criteria or a maximum of three pairs of practice blocks were completed. If the above criteria were met then a fixed set of three test block pairs were administered. If not, then participants were thanked, debriefed and dismissed.

Faking instructions. Following the first IRAP a set of instructions appeared onscreen, the content of which differed for those in the control and faking conditions. Participants in the faking condition were informed that they would complete a similar task as before, but this time their goal was to "*strategically alter their performance in such a way that 'fooled' or 'tricked' the computer into thinking that you considered good words to be bad and bad words to be good. That is, you should respond in a way that would lead the researcher to think that you find words like 'murder' and 'incest' to be good and other words such as 'love' and 'peace' to be bad.*" These instructions also emphasized that participants would still need to meet speed and accuracy criteria during each block of trials. The researcher then checked that the participants understood the above instructions and provided corrective feedback where necessary.

Participants in the control condition were also provided with instructions that were matched in length but which differed in content. These instructions provided background information about the procedure and suggested that they would need to complete the task once more so that its test-retest reliability could be ascertained. Any questions were subsequently addressed by the researcher, after which, the post-instructions IRAP was initiated. This second IRAP was identical in all respects to the baseline measure.

Results

Analytic Strategy

To determine whether IRAP performance differed as a function of instructions received (independent variable), a series of repeated measures ANOVAs and post-hoc tests were carried out on the data from the IRAP's four trial-types (dependent variable).

Data Preparation

Participant exclusion. Two participants failed to meet the accuracy and latency criteria during both the pre- and post-instructions IRAPs. Another participant failed the pre-instructions IRAP while five failed the post-instructions IRAP. The data from these eight individuals were removed prior to analysis[3].

IRAP scoring. The primary datum 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. To minimize contamination by individual differences associated with age, motor skills, and/or cognitive ability, response latency data were transformed into difference (*D*) scores using an adaptation of Greenwald, Nosek and Banaji's (2003) *D* algorithm (see Appendix C for a detailed overview of the steps involved in calculating *D*-IRAP scores). Four *D* scores were calculated for each participant, one for each of the trial-types that comprised the task (i.e., *Positive-Positive*; *Negative-Negative*; *Positive-Negative*; *Negative-Positive*). Positive scores indicated that participants were quicker to endorse the belief that positive stimuli were positive and reject the belief that they were negative. Negative scores indicated that participants were quicker to endorse the belief that negative stimuli were negative

[3] In-line with previous work (Nicholson & Barnes-Holmes, 2012) whenever participants failed to maintain accuracy criterion on one of the six test blocks all the data from that test block pair was excluded and analyses conducted on the remaining two test block pairs. In Experiment 1 this was the case for six participants in the faking condition and eight participants in the control condition.

and reject the belief that they were positive. Neutral scores indicated the absence of either response bias. Finally, we calculated a change score for each of the four trial-types by subtracting pre- from post-instructions *D* scores. This allowed us to examine whether task performance varied from one IRAP to the other. Positive change scores indicate an increase in the magnitude of IRAP effects across test sessions while negative scores indicate the opposite.

Hypothesis Testing

Pre- and Post-Instruction IRAPs. We expected participants to automatically endorse the belief that positive stimuli are positive and reject the belief that positive stimuli are negative (i.e., produce a positive effect on the *Positive-Positive* and *Positive-Negative* trial-types). We also expected participants to endorse the belief that negative stimuli are negative and reject the belief that negative stimuli are positive (i.e., produce a negative effect on the *Negative-Negative* and *Negative-Positive* trial-types).

Submitting *D* scores to a 2 (*Instruction Type*: faking vs. control) x 2 (*Test Time*: pre vs. post instructions IRAP) x 4 (*Trial-Type*) mixed-models ANOVA revealed a main effect for Trial-Type, $F(3, 50) = 31.44$, $p < .001$, $\eta^2_{partial} = .39$, as well as a two-way interaction between Trial-Type and Time, $F(3, 50) = 9.15$, $p < .001$, $\eta^2_{partial} = .16$ (no main or interaction effects emerged for Instruction Type). During the pre-instructions IRAP, participants showed the expected effect on all four trial-types, with follow-up, one-sample *t*-tests indicating that these effects significantly differed from zero (*ps* < .002), with the exception of the *Negative-Negative* trial-type, $t(51) = 1.25$, $p = .22$. During the post-instructions IRAP, participants continued to endorse the belief that positive words were positive (*Positive-Positive*), $t(51) = 7.48$, $p < .001$, and reject the belief that positive words were negative (*Positive-Negative*) trial-types, $t(51) = 2.77$, $p < .01$. However, there was no evidence for an effect on either the *Negative-Positive* (*p* =

.49) or *Negative-Negative* trial-types (p = .57) indicating that certain IRAP effects diminished in magnitude across repeated test administrations (see Table 1).

Change from Pre- to Post-Instruction IRAP. Submitting change scores to a 2 (*Instruction Type*) x 4 (*Trial-Type*) repeated measures ANOVA revealed that trial-type effects for those in the faking and control conditions did not differ significantly across the two test sessions (all *ps* > .3). When data from all participants was considered, effects on the *Positive-Positive*, $t(51) = 3.03$, p = .004, and *Negative-Positive* trial-types, $t(51) = 3.54$, p = .001, were found to significantly attenuate from one test session to the next.

IRAP Reliability

To assess the internal consistency of the IRAP, four split-half reliability scores were calculated, one for each of the four trial-types. These scores was generated by applying the *D* algorithm separately to odd and even numbered trials. With respect to the pre-instruction IRAP, split-half correlations between odd and even scores, applying Spearman-Brown corrections, were as follows: *Positive-Positive*, r = .40, *Positive-Negative*, r = .24, *Negative-Positive*, r = .34, and *Negative-Negative*, r = .58. Internal consistency of the post-instructions IRAP was as follows: *Positive-Positive*, r = .34, *Positive-Negative*, r = .47, *Negative-Positive*, r = .19, and *Negative-Negative*, r = .39.

Discussion

In-line with McKenna et al. (2007) we found that instructing participants to fake their evaluations of positive and negative adjectives – without providing a viable strategy to do so – did not eliminate or reverse their IRAP effects. During the pre-instructions IRAP participants responded in-line with our expectations – automatically endorsing the belief that positive words were positive and that negative words were negative while simultaneously rejecting the belief

that negative words were positive. During the post-instructions IRAP participants continued to respond as they had before. However, the magnitude of their effects diminished across repeated measurement, such that they continued to produce significant effects on the *Positive-Positive* and *Positive-Negative* trial-types but showed no evidence for effects on the *Negative-Negative* or *Negative-Positive* trial-types. There was no evidence to suggest that these attenuated effects were due to the type of instructions participants received.

EXPERIMENT 2

Only two studies have explored the fakeability of the IRAP (Experiment 1 and McKenna et al., 2007) and both have relied on a similar set of (generic) valenced stimuli. Therefore, in Experiment 2, we sought to rule out the possibility that the IRAP's insensitivity to manipulation was due to the specific stimuli employed or the domain tested. We set evaluations of generic stimuli to the side and instead focused our attention on a clinically-relevant domain (disgust) which has previously been shown to produce robust IRAP effects (Nicholson & Barnes-Holmes, 2012). At the same time we also modified the faking instructions so that they now orientated participants towards those task parameters that would allow them to derive a viable response strategy for themselves.

Method

Participants

Sixty students at Ghent University (43 women), ranging in age from 18 to 47 years (M = 22.7, SD = 5.3) completed the study in exchange for €5 or course credit. The order of IRAP test blocks as well as assignment of participants to the faking or control conditions was counterbalanced across participants. Students reported that they had completed either no, or a single, IRAP prior to the study.

Measures

IRAP stimuli. The task was identical to that used in Experiment 1 with the exception of the stimuli employed. There were a total of twelve label stimuli; six of which involved a negative appraisal of a target ('*I find it horrible*', '*I think it is disgusting*', '*It looks nasty*', '*I feel revolted*', '*It makes me sick*', and '*I am repulsed*') and six which involved a positive appraisal of a target ('*I think it's pleasant*', '*I find it nice*', '*It's good*', '*It makes me feel happy*', '*It look's nice*', '*It makes me feel great*'). Six color photographs of items designed to evoke disgust (rotten meat, a large maggot, bloody hand, diseased mouth cavity, toilet with feces, burnt face) and six pleasant images (baby, puppies, nature scenes, kittens, bunnies) were selected variously from the Internet and from the International Affective Picture System (IAPs: Lang, Bradley, & Cuthbert, 2008) to serve as target stimuli. Participants rated the six disgusting items (M = 2.02, SD = 1.09) and six non-disgusting items (M = 6.57, SD = .83) using a scale ranging from 1 (disgusting) to 7 (not disgusting) with 4 as a neutral point. The words 'True' and 'False' served as two response options while the rules for responding were Rule A ("*Please answer AS IF disgusting things are disgusting*") and Rule B ("*Please answer AS IF disgusting things are pleasant*"). This led to the following four trial-types: *Pleasant-Positive*, *Pleasant-Negative*, *Disgusting-Positive*, *Disgusting-Negative*. Finally, the response options were varied in a quasi-random order within each block.

Procedure

Faking Instructions. Similar to Study 1 participants were allocated to either a control or faking instruction condition. This time however, the faking condition was provided with a more elaborate set of instructions that orientated attention towards those properties of the procedure (i.e., speeded responding) that would facilitate strategic manipulation of task performance.

Specifically, participants were informed that "*In the next part of the experiment you will have to try to trick the computer into thinking that you like disgusting things and dislike pleasant things. You should respond in a way that would lead the researcher to think that you find disgusting pictures to be pleasant and pleasant pictures to disgusting. I can't provide specific instructions on how you can do this, but the only way to fake your performance is to figure out how the task works. Pay attention to how you respond during the task in order to figure out how you can fake it."* The researcher subsequently checked that participants understood the above instructions and provided corrective feedback where necessary.

Results

Analytic Strategy

The analytic strategy adopted here was similar to that in Experiment 1.

Data Preparation

Participant exclusion. Four participants failed to meet the mastery criteria during both the pre- and post-instructions IRAPs. Another two failed the pre-instructions IRAP while seven failed the post-instructions IRAP. The data from these thirteen individuals was removed prior to analysis. Of the remaining participants nine in the faking condition and seven in the control condition failed to maintain accuracy criterion on one of the six test blocks. For these participants *D* scores were calculated based on the remaining two test block pairs.

IRAP scoring. *D* scores for each of the four trial-types (*Pleasant-Positive*, *Pleasant-Negative*, *Disgusting-Positive*, *Disgusting-Negative*) were calculated in a similar manner as Experiment 1. Positive scores indicated that participants were quicker to endorse the belief that stimuli were pleasant and reject the belief that they were disgusting. Negative scores indicated that participants were quicker to endorse the belief that stimuli were disgusting and reject the

belief that they were positive. Neutral scores indicated the absence of either response bias. Finally, change scores were also calculated as in Experiment 1.

Hypothesis Testing

Pre- and Post-Instruction IRAPs. Based on previous work (Nicholson & Barnes-Holmes, 2012) we expected participants to endorse the belief that pleasant images were positive and reject the belief that those same images were negative (i.e., produce positive effects on the *Pleasant-Positive* and *Pleasant-Negative* trial-types). We also expected them to endorse the belief that disgusting images were negative and reject the belief that those same images were positive (i.e., to produce negative effects on the *Disgusting-Positive* and *Disgusting-Negative* trial-types). Submitting *D* scores to a 4 (*Trial-Type*) x 2 (*Test Time*) x 2 (*Instruction Type*) mixed-models ANOVA revealed a main effect for Trial-Type, $F(3, 45) = 27.21$, $p < .001$, $\eta^2_{partial} = .38$, a two-way interaction between Trial-Type and Time, $F(3, 45) = 11.37$, $p < .001$, $\eta^2_{partial} = .20$, and a three-way interaction between Trial-Type, Time, and Instruction Type, $F(3, 45) = 10.29$, $p < .001$, $\eta^2_{partial} = .19$. To qualify this three-way interaction we explored the impact of Time and Trial-Type separately for the faking and control conditions.

With respect to the control group, analyses revealed a main effect of Trial-Type, $F(3, 24) = 31.12$, $p < .001$, $\eta^2_{partial} = .57$, but no main or interaction effects with Time. In other words, participants showed the expected effects on all four trial types during the pre-and post-instructions IRAPs, with follow-up one-sample *t*-tests indicating that these effects differed significantly from zero (*ps* < .03) (with one exception: the *Disgusting-Positive* trial-type failed to reach significance during the post-instructions IRAP, $t(24) = 0.46$, $p = .65$). A very different picture emerged for those in the faking condition. Analyses revealed a main effect for Trial-Type, $F(3, 21) = 6.21$, $p = .001$, $\eta^2_{partial} = .23$, as well as a two-way interaction between Trial-

Type and Time, $F(3, 21) = 13.61$, $p < .001$, $\eta^2_{partial} = .39$. During the pre-instructions IRAP, participants showed the expected effect on all four trial-types, with follow-up *t*-tests indicating that these effects significantly differed from zero (*ps* < .001), with the exception of the *Disgusting-Positive* trial-type, $t(21) = 1.47$, $p = .16$. Yet during the post-instructions IRAP none of the effects differed significantly from zero with the exception of the *Disgusting-Positive* trial-type, $t(21) = 2.45$, $p = .02$, which was now in precisely the opposite direction as before (see Table 2).

Change from Pre- to Post-Instruction IRAP. When changes scores were submitted to a 2 (*Instruction Type*) x 4 (*Trial-Type*) repeated measures ANOVA a main effect emerged for Instruction Type, $F(1, 45) = 14.29$, $p < .001$, $\eta^2_{partial} = .24$. Participants in the control condition did not differ in how they responded on the *Pleasant-Positive* ($M = .04$, $SD = .47$), *Pleasant-Negative* ($M = -.06$, $SD = .43$), *Disgusting-Positive* ($M = .15$, $SD = .39$) or *Disgusting-Negative* trial-types ($M = .09$, $SD = .56$) from one IRAP to the next (all *ps* > .07). In contrast, the magnitude of the *Pleasant-Positive* ($M = -.51$, $SD = .69$), $t(21) = 3.44$, $p = .002$, *Pleasant-Negative* ($M = -.46$, $SD = .52$), $t(21) = 4.14$, $p < .001$, *Disgusting-Positive* ($M = .46$, $SD = .67$), $t(21) = 3.23$, $p = .004$, and *Disgusting-Negative* trial-types ($M = .55$, $SD = .77$), $t(21) = 3.35$, $p = .003$, significantly changed once participants received instructions to fake their performance.

IRAP Reliability

With respect to the pre-instruction IRAP, internal consistency for the four trial-types was as follows: *Positive-Positive*, $r = .38$, *Positive-Negative*, $r = .44$, *Negative-Positive*, $r = .62$, and *Negative-Negative*, $r = .53$. Internal consistency of the post-instructions IRAP was as follows: *Positive-Positive*, $r = .72$, *Positive-Negative*, $r = .71$, *Negative-Positive*, $r = .51$, and *Negative-Negative*, $r = .62$.

Discussion

In Experiment 2 we measured implicit beliefs about disgusting and pleasant stimuli on two occasions. In between test administrations we provided half of the participants with simple faking instructions that orientated their attention towards those properties of the procedure that would enable them to manipulate their effects (the other half received a set of control instructions). Participants responded in-line with our expectations. During the pre-instructions IRAP they endorsed the belief that pleasant images are positive and disgusting images are negative while simultaneously rejecting the belief that pleasant images are negative and disgusting images are positive. During the post-instructions IRAP participants in the control condition continued to respond in a similar way. Critically, however, their counterparts in the faking condition successfully eliminated all traces of their beliefs in the second IRAP. Thus it seems that providing instructions to fake without highlighting a viable strategy to do so only allows participants to dismantle rather than reverse the direction of their IRAP effects.

EXPERIMENT 3

Whereas the previous two studies focused on automatic evaluations of a non-socially sensitive nature it seems reasonable to assume that motivation to strategically manipulate one's performance may be relatively higher in socially sensitive situations. Indeed, indirect procedures are typically deployed under the assumption that motivations to bias one's performance will ultimately meet with failure even in socially-sensitive contexts (although see Fiedler & Bluemke, 2005). Demonstrating that IRAP effects can be strategically altered in such situations would provide even stronger evidence for the fakeability of the measure. Therefore in Experiment 3 we sought to replicate and extend our previous results into a third domain (implicit race evaluations). Half of the participants were provided with a similar set of 'orientating' faking

instructions as in Experiment 2 with the assumption that this would serve to eliminate (rather than reverse) their IRAP effects. The other half were given a detailed set of faking instructions which clearly stated how they could successfully manipulate the direction of their effects.

Method

Participants

Forty nine students at an Irish university completed the study in exchange for a candy bar. The majority of those who provided gender information (n = 40) self-identified as female (65%), while the total sample ranged in age from 18 to 52 (M = 25.33, SD = 9.13). The order of IRAP test blocks as well as assignment to the simple or detailed faking conditions was counterbalanced across participants. Students reported that they had completed between 0 and 15 IRAPs prior to the study. It is worth noting, however, that median IRAP experience was 0.

Measures

IRAP stimuli. Eight label stimuli were used; four negative ('*Dangerous*', '*Aggressive*', '*Rude*' and '*Violent*') and four positive adjectives ('*Safe*', '*Friendly*', '*Polite*' and '*Kind*'), that were each presented twice during a block of IRAP trials. Eight color photographs of black individuals (four male and four female) as well as eight images of white individuals (four male and four female) were obtained from the CAL/PAL Face Database (Minear & Park, 2004) and served as target stimuli. The words "True" and "False" served as the two response options and the response rules were as follows: Rule A ("*Please answer AS IF Black people are dangerous and White people are safe*") and Rule B ("*Please answer AS IF White people are dangerous and Black people are safe*"). This led to the following four trial-types: *Black-Positive*, *Black-Negative*, *White-Positive*, and *White-Negative*. Finally, response options were varied in a quasi-random order within each block.

Procedure

Faking Instructions. The simple faking instructions were identical to those used in Experiment 2. The detailed faking instructions indicated that participants should respond quickly on certain blocks and slowly on others, while ensuring that they still met the overall latency/accuracy criteria (see Appendix B). In addition, a short message was provided before each block that reminded participants that they should respond either quickly ("***Note: please respond quickly***") or slowly ("***Note: please respond slowly***") during the subsequent block of trials.

Results

Analytic Strategy

The analytic strategy adopted here was similar to that in Experiments 1-2.

Data Preparation

Participant exclusion. Five participants failed to pass both IRAPs while an additional participant failed the post-instructions IRAP. Data for these six participants was discarded prior to analysis. Of the remaining participants five in the simple "orientating" faking condition and four in the detailed faking condition failed to maintain accuracy criterion on one of the six test blocks. For these participants *D* scores were calculated based on the remaining two test block pairs.

IRAP scoring. *D* scores for each of the four trial-types (*Black-Positive*, *Black-Negative*, *White-Positive*, and *White-Negative*) were calculated in a similar manner as Experiments 1-2. Positive scores indicate an endorsement of the belief that a racial group is positive or a rejection of the belief that they are negative. Negative scores indicate an endorsement of the belief that a

group is negative or a rejection of the belief that they are positive. Neutral scores indicate a lack of a racial bias. Change scores were also calculated as in Experiments 1-2.

Hypothesis Testing

Pre- and Post-Instruction IRAPs. Based on previous findings (e.g., Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010) we expected participants to endorse the belief that White people and Black people are positive, and reject the belief that White people were negative (i.e., produce positive effects on the *White-Positive*, *White-Negative,* and *Black-Positive* trial-types). We also expected them to endorse the belief that Black people are negative (produce a negative effect on the *Black-Negative* trial-type). Submitting *D* scores to a 4 (*Trial-Type*) x 2 (*Test Time*) x 2 (*Instruction Type*) mixed-models ANOVA revealed a main effect for Trial-Type, $F(3, 41) = 25.69$, $p < .001$, $\eta^2_{partial} = .39$, and Test Time, $F(1, 41) = 4.05$, $p = .05$, $\eta^2_{partial} = .09$, a two-way interaction between Trial-Type and Instruction, $F(3, 41) = 21.22$, $p < .001$, $\eta^2_{partial} = .34$, Trial-Type and Test Time, $F(3, 41) = 53.32$, $p < .001$, $\eta^2_{partial} = .57$, as well as a three-way interaction between Trial-Type, Test Time and Instruction Type, $F(3, 41) = 31.75$, $p < .001$, $\eta^2_{partial} = .44$. To qualify this three way interaction, the impact of Test Time and Trial-Type was examined separately for the simple and detailed faking conditions.

With respect to the simple "orientating" faking condition, analyses revealed a marginally significant effect for Trial-Type, $F(3, 19) = 2.53$, $p = .07$, $\eta^2_{partial} = .12$, but no main or interaction effect for Test Time. During the pre-instructions IRAP participants were quicker to endorse the belief that White people were positive (*White-Positive*), $t(19) = 4.12$, $p = .001$, and that Black people were positive (*Black-Positive*), $t(19) = 3.01$, $p = .007$ (however they showed no effects on the *White-Negative* or *Black Negative* trial-types; both *ps* > .07). During the post-instructions IRAP, and similar to Experiment 2, participants showed no evidence for any effect with the

exception of the *Black Positive* trial-type, $t(19) = 3.62$, $p = .002$ (all other *ps* > .38). In contrast, the detailed faking condition showed evidence of a main effect for Trial-Type, $F(3, 22) = 39.65$, $p < .001$, $\eta^2_{partial} = .64$, as well as a two-way interaction between Trial-Type and Test Time, $F(3, 22) = 74.55$, $p < .001$, $\eta^2_{partial} = .77$. When these participants completed the pre-instructions IRAP they strongly endorsed the belief that White people are positive (*White-Positive*), $t(22) = 3.75$, $p = .001$ (all other *ps* > .24). Yet when they completed the post-instructions IRAP they responded in-line with the faking instructions received. Specifically, they now endorsed the belief that White people were negative (*White-Negative*), $t(22) = 6.89$, $p < .001$, and rejected the belief that White people were positive (*White-Positive*), $t(22) = 8.09$, $p < .001$. They were also quicker to endorse the belief that Black people were positive (*Black-Positive*), $t(22) = 9.68$, $p < .001$, and reject the belief that Black people were negative (*Black-Negative*), $t(22) = 6.89$, $p < .001$ (see Table 3).

Change from Pre- to Post-Instruction IRAP. When changes scores were submitted to a 2 (*Instruction Type*) x 4 (*Trial-Type*) repeated measures ANOVA a main effect for Trial-Type, $F(3, 41) = 53.25$, $p < .001$, $\eta^2_{partial} = .57$, as well as a two-way interaction between Trial-Type and Instruction Type was observed, $F(3, 41) = 31.72$, $p < .001$, $\eta^2_{partial} = .44$. For participants in the simple "orientating" faking condition the *White-Positive* trial-type effect diminished in magnitude across the two IRAPs ($M = -.22$, $SD = .46$), $t(19) = 2.08$, $p = .05$, whereas the *White-Negative* ($M = -.19$, $SD = .58$), *Black-Positive* ($M = .10$, $SD = .49$) and *Black-Negative* ($M = .07$, $SD = .57$) effects did not differ from one test administration to the next (all *ps* > .16). In contrast, their counterparts in the detailed faking condition significantly reversed the direction of their IRAP effects in-line with the instructions provided. Specifically, they showed a reversed effect on the *White-Positive* ($M = -1.33$, $SD = .77$), $t(22) = 8.29$, $p < .001$, *White-Negative* ($M = -.97$,

SD = .58), $t(22) = 8.01$, $p < .001$, *Black-Positive* (*M* = 1.06, *SD* = .66), $t(22) = 7.66$, $p < .001$, and *Black-Negative* trial-types (*M* = 1.02, *SD* = .52), $t(22) = 9.47$, $p < .001$[4].

IRAP Reliability

With respect to the pre-instruction IRAP, internal consistency for the four trial-types was as follows: *White-Positive*, *r* = .05, *White-Negative*, *r* = .66, *Black-Positive*, *r* = .04, and *Black-Negative*, *r* = .18. Internal consistency of the post-instructions IRAP was as follows: *White-Positive*, *r* = .91, *White-Negative*, *r* = .91, *Black-Positive*, *r* = .89, and *Black-Negative*, *r* = .83.

Discussion

In Experiment 3 we provided half of the participants with faking instructions that orientated them towards those task properties which would allow them to generate a faking strategy for themselves. Similar to Experiment 2 we found that doing so helped people to eliminate rather than reverse the direction of their IRAP effects. The other half of the sample were provided with a concrete faking strategy which was then reiterated from block to block. This resulted in strong evidence for faking. During the pre-instructions IRAP participants strongly endorsed the belief that White people are positive and showed no other IRAP effects. Yet during the post-instructions IRAP they responded in precisely the opposite way: now endorsing the belief that White people are negative and that Black people are positive while simultaneously rejecting the belief that White people are positive and Black people are negative.

EXPERIMENT 4

In the previous experiment we provided participants with faking instructions that allowed them to manipulate their task performance in an effective but rather unsophisticated way. By informing them to speed up on certain blocks and slow down on others we provided a strategy

[4] Given that participants varied in their prior IRAP experience we re-ran the above analyses while including that experience as a covariate. Results indicated that there was still significant main and interaction effects for, and between, test time and instruction type (even after IRAP experience was controlled for).

that influenced their performance on all four trial-types. However, a more convincing and difficult to detect form of manipulation would involve strategically altering performance on individual trial-types. This may play an important role in psychologically sensitive and applied domains. For instance, child sexual offenders might be motivated to alter their responses on *Child-Sexual* rather than *Adult-Sexual* trial-types when sexual preferences are being assessed (e.g., Dawson et al., 2009). Likewise, those looking to present themselves in an egalitarian light might attempt to manipulate their performance on the *Black-Negative* rather than *White-Positive* trial-types. Experiment 4 set out to determine if this subtle form of manipulation is actually possible in the context of implicit beliefs about sexuality. Specifically, we asked a group of male students who self-identified as homosexual to mimic how their heterosexual counterparts respond on the IRAP and vice-versa. In each case a refined set of instructions indicated how students could (a) reverse the direction of one trial-type effect, (b) attenuate the effect on a second trial-type, while (c) leaving the remaining two trial-types untouched. If participants can exert sophisticated control over their implicit sexual beliefs, then we would expect to see a reversed effect on one, an attenuated effect on another, and no change across repeated test administrations on the final two trial-types.

Method

Participants

Forty one male students at an Irish university were recruited on the basis of their self-reported sexuality. Twenty participants, ranging in age from 18 to 51 years (M = 24.95, SD = 9.24) self-identified as heterosexual while another twenty one, ranging from 18 to 34 years (M = 20.95, SD = 3.34) self-identified as homosexual. Heterosexual men were operationally defined as those with a score between 1 and 3 on the Klein Sexual Orientation Grid (KSOG; Klein, 1993)

while homosexual men were defined as those with a score between 5 and 7 on that same scale. The order of IRAP test blocks was counterbalanced across participants. Students reported that they had completed between 0 to 6 IRAPs prior to the study. It is worth noting, however, that median IRAP experience was a single prior exposure.

Measures

IRAP stimuli. Ten label stimuli were used; five 'sexually attractive' terms (*'Arousing, 'Erotic', 'Attractive', 'Sensual', and 'Exciting'*) and five 'sexually unattractive' terms ('Awful', 'Repulsive', 'Repelling', 'Repugnant', and 'Repellent'). Five color images of nude males (4460, 4500, 4534, 4550, 4561) as well as five images of nude females (4141, 4142, 4210, 4240, 4332) were taken from the IAPS and served as target stimuli. The words "True" and "False" served as two response options while the following response rules were used: Rule A ("*Please answer AS IF Women are attractive and Men are unattractive*") and Rule B ("*Please answer AS IF Men are attractive and Women are unattractive*"). This led to the following four trial-types: *Men-Attractive, Men-Unattractive, Women-Attractive, Women-Unattractive*. Finally, the location of response options was fixed within each block of trials.

Klein Sexual Orientation Grid. The current study used the KSOG to gather information about seven different aspects of sexual orientation; 'sexual attraction', 'sexual behavior', 'sexual fantasies', 'emotional preference', 'social preference', 'hetero/gay lifestyle' and 'self-identification'. Participants are asked to respond to items from the first five dimensions using a scale ranging from 1 (other-sex only) to 7 (same-sex only), with a midpoint of 4 (both sexes equally). They were also asked to respond to items from the final two dimensions using a scale ranging from 1 (Heterosexual only) to 7 (Gay/Lesbian only) with a midpoint of 4 (Hetero/Gay-Lesbian equally). Although responses are typically assessed across three different time periods

(past, present, and ideal) we only focused on a single temporal period ('present') given the methodological purposes of the current study.

Procedure

Faking Instructions. Participants were administered a detailed set of instructions which provided them with information on how to strategically alter their performance on specific IRAP trial-types. These instructions indicated that homosexual males should respond like their heterosexual counterparts and vice-versa. For instance, heterosexual participants were asked to "*respond as if they found images of naked men attractive and naked women unattractive*". To help them do so, instructions prior to Rule A blocks indicated that they should respond slowly on the *Men-Attractive* and *Women-Attractive* trial-types whereas instructions before the Rule B block indicated that they should respond slowly during *Women-Attractive* and quickly during *Men-Attractive* trial-types. Note homosexual participants received a comparable but modified set of instructions that would enable them to respond as a heterosexual male presumably would (see Appendix B).

Results

Analytic Strategy

The analytic strategy adopted here was similar to that in Experiments 1-3.

Data Preparation

Participant exclusion. Four participants failed both IRAPs while another six failed the post-instructions IRAP. Data for these ten participants were discarded prior to analysis. Of the remaining participants two students who self-identified as homosexual failed to maintain accuracy criterion on one of the six test blocks. Their *D* scores were calculated based on the remaining two test block pairs.

IRAP. *D* scores for each of the four trial-types (*Men-Attractive*, *Men-Unattractive*, *Women-Attractive*, *Women-Unattractive*) were calculated in a similar manner as Experiments 1-3. Positive scores indicate an endorsement of the belief that a particular gender is attractive and a rejection of that gender as unattractive. Negative scores indicate an endorsement of the belief that a gender is unattractive and a rejection of the belief that they are attractive. Neutral scores indicate a lack of either response bias. Change scores were calculated as in Experiments 1-3.

KSOG. The data from the 'emotional preference' and 'social preference' dimensions were not included given that they are thought to measure something other than sexual orientation (see Weinrich, Snyder, Pillard, Grant, Jacobson, Robinson, & McCutchan, 1993). Scores from the other five dimensions were averaged to create a single mean index of sexual orientation, with lower values (1-3) reflecting an overall preference for the other sex, higher values (5-7) an overall preference for the same sex while a score of 4 indicated a lack of preference for either sex. Submitting these scores to a one-way ANOVA produced a pattern of findings that were consistent with participants self-reported sexual orientation, $F(1, 28) = 299.59$, $p < .001$, $\eta^2_{partial} = .92$, with heterosexuals students scoring between 1 and 3 ($M = 1.41$, $SD = 0.37$) and homosexual students scoring between 5 and 7 ($M = 5.9$, $SD = 0.93$). Note that no participant obtained a score of 4.

Hypothesis Testing

Pre- and Post-Instruction IRAPs. Based on previous work on implicit beliefs about sexuality (e.g., Rönspies et al., 2015) we expected heterosexual students to endorse the belief that women are attractive and men are unattractive while rejecting the belief that women are unattractive and male are attractive (i.e., produce positive effects on the *Women-Attractive* and *Women-Unattractive* and negative effects on the *Men-Attractive* and *Men-Unattractive* trial-

types). We expected a different pattern of responding for their homosexual counterparts. That is, they should endorse the belief that men are attractive and reject the belief that men are unattractive (i.e., produce positive effects on the *Men-Attractive* and *Men-Unattractive* trial-types). However, we did not expect them to produce any effects on the *Women-Attractive* and *Women-Unattractive* trial-types.

Submitting data to a 4 (*Trial-Type*) x 2 (*Sexuality*) x 2 (*Test Time*) mixed-models ANOVA revealed a main effect for Trial-Type, $F(3, 29) = 9.73$, $p < .001$, $\eta^2_{partial} = .25$, and Test Time, $F(1, 29) = 10.43$, $p = .003$, $\eta^2_{partial} = .27$, a two-way interaction between Trial-Type and Sexuality, $F(1, 29) = 4.93$, $p = .003$, $\eta^2_{partial} = .15$, Trial-Type and Test Time, $F(3, 29) = 4.45$, $p = .006$, $\eta^2_{partial} = .13$, as well as a three-way interaction between Trial-Type, Test Time, and Sexuality, $F(3, 29) = 40.29$, $p < .001$, $\eta^2_{partial} = .58$. In order to specify this three way interaction, the impact of Trial-Type and Test Time were examined separately for heterosexual and homosexual students.

With respect to heterosexual students, a main effect for Trial-Type was observed, $F(3, 14) = 5.29$, $p = .003$, $\eta^2_{partial} = .27$, along with a two-way interaction between Trial-Type and Test Time, $F(3, 14) = 16.46$, $p < .001$, $\eta^2_{partial} = .54$. During the pre-instructions IRAP, heterosexual students endorsed the belief that women were sexually attractive (*Women-Attractive*), $t(14) = 6.59$, $p < .001$, and rejected the belief that women were sexually unattractive (*Women-Unattractive*), $t(14) = 6.62$, $p < .001$. However, they showed no effects on the *Men-Attractive* or *Men-Unattractive* trial-types (both $ps > .07$). During the post-instructions IRAP, heterosexual students displayed the expected pattern of faking effects: they attenuated their effect on the *Women-Attractive* trial-type, $t(14) = 2.85$, $p = .013$, showed no effect on the *Women-Unattractive*

or *Men-Unattractive* trial types (both *ps* > .08) and now showed a strong effect on the *Men-Attractive* trial-type, $t(14) = 3.66$, $p < .003$, which was directly targeted in the faking instructions.

Submitting the data from the homosexual group to a similar set of analyses revealed a main effect for Trial-Type, $F(3, 15) = 10.14$, $p < .001$, $\eta^2_{partial} = .40$, Test Time, $F(1, 14) = 14.11$, $p = .002$, $\eta^2_{partial} = .49$, and a two-way interaction between Trial-Type and Test Time, $F(3, 14) = 29.29$, $p < .001$, $\eta^2_{partial} = .66$. During the pre-instructions IRAP homosexual students endorsed the belief that men were sexually attractive (*Men-Attractive*, $t(15) = , 9.92$ $p < .001$) and that women were sexually unattractive (*Women-Unattractive*, $t(15) = 2.22$, $p = .04$). They were also quicker to reject the belief that men were sexually unattractive (*Men-Unattractive*), $t(15) = 5.90$, $p < .001$, but showed no effect on the *Women-Attractive* trial-type ($p = .7$). During the post-instructions IRAP those same participants showed the expected pattern of faking effects: they attenuated their effect on the *Men-Attractive* trial-type, $t(15) = 3.00$, $p = .01$, showed no effect on the *Male-Unattractive* ($p = .3$) or *Women-Unattractive* trial-types ($p = .06$), and now showed a strong effect the *Women-Attractive* trial-type, $t(15) = 9.11$, $p < .001$, which was directly targeted by the faking instructions (see Table 4).[5]

Changes from Pre- to Post Instructions IRAP. If our instructions were successful, then we would expect homosexual students to strongly endorse the belief that women were attractive (*Woman-Attractive*) and show a reduced belief that men were attractive (*Men-Attractive*). We would also expect heterosexual students to strongly endorse the belief that men are attractive (*Men-Attractive*) and show attenuated belief that women are attractive (*Women-Attractive*). To test this hypothesis, change scores were submitted to a 2 (*Sexuality*) x 4 (*Trial-type*) repeated measures ANOVA. Analyses revealed a main effect for Trial-type, $F(3, 29) = 4.45$, $p = .006$,

[5] Note that in Experiments 1-4 we also averaged the *D* scores from the four IRAP trial-types to create a single overall *D* score. Submitting the overall *D* scores from the pre and post-instructions IRAPs to a similar set of statistical analyses as outlined above led to comparable findings as seen at the trial-type level.

$\eta^2_{partial}$ = .13, as well as a two-way interaction between Trial-Type and Sexuality, $F(3, 29)$ = 40.29, $p < .001$, $\eta^2_{partial}$ = .58. Consistent with our first prediction, homosexual students completely reversed the direction of their effect on the *Women-Attractive* trial-type (M = 1.21, SD = .53), unlike their heterosexual counterparts, who attenuated their effect across repeated test administrations (M = -.36, SD = .57). Consistent with our second prediction, heterosexual students reversed the direction of their *Men-Attractive* effect (M = .88, SD = .80), unlike their homosexual counterparts, whose attenuated their effect from one IRAP to the next (M = -.25, SD = .71). Unexpectedly, homosexual students also reversed their effect on the *Women-Unattractive* trial-type (M = .41, SD = .45), while heterosexual students showed an attenuated effect across the two IRAPs (M = -.29, SD = .59). Similarly, heterosexual students (M = .29, SD = .43) reversed the direction of their effect on the *Men-Unattractive* trial-type, unlike their homosexual counterparts, who showed an attenuated effect (M = -.29, SD = .36).[6]

IRAP Reliability

With respect to the pre-instruction IRAP, internal consistency for the four trial-types was as follows: *Men-Attractive*, r = .72, *Men-Unattractive*, r = .63, *Women-Attractive*, r = .63, and *Women-Unattractive*, r = .79. Internal consistency of the post-instructions IRAP was as follows: *Men-Attractive*, r = .87, *Men-Unattractive*, r = .43, *Women-Attractive*, r = .76, and *Women-Unattractive*, r = .85.

Discussion

In Experiment 4 we sought to produce patterns of faking that would be difficult to detect in experimental settings. Specifically, we instructed participants how they could attenuate their effects on one trial-type, reverse their effects on a second, and show no change on the other two.

[6] Similar to Experiment 3 we re-ran our analyses while including prior IRAP experience as a covariate. A comparable set of findings emerged even after IRAP experience was controlled for.

Our findings suggest that this outcome is possible. At baseline, homosexual students automatically endorsed the belief that men are sexually attractive and women are sexual unattractive (they also rejected the belief that men were sexually unattractive). Yet after receiving faking instructions, those same students attenuated their *Men-Attractive* effects and now responded to women as being sexually attractive. A similar set of outcomes were also obtained for their heterosexual counterparts. At baseline these students endorsed the belief that women are attractive, rejected the belief that women are unattractive, and showed no effect on either of the men related trial-types. Following faking instructions those same students attenuated their *Women-Attractive* effect and now responded to men as being sexually attractive. Therefore it seems that participants are capable of a subtle form of faking that involves exerting control over individual IRAP effects.

General Discussion

By systematically varying the nature of the instructions provided, and by generalizing our findings across a variety of domains, the current work reveals that IRAP effects, like those obtained from the IAT and priming, are not impervious to strategic manipulation. Consistent with McKenna et al. (2007) we found that merely informing participants to ‘fake’ their performance without providing a concrete strategy to do so did not eliminate, reverse, or in any way alter the obtained outcomes, even when participants had immediate prior experience with the task (Experiment 1). However, when instructions orientated attention towards the core parameters of the procedure, participants spontaneously derived a strategy that allowed them to eliminate, but not reverse, their IRAP effects (Experiment 2). Indeed, participants were only able to reverse the direction of their effects when they were provided with a specific response strategy that was continually reiterated throughout the task (Experiment 3). By refining the nature of the

instructions provided, we managed to produce subtle patterns of faking that would be difficult to identify in natural settings (Experiment 4). Taken together, it seems that IRAP performance can be strategically manipulated. However, the degree to which control can be exerted over one's performance is contingent upon (a) the nuanced nature of that control (whether one is attempting to fake overall or trial-type effects) and (b) whether one is provided with a viable response strategy or has to devise such a strategy for themselves.

Several issues are worth noting here. First, we recognize that the instructed faking strategies used in Experiments 3-4 are unlikely to be spontaneously devised by uninformed participants in natural settings. Indeed, in ecologically valid situations a sophisticated faking pattern would require participants to manipulate individual trial-types on the basis of their own self-generated response strategy. Drawing on the findings of Barnes-Holmes et al. (2010), for example, if participants wanted to conceal evidence of automatic racial bias, they would need to reverse their effect on a single trial-type (*Black-Negative*) while leaving the other three untouched. Yet based on Experiment 4, it seems that sophisticated levels of manipulation, while certainly possible, are among the most difficult to produce. In other words, the current paper was not concerned with the degree to which participants habitually exert control over their IRAP performance – simply whether such control is possible.

Second, researchers continue to use indirect procedures like the IAT and evaluative priming despite their susceptibility to manipulation because they (a) capture automatic behaviors that often allude self-report procedures and (b) predict thoughts, feelings and actions across a wide variety of settings (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). Demonstrating that the IRAP is also susceptible to control does not threaten its utility insofar as it also captures and predicts those same classes of behavior (see Hughes & Barnes-Holmes, 2013; Vahey et al.,

2015). Third, it appears that unlike the IAT and the AMP, participants were unable to spontaneously generate a strategy to alter their performance without direct intervention by the researcher, despite the fact that they had just completed an IRAP several minutes before (Experiment 1). Even when attention was drawn to aspects of their behavior that they should change in order to manipulate their effects (Experiment 2) the ensuing response strategy attenuated, rather than reversed the direction of those outcomes. Only when participants were equipped with detailed instructions and those instructions were constantly reiterated throughout the task was evidence of robust faking observed. Thus controlling the direction and magnitude of one's IRAP effects seems to be relatively more difficult to achieve when compared to alternative indirect procedures. Critically, however, support for this claim necessitates future research in which the relative sensitivity to manipulation of IRAP, IAT, and AMP effects, and other newly introduced indirect procedures (e.g., DeHouwer et al., 2015; O'Shea, et al., 2015) are directly compared. Surprisingly, past work has tended to focus on the degree to which a single implicit measure can be faked, and in some instances, compared performances on indirect to direct procedures. Yet researchers have never compared the degree to which one implicit measure is more or less sensitive to faking than another, or compared their respective dependency on factors such as prior task experience or instructed response strategies. Systematic investigation of these factors could allow for indirect procedures to be arranged along a continuum from higher to lower controllability and provide researchers with a means to select between tasks when controllability is an issue (e.g., in applied or diagnostic settings).

Open Questions and Future Directions

Although our research sheds new light on the IRAP's sensitivity to strategic manipulation several important questions still need to be addressed. We know very little about the contextual

conditions that either increase or decrease a person's ability to strategically influence their performance. It may be that, just like the IAT, participants are better able to exert control over their behavior when they have previously encountered the task, are given repeated opportunities to manipulate their performance, or when they are exposed to multiple IRAPs in close temporal succession. For instance, in Experiments 1-4 participants always completed a baseline IRAP before receiving instructions to fake their performance on a second IRAP. It may be that the observed faking performances were – in part – contingent upon this prior experience with the task. It is also possible that fixing versus randomizing the location of the IRAP's response options (or even the content of those response options) may influence the success of any faking attempt. For instance, when fixed response options are employed participants no longer need to 'track' the changing location of a response, thus providing them with additional time to exert control over their task performance. Yet keeping response options in constant flux across trials may make it more difficult to meet the mastery criteria, and adhere to the response rule operating in that block of trials, while also attempting to manipulate one's performance. Put simply, faking may be facilitated in the former and undermined in the latter situation. We should note that this property of the IRAP was varied in an unsystematic manner across studies insofar as fixed response locations were used in Experiments 1 and 4 whereas random locations were used in Experiments 2 and 3. Although we observed evidence of strategic manipulation across three of these four studies future work could systematically vary this property to determine its potential impact on faking success. The same goes for the manner in which target and label stimuli have to be related during each IRAP trial. It may be that increasing the complexity of the presented stimuli, or the ways in which they have to be related, decreases the likelihood of successful faking performances. For instance, participants in Experiment 1 had to relate stimuli as either

'Similar' or 'Opposite', whereas their counterparts in Experiments 2-4 had to relate stimuli using 'True' and 'False'. The impact of using different response options in this way currently remains unknown. Similarly, while we know that faking one IAT increases the probability of faking another (Röhner et al., 2011), the extent to which this is also true for the IRAP remains to be seen. When answering the above questions it will be important to control for other possible moderators, such as domain of interest (socially versus non-socially sensitive), its relevance to the individual, along with their current physiological and psychological state. Thus new studies are needed which systematically manipulate prior experience with the IRAP (both within and between experimental sessions) while accounting for the above factors.

Another question concerns the degree to which people devise and utilize response strategies in natural settings. In other words, how likely is faking in everyday research using the IRAP? The ability to implement externally-conveyed faking instructions, while certainly informative, does not tell us whether people are capable of discovering and applying such strategies for themselves. Nor does it tell us what strategies they tend to devise. It may be that some participants alter their speed on 'Rule A' blocks, 'Rule B' blocks or both. In addition, by implementing the above strategies at the trial-type (rather than block level) participants could alter their outcomes in a manner that is difficult to detect and correct for. Therefore future research should identify those aspects of the procedure, participant, and context which increase the likelihood of self-initiated attempts to manipulate IRAP effects. For instance, motivation to fake could be manipulated by exposing participants to a mock interview with an internationally respected company and then informing them that their job would involve daily contact with a specific racial group (see Teige-Mocigemba et al., 2015 for such work with the AMP). Participants could also be informed that the mock hiring decision would be based on how they

answer a series of questionnaires as well as the IRAP. The hiring agent (a confederate) could "help" the participant by telling them that their chances of employment would increase if they "acted as if they liked the racial group on the IRAP". Researchers could compare the extent to which this information leads to modified IRAP effects relative to a participant who is simply exposed to the IRAP on two separate occasions.

Finally, past work has shown that it is possible to statistically detect and correct for successful manipulation attempts on the IAT. Yet no such efforts have been taken with regard to the IRAP. Future work could not only identify the environmental conditions that undermine the ability to control task performance but also assess whether fakers have a 'signature' response style which could be used to identify and control for their biased performances. Again, the relative ease of detecting faking could also be compared across different indirect procedures. When it comes to the IRAP two factors may serve as important diagnostic markers for those attempting to manipulate their effects. The first is the ability to maintain accuracy and latency criteria across successive blocks of trials. It may be that some participants can implement a faking strategy but that the parallel requirement to maintain a strict speed and accuracy criteria causes them to occasionally dip below these criteria from time to time. A second marker for faking attempts could be attrition rates. Experiments 3 and 4 provided the clearest evidence of faking and yet the highest number of participants who failed to pass the task. It may simply be too difficult for some participants to follow one response rule (i.e., "fake performance"), implement another (i.e., either "Rule A" or "Rule B") while simultaneously meeting the IRAPs mastery criteria. This would explain why attrition levels increased in-line with the complexity of the faking instructions. Researchers could replicate the aforementioned studies while including a between-subjects control condition in order to determine if the observed attrition rates are a

function of repeated task completion or attempts to strategically edit one's effects at increasing levels of complexity.

References

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2011). Detecting fakers of the autobiographical IAT. *Applied Cognitive Psychology, 25*, 299-306.

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.

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., & Snowden, R. J. (2010). Faking of the Implicit Association Test is statistically detectable and partly correctable. *Basic and Applied Social Psychology, 32,* 302–314.

Dawson, D. L., Barnes-Holmes, D., Gresswell, D. M., Hart, A. J. P., & Gore, N. J. (2009). Assessing the implicit beliefs of sexual offenders using the Implicit Relational Assessment Procedure: A first study. *Sexual Abuse: A Journal of Research and Treatment, 21,* 57–75.

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.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology, 27,* 307–316.

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.

Golijani-Moghaddam, N., Hart, A., & Dawson, D. (2013). The implicit relational assessment procedure: emerging reliability and validity data. *Journal of Contextual Behavioral Science, 2*, 105-119.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. (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., & 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.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin 30,* 161–172.

Hughes, S., & Barnes-Holmes, D. (2013). A Functional Approach to the study of implicit cognition: The Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. In Dymond, S & Roche, B. (Eds.). Advances in Relational Frame Theory & Contextual Behavioral Science: Research & Application. Oakland, CA: New Harbinger Publications.

Kim, D. (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly*, *66,* 83–96.

Klein, F. (1993). *The bisexual option*. New York: Haworth Press.

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual.* Technical Report A-8. University of Florida, Gainesville, FL.

Langner, O., Ouwens, M., Muskens, M., Trumpf, J., Becker, E. S., & Rinck, M. (2010). Faking on direct, indirect, and behavioural measures of spider fear: Can you get away with it? *Cognition and Emotion, 24,* 549-558.

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.

Minear, M. & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers, 36,* 630-633.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., et al. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods, 45,* 169–177.

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 OCD. *Journal of Behavior Therapy and Experimental Psychiatry, 43,* 922-930.

O'Shea, B., Watson, D. G. & Brown, G. D. A. (2015). Measuring implicit attitudes: A positive framing bias flaw in the Implicit Relational Assessment Procedure (IRAP). *Psychological Assessment.* Available on-line at http://dx.doi.org/10.1037/pas0000172

Payne, B. K., Cheng, S. M., Goverun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89,* 277–293.

Remue, J., De Houwer, J., Barnes-Holmes, D., Vanderhasselt, M.-A., & De Raedt, R. (2013). Self-esteem revisited: Performance on the implicit relational assessment procedure as a measure of self- versus ideal self-related cognitions in dysphoria. *Cognition & Emotion, 27,* 1441-9.

Roddy, S., Stewart, I., & Barnes-Holmes, D. (2011). Facial reactions reveal that slim is good but fat is not bad: Implicit and explicit measures of body size bias. *European Journal of Social Psychology, 41,* 488-494.

Röhner, J., Schröder-Abè, M., & Schütz, A. (2011). Exaggeration is harder than understatement, but practice makes perfect! Faking success in the IAT. *Experimental Psychology, 58,* 464-472.

Röhner, J., Schröder-Abé, M., & Schütz, A. (2013). What do fakers actually do to fake the IAT? An investigation of faking strategies under different faking conditions. *Journal of Research in Personality, 47,* 330-338.

Steffens, M. C. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology, 51,* 165–179.

Stieger, S., Göritz, A. S., Hergovich, A., & Voracek, M., (2011). Intentional faking of the Single Category Implicit Association Test and the Implicit Association Test. *Psychological Reports, 109,* 219-230.

Teige-Mocigemba, S., Penzl, B., Becker, M., Henn, L., Klauer, K. (2015). Controlling the "Uncontrollable": Faking Effects on the Affect Misattribution Procedure. *Unpublished manuscript*.

Teige-Mocigemba, S. & Klauer, K. C. (2013). On the controllability of evaluative-priming effects: Some limits that are none. *Cognition & Emotion, 27,* 632-657.

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally orientated self-control. *Perspectives of Psychological Science, 5,* 243 –262.

Verschuere, B., Prati, V., & De Houwer, J. (2009). Cheating the lie-detector: Faking the autobiographical IAT. *Psychological Science, 20,* 410–413.

Weinrich, J. D., Snyder, P. J., Pillard, R. C., Grant, I., Jacobson, D. L., Robinson, S. R., et al. (1993). A factor analysis of the Klein Sexual Orientation Grid in two disparate samples. *Archives of Sexual Behavior, 22,* 157–168.

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology, 55,* 493–518.

Wittenbrink, B. (2007). Measuring attitudes through priming. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 17–58). New York: Guilford Press.

Appendix A

Table 1

Mean and standard deviation scores for the four IRAP trial-type effects as a function of instruction type (faking vs. control) and test time (pre- versus post-instructions).

	Faking Instruction		No Faking Instructions		Total	
	M	SD	M	SD	M	SD
Pre-Instructions IRAP						
Positive-Positive Trial Type	0.64*	0.37	0.60*	0.42	0.62*	0.39
Positive-Negative Trial Type	0.34*	0.42	0.19*	0.39	0.26*	0.41
Negative-Positive Trial Type	-0.26*	0.39	-0.14	0.43	-0.19*	0.42
Negative-Negative Trial Type	-0.21*	0.47	0.03	0.46	-0.08	0.48
Post-Instructions IRAP						
Positive-Positive Trial Type	0.51*	0.47	0.36*	0.35	0.43*	0.41
Positive-Negative Trial Type	0.25*	0.40	0.09	0.45	0.17*	0.43
Negative-Positive Trial Type	0.10	0.49	-0.01	0.41	0.04	0.45
Negative-Negative Trial Type	-0.04	0.37	0.09	0.39	0.03	0.38

Note. * indicates that the corresponding IRAP effect differed significantly from zero ($p < .05$).

Table 2

Mean and standard deviation scores for the four IRAP trial-type effects as a function of instruction type (faking vs. control) and test time (pre- versus post-instructions).

	Faking Instruction		No Faking Instructions		Total	
	M	SD	M	SD	M	SD
Pre-Instructions IRAP						
Pleasant-Positive Trial Type	0.41*	0.42	0.27*	0.42	0.34*	0.42
Pleasant-Negative Trial Type	0.39*	0.45	0.31*	0.38	0.35*	0.41
Disgusting-Positive Trial Type	-0.15	0.48	-0.18*	0.38	-0.17*	0.43
Disgusting-Negative Trial Type	-0.64*	0.33	-0.47*	0.41	-0.55*	0.38
Post-Instructions IRAP						
Pleasant-Positive Trial Type	-0.10	0.68	0.31*	0.45	0.12	0.59
Pleasant-Negative Trial Type	-0.07	0.62	0.25*	0.50	0.10	0.58
Disgusting-Positive Trial Type	0.31*	0.59	-0.03	0.30	0.13	0.49
Disgusting-Negative Trial Type	-0.10	0.71	-0.56*	0.37	-0.34*	0.59

Note. * indicates that the corresponding IRAP effect differed significantly from zero ($p < .05$).

Table 3

Mean and standard deviation scores for the four IRAP trial-type effects as a function of instruction type (simple vs. detailed faking) and test time (pre- versus post-instructions).

	Simple Faking		Detailed Faking		Total	
	M	SD	M	SD	M	SD
Pre-Instructions IRAP						
White-Positive Trial Type	0.32*	0.35	0.35*	0.45	0.34*	0.40
White-Negative Trial Type	0.13	0.29	-0.00	0.33	0.06	0.32
Black-Positive Trial Type	0.19*	0.29	0.08	0.33	0.14*	0.32
Black-Negative Trial Type	-0.01	0.37	-0.03	0.35	-0.02	0.35
Post-Instructions IRAP						
White-Positive Trial Type	0.10	0.51	-0.98*	0.58	-0.48*	0.77
White-Negative Trial Type	-0.06	0.55	-0.97*	0.67	-0.55*	0.77
Black-Positive Trial Type	0.29*	0.37	1.14*	0.57	0.75*	0.64
Black-Negative Trial Type	0.06	0.48	0.98*	0.54	0.55*	0.69

Note. * indicates that the corresponding IRAP effect differed significantly from zero ($p < .05$).

Table 4

Mean and standard deviation scores for the four IRAP trial-type effects as a function of sexuality (heterosexual vs. homosexual) and test time (pre- versus post-instructions).

	Heterosexual		Homosexual		Total	
	M	SD	M	SD	M	SD
Pre-Instructions IRAP						
Women-Attractive Trial Type	0.73*	0.43	0.03	0.38	0.37*	0.54
Women-Unattractive Trial Type	0.55*	0.32	-0.15*	0.28	0.19*	0.46
Men-Attractive Trial Type	-0.16	0.54	0.69*	0.28	0.28*	0.61
Men-Unattractive Trial Type	-0.16	0.32	0.39*	0.27	0.13	0.41
Post-Instructions IRAP						
Women-Attractive Trial Type	0.37*	0.50	1.23*	0.54	0.81*	0.67
Women-Unattractive Trial Type	0.25	0.51	0.25	0.49	0.25*	0.49
Men-Attractive Trial Type	0.71*	0.75	0.45*	0.59	0.58*	0.68
Men-Unattractive Trial Type	0.13	0.37	0.10	0.39	0.12	0.38

Note. * indicates that the corresponding IRAP effect differed significantly from zero ($p < .05$).

Appendix B

Detailed Faking Instructions (Experiment 2)

To fake your responses on this task you need to do two things First, try to respond slowly on those blocks that require you to respond as if "Black is Bad" and "White is Good". For instance, when blocks ask you to respond to images of White people as "Good" and Black people as "Bad" please respond slowly. Second, try to respond quickly on those blocks that require you to respond as if "Black is Good" and "White is Bad". For instance, when blocks ask you to respond to images of White people as "Bad" and images of Black people as "Good" please respond quickly. ** Note: You still have to respond within the speed and accuracy criteria that you encountered in the previous version of this task (i.e., try to avoid the red X and the !** You will be reminded before each block when to go quickly and when to go slowly.

Detailed Faking Instructions – Heterosexual Condition (Experiment 3)

To fake your responses on this task you need to do two things First, on the "women are attractive and men are unattractive" blocks, try to respond slowly on men attractive and women attractive trials but quickly on the other trials. Second, on the "men are attractive and women are unattractive" blocks, try to respond slowly on women attractive trials but quickly on the other trials.' ** Note: You still have to respond within the speed and accuracy criteria that you encountered in the previous version of this task (i.e., try to avoid the red X and the !** You will be reminded before each block when to go quickly and when to go slowly.

Appendix C

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 10,000ms 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 300ms; (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 choses.