Psychol Rec (2016) 66:153–163
DOI 10.1007/s40732-015-0160-1

ORIGINAL ARTICLE

Testing the IRAP: Exploring the Reliability and Fakability of an Idiographic Approach to Interpersonal Attitudes

Chad E. Drake[1] · **Kail H. Seymour**[1] · **Reza Habib**[1]

Published online: 30 December 2015

Abstract Although multiple studies have demonstrated that the Implicit Association Test (IAT) can be successfully faked, a single study using the Implicit Relational Assessment Procedure (IRAP; McKenna, Barnes-Holmes, Barnes-Holmes, & Stewart, International Journal of Psychology and Psychological Therapy, 7(2): 253–268, 2007) concluded that this behavior-analytic alternative to implicit measures may be relatively immune to faking. The current study examined the fakability of the IRAP using more accessible faking instructions and an idiographic approach to stimulus selection. The methodology also provided an opportunity to examine split-half and test–retest reliability, an underreported statistic among existing IRAP publications. Three IRAPs were delivered in succession to 102 undergraduates randomly assigned to one of three conditions: three real IRAPs, two real IRAPs followed by a faked IRAP, or one real IRAP followed by two faked IRAPs. Split-half reliabilities were acceptable for four of the six real IRAPs and excellent for the three faked IRAPs. For the group receiving three real IRAPs, consecutive IRAPs generated significant test–retest correlations; the group engaging in two consecutive faked IRAPs also generated a significant test–retest correlation. Furthermore, all groups that received faking instructions subsequently demonstrated significantly reversed IRAP performance compared to previous real IRAPs and to real IRAPs in other conditions: Faking was robustly demonstrated. These results may have implications with regard not only to the fakability but also the reliability of the measure. Future research might focus on methods of detecting and/or preventing faking of the procedure.

Keywords Implicit Relational Assessment Procedure · Faking · Reliability · Idiographic · Relationships

In the decades that have passed since Allport (1935) wrote about the importance of attitude, empirical psychology has produced extensive research on this hypothetical construct. The most common means of measuring attitude involves responding to statements by selecting options from a Likert-type array. In other words, the respondent provides a public, verbal estimation of their private behavior. This approach is known by various monikers, including self-report, explicit, and direct measures (De Houwer, 2006; Greenwald & Banaji, 1995). Although the self-report approach strategy has proven popular and scientifically productive, the nature of self-reports presupposes that respondents both can and will report their thoughts and beliefs honestly. However, multiple studies have revealed evidence of presentational biases as well as inconsistencies between self-reports and other means of assessment (Nisbett & Wilson, 1977; Paulhus, 1989; see also Greenwald & Banaji; Hughes, Barnes-Holmes, & De Houwer, 2011). In such cases, self-reports are unlikely to be sufficient.

Some measurement methodologies have shown potential for reducing problems with the self-report paradigm; most of these attempt to measure a behavior of interest rather than self-reports about said behavior of interest. As such, these measures are often referred to as behavioral, implicit, or indirect measures (De Houwer, 2006; Greenwald & Banaji, 1995). In recent decades, a number of these behavioral measures have gained prominence in attitudinal research, as it has become possible to use computerized tasks to assess response time

✉ Chad E. Drake
chad.e.drake@gmail.com

[1] Department of Psychology, Southern Illinois University, 1125 Lincoln Drive, Mail Code 6502, Carbondale, IL 62901-6502, USA

latencies with respect to complex arrays of stimuli (see Hughes et al., 2011; Nosek, Hawkins, & Frazier, 2011). Arguably, the measure that has been most responsible for this shift is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). The IAT has been used to show attitudinal biases that were less apparent when assessed with self-reports, especially for socially sensitive attitudes such as racism (e.g., Green et al., 2007) and homophobia (e.g., Steffens, 2005).

Along with studies showing a lack of convergence between IAT scores and self-reported attitudes, researchers have directly addressed the manipulability of IAT performance by evaluating the ability of respondents to purposefully "fake" biases on the task. Faking a bias may involve intentionally manipulating one's responding in order to misrepresent the response patterns that would occur in the absence of such intentions. In most published studies of faking with the IAT, results have shown that the task is readily faked when respondents are provided with specific instructions for doing so, and that non-specific instructions are less effective in engendering faked response patterns (Agosta, Ghirardi, Zogmaister, Castiello, & Sartori, 2011; Kim, 2003; Steffens, 2004; Stieger, Goritz, Hergovich, & Voracek, 2011; Tulbure, 2006). Nevertheless, some studies have shown that even non-specific instructions may lead to successful faking (De Houwer, Beckers, & Moors, 2007; Fiedler & Blumke, 2005; McDaniel, Beier, Perkins, Goggin, & Frankel, 2009; Röhner, Schröder-Abé, and Schütz, 2011, 2013).

The demonstrated fakability of the IAT has led some researchers to examine various means of distinguishing faked from non-faked performance. The results of these efforts have been mixed, with one study demonstrating up to 87 % accuracy in identifying fakers (Agosta et al., 2011) and another reporting no better than 58 % accuracy (Fiedler & Bluemke, 2005). These studies also controlled for practice effects, attempting to induce faking in certain conditions before respondents had any familiarity with the procedure, and also showed mixed results. Among participants who were naïve to the procedure but explicitly informed of how to fake it, Agosta and colleagues reported some success at faking an IAT loaded with accurate and inaccurate autobiographical information, whereas Fiedler and Bluemke reported no success at producing faking with an IAT assessing social identity. At least one IAT study conducted in a naturalistic setting has revealed the occurrence of distorted responding in the absence of any faking instructions. Vecchione, Dentale, Alessandri, and Barbaranelli (2014) showed that employees who were told that their performance would not be anonymous produced more socially desirable IAT scores reflecting the Big Five traits compared to a group of participants whose anonymity was assured.

Attitude researchers today have a variety of implicit measures to consider in addition to the IAT (Nosek et al., 2011). However, most of these are grounded in associative theories that describe performance on these measures in mechanistic and hypothetical terms (e.g., Greenwald et al., 2002). As such, an interpretation of task performance in descriptive, inductive behavioral terms can be challenging. An alternative implicit measure grounded in behavior-analytic theory is the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes et al., 2006). Developed within the context of a program of research on Relational Frame Theory (Hayes, Barnes-Holmes, & Roche, 2001), the IRAP has been described as a measure of brief and immediate relational responding (BIRR) and extended and elaborated relational responding (EERR), repertoires that are more readily accounted for by conditioning histories and contextual variables (Hughes & Barnes-Holmes, 2013). As with the IAT, the IRAP has demonstrated utility as a measure of cognitive repertoires for a variety of content domains, including social attitudes (e.g., Cullen, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009; Roddy, Stewart, & Barnes-Holmes, 2010) and clinical concerns (e.g., Carpenter, Martinez, Vadhan, Barnes-Holmes, & Nunes, 2012; Hussey & Barnes-Holmes, 2012).

To date, only one published study has directly examined the fakability of the IRAP (McKenna, Barnes-Holmes, Barnes-Holmes, & Stewart, 2007). As with many IAT studies on faking, a group design contrasted a no-instructions control condition with two faking conditions, one with non-specific and the other with specific faking instructions. Two IRAPs were administered in succession, each containing the words "pleasant" and "unpleasant" in conjunction with two groups of generically positive and negative words (e.g., "freedom", "love", "murder", "sickness"). In the applicable conditions, faking instructions were provided between the first and second IRAP. Analyses revealed a small reduction in expected bias from the first to the second IRAP, but not as a function of condition. The authors concluded that there was little evidence of any effect for faking. Furthermore, detailed inspection of individual performance did not correspond well to self-reported attitudes regarding success at faking the task.

Although the findings reported by McKenna and colleagues (2007) suggest that the IRAP may be more immune to demand effects than the IAT, the lack of evidence for faking might be explained in a variety of ways. Several of the participants reported that the faking instructions were too difficult to follow, and indeed, the instructions appear to have been more complex than those provided in most IAT studies. Perhaps the participants were not provided an adequate strategy for faking, or were not suitably invested in attempting to fake repertoires with respect to the stimuli contained in the IRAP. Rather than evidence in support of the stability of the measure, the results may instead reflect null findings resulting from insufficient control over variables that would facilitate faking. Additional research on the fakability of the IRAP may further illuminate factors that could enhance

or degrade the utility of the measure. In addition, studies on faking typically involve multiple administrations of the task, providing an opportunity to explore test–retest reliability, a fundamental psychometric variable for any measure. Unfortunately, reliability data are rarely reported in fakability studies, and are generally under-addressed in the implicit cognition literature (for a recent review, see Golijani-Moghaddam, Hart, & Dawson, 2013).

The current study was designed to explore an idiographic approach to stimulus selection, a more accessible set of faking instructions, and the reliability psychometrics of the IRAP. Across three different conditions, three IRAPs were successively administered in which attitudes regarding two individuals known to each respondent were idiographically assessed. This idiographic strategy was implemented under the assumption that the use of stimuli that were more personalized and evocative than was typically the case might enhance the reliability of the measure and possibly render it more resistant to faking. The three conditions differed with respect to when faking instructions were provided (including one condition in which faking instructions were never provided), which allowed for assessment of differential acquisition of faking performance among the conditions over time. In addition to planned analyses for reliability, a number of self-report measures were also administered in order to allow an evaluation of the function of the stimuli entered into the IRAPs.

Method

Participants

One hundred two participants were recruited from an introductory psychology class at a Midwestern university. All participants received course credit for participating. The average age of the sample was 19.0 years ($SD = 1.70$), with 67 participants (65.7 %) identifying as female. Fifty-six participants (54.9 %) identified as white or Caucasian, 38 (37.3 %) identified as black or African-American, and 16 (15.7 %) reported various other racial affiliations; seven participants (6.9 %) selected more than one racial category (which accounts for the summed percentages exceeding 100), and two participants did not endorse any category.

Measures

Demographic Information Participants completed a brief questionnaire that contained items for a variety of demographic variables, including age, sex, and race.

Names Questionnaire This form asked for the names of two individuals personally known to the participant. One was of "someone who has affected you in a very positive way" and the other was "someone who has affected you in a very negative way". Although participants provided names for a variety of relationships (see Table 1), they will be categorized simply as "friend" or "enemy" throughout the remainder of this report.

Evaluative Stimulus Rating Scale (ESRS) The 12-item ESRS assessed the participant's perception of the 12 evaluative words used as target stimuli in the IRAP procedure. Each evaluative word was presented with a Likert-type scale ranging from −5 (*extremely negative*) to +5 (*extremely positive*). An average score was computed for positive (ESRS Positive) and negative words (ESRS Negative). Cronbach's alpha values for these two scores were 0.807 for positive and 0.903 for negative words.

Positive Relationship Questionnaire (PRQ) Each participant answered three questions regarding the positively regarded person whose name had been provided on the Names Questionnaire. The first inquired about the nature of the relationship (e.g., mother, friend), and the remaining two questions were designed to assess the participant's perception of the individual. One item stated, "This person is a good person.", and the other stated, "This person is a bad person." Response options ranged from −3 (*strongly disagree*) to +3 (*strongly agree*) for both questions. These items were significantly correlated (Spearman's $r = -0.595$, $p < 0.001$). A bias-corrected bootstrapping method using 1000 samples resulted in a 95 % confidence interval of this correlation that ranged from −0.41 to −0.75. Answers to the second question were multiplied by −1 and added to the answer to the first question, providing a composite score for the participant's perception of this person.

Negative Relationship Questionnaire (NRQ) The NRQ was identical to the PRQ, but the questions pertained to the negatively regarded person. The two items were significantly correlated (Spearman's $r = -0.812$, $p < 0.001$). A bias-corrected bootstrapping method using 1000 samples resulted in a 95 % confidence interval of this correlation ranging of

Table 1 Relationship Categories Reported by Participants

	Full Sample ($N = 102$)		Analyzed Sample ($n = 73$)	
	Positive	Negative	Positive	Negative
Relative	48	19	32	16
Romantic Partner	13	12	12	8
Friend	23	39	17	27
Named Individual	14	17	10	12
Other	2	14	0	9
No Answer	2	1	2	1

between −0.70 and −0.89. A composite score of the two perception items was calculated in the same manner as with the PRQ.

Faking Quiz and Questionnaire (FQQ) Participants assigned to either of the faking conditions completed different portions of the FQQ via paper and pencil before and after attempting to fake the IRAP (see details in "Procedure"). The front of the FQQ contained two multiple-choice questions designed to assess understanding of the faking instructions. The first question pertained to a faking strategy for the pro-friend/anti-enemy block-type, and the second to the pro-enemy/anti-friend block-type (see details in the "Implicit Relational Assessment Procedure" section below). The back of the FQQ contained Likert-type questions about the participant's performance in faking the IRAP, including items for Following Instructions and Success at Faking. The Following Instructions item stated, "Overall, how closely did you follow the faking instructions for the task?" and was paired with a scale ranging from 0 (*not at all*) to 10 (*perfectly*). The Success at Faking item stated "How successful do you think you were at faking your performance on the task?" and was paired with a scale ranging from 0 (*not at all successful*) to 10 (*extremely successful*).

Implicit Relational Assessment Procedure (IRAP) The IRAP is a response time measure of derived relational repertoires (Barnes-Holmes et al., 2006; Hughes & Barnes-Holmes, 2013). The current study populated the procedure with two samples and 12 target stimuli. The samples were the names provided by the participants on the Names Questionnaire. The targets were two groups of evaluative words, one generally viewed as positive ("caring", "friend", "good", "nice", "safe", and "trustworthy") and the other generally viewed as negative ("bad", "cruel", "dangerous", "enemy", "hateful", and "selfish"). During each block of the procedure, each name was paired once with each evaluative word, generating 24 unique trials per block. These 24 trials were presented in random order in each block of the procedure, with the words "true" and "false" offered as response options for each trial. The 24 trials can be categorized into four groups, known as trial-types: friend-positive, friend-negative, enemy-positive, and enemy-negative (see Fig. 1).

The blocks of the procedure containing these trials can be categorized into two groups, known as block-types: pro-friend/anti-enemy, and pro-enemy/anti-friend. The block-type determines the pattern of responses required for each trial and, ultimately, completion of each block. For the pro-friend/anti-enemy block-type, participants were required to provide answers likely to be consonant with their learning history. Thus, the friend-positive and enemy-negative trial-types required the selection of "true", while the friend-negative and enemy-positive trial-types required the selection of "false". The selection of these "correct" answers resulted in a blank screen for 400 ms, followed by presentation of the next trial (and eventually completion of the block); failure to select these responses resulted in a red "X" appearing in the middle of the computer screen until a "correct" answer was selected. For the pro-enemy/anti-friend block-type, the opposite response pattern was required (e.g., selecting "true" for the enemy-positive trial-type). Throughout the procedure, the block-type alternated in succession, with each new block requiring an opposing pattern of responses compared with that of the previous block. After a series of practice blocks, participants engaged in six test blocks.

Three IRAPs were administered in succession. The pro-friend/anti-enemy block-type was administered first for all IRAPs in all conditions. For the first IRAP, up to six practice blocks were administered with the assistance of the experimenter. Prior to engaging the first practice block of the first IRAP, the experimenter provided instructions regarding the content and performance requirements of the task. Participants were required to exhibit greater than 78 % accuracy, with a median latency below 2000 ms, on two successive blocks (a pro-friend block and a subsequent pro-enemy block) before proceeding to the test blocks. The experimenter observed participants' performance from a distance once they began the test blocks. If a participant failed all six practice blocks, the procedure was closed and restarted, and test blocks were administered without any practice blocks. The second and third IRAPs also were administered without practice blocks (a detailed protocol is available from the first author).

Procedure

All experimental materials were administered to individual participants by a single experimenter in a small room. One hour was allotted for each participant to finish the entire study. After reading and signing an informed consent statement, participants filled out the Names Questionnaire via paper and pencil, and subsequently received either the remaining self-reports (demographics form, ESRS, PRQ, and NRQ) presented in random order via an online survey website or the three IRAPs. The order of self-reports and IRAPs was counterbalanced across participants. Participants were randomly assigned to one of three conditions: (a) the "real-real-real" condition, which involved the administration of the IRAP three times in succession; (b) the "real-real-fake" condition, which was identical to the real-real-real condition, except that the participant attempted to fake the third IRAP; or (c) the "real-fake-fake" condition, which involved attempts to fake both the second and third IRAPs.

Participants assigned to either of the faking conditions were provided with a rationale for the IRAP assessment strategy before the second or third IRAP (depending on condition), and were subsequently asked if they would be willing to

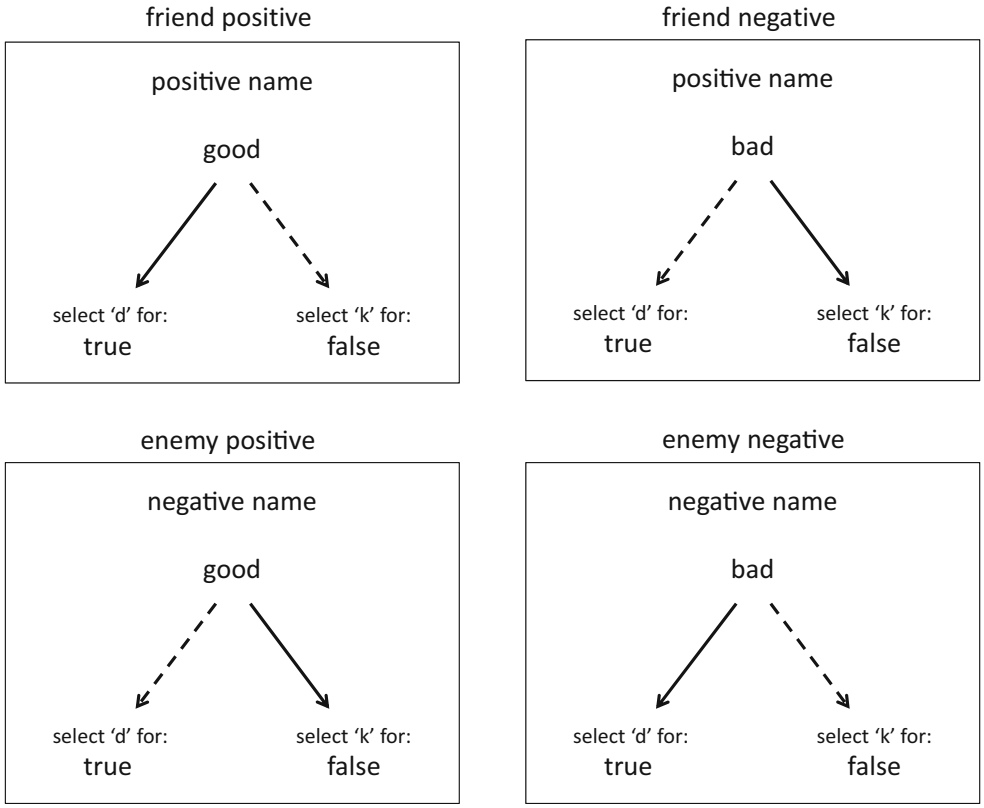

Fig. 1 IRAP trial-types, with *solid arrows* denoting correct answers for pro-friend/anti-enemy blocks and *dashed arrows* denoting correct answers for pro-enemy/anti-friend blocks (arrows did not appear during actual trials)

attempt to fake the measure on the next administration. All participants agreed to attempt faking. Before engaging the first test block (a pro-friend/anti-enemy block), participants were given a faking strategy (i.e., perform within the time limit but otherwise do not attempt to finish quickly) and were provided with the first question of the FQQ via paper and pencil. After clarification of any confusion, if needed, the participant engaged the first test block. Once this was completed, the participant was given a faking strategy for the second block (i.e., rehearse the necessary answers first, then perform as well as possible) and was provided with the second question of the FQQ. After clarification of any confusion, the participant engaged the second block. Subsequently, participants were informed that they would be allowed to proceed through the remainder of the IRAP on their own, and to continue attempting to fake the task by following the faking instructions. After finishing the faked IRAP, participants completed the remaining questions on the FQQ regarding their adherence to the faking instructions and their perceived success in faking the IRAP. Participants assigned to the real-fake-fake condition also were asked to continue attempting to fake the task before engaging in the third IRAP, but otherwise received no additional instruction or guidance. After they had completed all IRAPs and all self-report measures, participants were debriefed about the study and were credited for their participation.

Data Processing

As has been detailed in other studies, IRAP data were processed via the D-score algorithm (Barnes-Holmes, Murtagh, Barnes-Holmes, & Stewart, 2010), which uses a series of calculations involving average latency scores for each trial-type of each block of the procedure, to produce an overall D-score for each IRAP. Positive D-scores would represent an expected pro-friend/anti-enemy bias, whereas negative D-scores would represent an unexpected pro-enemy/anti-friend bias. Average latencies for each block-type were also retained for analysis. In order to provide for analyses of split-half reliability, the D-score algorithm was also applied separately to odd-numbered and even-numbered trials within each trial-type, producing two overall D-scores, one for odd and one for even trials.

Results

Attrition

Among the 102 participants, ten (9.8 %) failed to complete all three IRAPs within the allotted hour. One of these participants was assigned to the real-real-real condition, two were assigned to the real-real-fake condition, and seven to the real-fake-fake condition. A chi-square analysis showed that completion

failure differed by condition, χ^2 (2, $N = 102$) = 6.939, $p = 0.031$. The additional time consumed by the provision of the faking instructions and the FQQ, as well as engagement in the faking strategy (e.g., slowing down on consistent blocks), seems to have disproportionately impacted the real-fake-fake condition. An additional 19 participants (18.6 %) failed to produce accuracy scores of 70 % or higher on all test blocks in all three IRAPs. Six participants failed to maintain accuracy in the real-real-real condition, six failed in the real-real-fake condition, and seven in the real-fake-fake condition. A chi-square analysis showed that failure to maintain accuracy was not a function of condition, χ^2 (2, $N = 92$) = 0.729, $p = 0.695$. Faking does not appear to have differentially impacted performance accuracy in the IRAP. All data from these 10 and 19 participants, respectively, were excluded from subsequent analyses, leaving 29 participants in the real-real-real condition, 24 participants in the real-real-fake condition, and 20 participants in the real-fake-fake condition.

Descriptive Statistics and Condition Comparisons

Descriptive statistics for all measures are displayed in Table 2. In order to characterize the sample, assess for the self-reported functions of the IRAP stimuli, and check for systematic differences across the three conditions, a number of analyses were conducted on the self-report measures. Each of the average scores for the PRQ, NRQ, ECRS Positive, and ECRS Negative were subjected to a one-sample *t* test. All were significantly different from zero in the expected directions (all $ps < 0.001$), suggesting that the ratings for the names and the evaluative words entered into the IRAP were viewed in a substantially positive or negative direction. An ANOVA revealed no differences for condition with respect to these measures (all $ps > 0.334$).

IRAP Reliability

Split-half Reliability The overall D-scores for odd and even trials within each trial-type were subjected to Pearson correlations in order to assess split-half reliability for each IRAP in each condition. Spearman–Brown corrected correlations for these analyses, along with their respective 95 % confidence intervals, are displayed in bold font in Table 3. Of the nine separate IRAPs administered (three IRAPs in each of three conditions), seven revealed significant correlations for split-half reliability. Non-significant correlations were obtained for the first IRAP of the real-real-fake condition ($p = 0.234$) and of the real-fake-fake condition ($p = 0.052$).

Test–Retest Reliability Within each condition, each possible pairing of overall D-scores (first and second IRAPs, second and third IRAPs, and first and third IRAPs) were subjected to a collection of Pearson correlations in order to assess test–retest reliability. The correlations for all nine analyses, along with their respective 95 % confidence intervals, are displayed in regular font in Table 3. Significant correlations were obtained in the real-real-real condition between the first and second IRAPs ($p = 0.004$) and between the second and third IRAPs ($p = 0.001$). A significant correlation was also obtained in the real-fake-fake condition between the second and third IRAPs ($p < 0.001$). The remaining six analyses were not significant (all $ps > 0.114$), including four pairings of real and faked IRAPs that would not necessarily be expected to exhibit test–retest reliability.

Analyses for Faking

IRAP Block-Type Latencies A 3 × 3 ANOVA with order (first IRAP, second IRAP, third IRAP) as a within-subjects factor and condition (real-real-real, real-real-fake, real-fake-fake) as a between-subjects factor was conducted for each IRAP block-type. For the pro-friend/anti-enemy block-type, there was no significant main effect of condition ($p = 0.424$), but there was a significant main effect for order [$F(1, 70) = 8.092$, $p = 0.006$, $\eta^2 = 0.104$]. There was also a significant interaction between order and condition [$F(2, 70) = 18.879$, $p < 0.001$, $\eta^2 = 0.350$]. For the pro-enemy/anti-friend block-type, there was no significant main effect of condition ($p = 0.119$), but there was a significant main effect for order [$F(1, 70) = 47.905$, $p < 0.001$, $\eta^2 = 0.166$]. The interaction between order and condition was not significant ($p = 0.752$). A post hoc analysis of the marginal means for order among the pro-enemy/anti-friend block-type revealed significant differences across all IRAP order designations (all $ps < 0.040$); the average block-type latency across the conditions decreased with each subsequent IRAP administration.

Given the interaction of order and condition for the pro-friend/anti-enemy block-type, simple-effects ANOVAs were conducted on each condition, followed by post hoc tests using a Šidák correction when applicable. For the real-real-real condition, the ANOVA was significant ($p < 0.001$), with post hoc comparisons revealing a significant difference between the first and second IRAP administrations ($p < 0.001$) and between the first and third IRAP administrations ($p < 0.001$), but not between the second and third IRAPs ($p = 0.104$). Average latencies were shorter with each successive IRAP, although not significantly so between the second and third IRAPs. For the real-real-fake condition, the ANOVA was significant ($p < 0.001$), with post hoc comparisons revealing significant differences across all IRAP administrations (all $ps < 0.003$): the initial two non-faked IRAPs showed a significant decrease in the average latency from the first to the second IRAP, while the third faked IRAP exhibited a significantly longer average latency, consistent with faking instructions. For the real-fake-fake condition, the ANOVA was not significant ($p = 0.066$); although the faked second and third

Table 2 Means (M) and Standard Deviations (SD) for all Measures for the Full Sample and Each Condition

	Full Sample		Condition					
			real-real-real		real-real-fake		real-fake-fake	
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
PRQ	5.46	1.01	5.52	1.09	5.36	1.14	5.50	0.76
NRQ	−1.85	3.68	−2.41	3.18	−1.35	3.98	−1.60	4.06
ESRS Positive	3.91	1.00	3.87	1.12	3.80	0.94	4.08	0.92
ESRS Negative	−3.46	1.71	−3.14	2.23	−3.48	1.42	−3.88	0.98
FQQ-Following Instructions	8.11	1.83	–	–	7.70	2.11	8.66	1.23
FQQ-Success at Faking	7.51	1.74	–	–	7.00	1.86	8.20	1.32
1st IRAP Consistent Latency	1476	278	1535	320	1458	228	1413	263
1st IRAP Inconsistent Latency	1554	286	1624	351	1548	216	1459	233
2nd IRAP Consistent Latency	1451	276	1426	331	1370	168	1586	254
2nd IRAP Inconsistent Latency	1460	330	1549	447	1429	181	1368	238
3rd IRAP Consistent Latency	1550	330	1376	257	1758	338	1551	274
3rd IRAP Inconsistent Latency	1402	323	1472	369	1416	341	1285	177
1st IRAP D-Score	0.18	0.20	0.19	0.20	0.21	0.21	0.13	0.20
2nd IRAP D-Score	0.00	0.54	0.22	0.29	0.17	0.27	−0.54	0.70
3rd IRAP D-Score	−0.33	0.72	0.19	0.24	−0.70	0.71	−0.63	0.74

Note: PRQ = Positive Relationship Questionnaire; NRQ = Negative Relationship Questionnaire; ESRS = Evaluative Stimulus Rating Scale; FQQ = Faking Quiz and Questionnaire. IRAP latencies are listed in milliseconds. "Consistent" refers to the pro-friend/anti-enemy block-type, and "Inconsistent" refers to the pro-enemy/anti-friend block-type. Full-sample descriptives for the FQQ variables did not include data from the real-real-real condition, as the FQQ was not administered

IRAPs obtained longer average latencies than the non-faked first IRAP, the order designations were not significantly different.

Given the interaction obtained between order and condition for the pro-friend/anti-enemy block-types, simple-effects ANOVAs were also conducted to compare average latencies between each condition for each order designation, followed by post hoc tests using a Šidák correction when applicable. As expected, results of the ANOVA for the first IRAP were not significant ($p = 0.300$). For the second IRAP, the ANOVA was significant ($p = 0.027$), with post hoc comparisons revealing a significant difference only between the real-real-fake and real-fake-fake conditions ($p < 0.027$). The average latency for the faked second IRAP of the real-fake-fake condition was longer than that for the non-faked second IRAP of the real-real-fake condition, but not that for the non-faked second IRAP of the real-real-real condition. For the third IRAP, the ANOVA was significant ($p < 0.001$), with post hoc comparisons revealing a significant difference only between the real-real-real condition and the real-real-fake condition ($p < 0.001$). The average

Table 3 Split-Half and Test–Retest Reliability Estimates and 95 % Confidence Intervals for Each IRAP in Each Condition

		IRAP					
Condition	IRAP	1st		2nd		3rd	
real	1st	**0.70****	(0.45, 0.85)				
real	2nd	0.51**	(0.17, 0.74)	**0.73****	(0.50, 0.87)		
real	3rd	0.30	(−0.04, 0.60)	0.59**	(0.29, 0.79)	**0.63***	(0.34, 0.81)
real	1st	**0.40**	(−0.01, 0.69)				
real	2nd	0.32	(−0.10, 0.64)	**0.62***	(0.29, 0.81)		
fake	3rd	−0.15	(−0.52, 0.27)	0.07	(−0.34, 0.46)	**0.97****	(0.93, 0.99)
real	1st	**0.61**	(0.23, 0.83)				
fake	2nd	−0.06	(−0.49, 0.39)	**0.98****	(0.95, 0.99)		
fake	3rd	−0.05	(−0.48, 0.40)	0.79**	(0.54, 0.91)	**0.97****	(0.93, 0.99)

Note: Values in parentheses display 95 % confidence intervals. Values in **bold** represent split-half reliability estimates after a Spearman-Brown correction. All other values are test–retest reliability estimates. *$p < 0.05$; **$p < 0.01$

latency for the faked third IRAP of the real-real-fake condition was longer than that for the non-faked third IRAP of the real-real-real condition; there was no difference between the faked third IRAP of the real-fake-fake condition and either the non-faked or faked IRAP of the other conditions. In sum, group differences were revealed for order designations in which there was a mixture of non-faked and faked IRAPs, although not all comparisons revealed significant differences.

IRAP D-scores IRAP D-scores were subjected to a 3 × 3 ANOVA with order (first IRAP, second IRAP, third IRAP) as a within-subjects factor and condition (real-real-real, real-real-fake, real-fake-fake) as a between-subjects factor. Results revealed a main effect for order [$F(1, 70) = 56.660$, $p < 0.001$, $\eta^2 = 0.447$] and a main effect for condition [$F(2, 70) = 19.282$, $p < 0.001$, $\eta^2 = 0.355$]. These effects were qualified by an interaction between order and condition [$F(2, 70) = 16.359$, $p < 0.001$, $\eta^2 = 0.319$]. Post hoc ANOVAs using Tukey comparisons were conducted on the D-scores separately for each IRAP administration in order to assess for expected differences based on condition. As expected, no differences were detected among the conditions for the first IRAP (all $ps > 0.434$). For the second IRAP, the real-fake-fake condition was significantly different from both the real-real-fake condition ($p < 0.001$) and the real-real-real condition ($p < 0.001$), while the real-real-real and real-real-fake conditions did not differ significantly ($p = 0.910$); only the faked IRAP was different among the conditions for the second IRAP administration. For the third IRAP, the real-real-real condition was significantly different from the real-real-fake condition ($p < 0.001$) and from the real-fake-fake condition ($p < 0.001$), while the real-real-fake and real-fake-fake conditions were not significantly different ($p = 0.921$); both faked IRAPs were different from the non-faked IRAP among the conditions for the third IRAP administration. In all cases, faked IRAPs generated substantially negative average D-scores (see Fig. 2 for D-score averages, 95 % confidence intervals, and Cohen's *d* effect sizes).

FQQ Analyses Scores for the Following Instructions and Success at Faking questions of the FQQ were subjected to independent-samples *t* tests comparing the two faking conditions. Due to experimenter errors, four individuals in the real-real-fake condition and five participants in the real-fake-fake condition did not receive the faking quiz, leaving 20 and 15 participants, respectively, for analyses. Results were significant for Success at Faking [$t(33) = 2.123$, $p = 0.041$, $d = 0.744$], suggesting that participants who engaged in two real IRAPs before attempting to fake a third IRAP reported less success at faking the IRAP ($M = 7.00$, $SD = 1.86$) than participants who engaged in only one real IRAP ($M = 8.20$, $SD = 1.32$). Results for Following Instructions were not significant ($p = 0.099$). In order to assess for any relationship between faked IRAP performance and explicit estimates of faking performance, the D-scores of the second and third IRAPs and the Following Instructions and Success at Faking questions of the FQQ were subjected to a collection of Pearson correlations. None of these analyses showed significance (all $ps > 0.277$).

Discussion

In contrast to the findings of McKenna and colleagues (2007) suggesting that the IRAP is relatively immune to faking intentions, the current findings suggest that the IRAP is readily faked when participants are provided with straightforward instructions for doing so. These results were produced with IRAPs populated to assess attitudes regarding two individuals with significant personal and emotional salience. This approach to stimulus selection was implemented based on the expectation that an idiographic approach might make it more difficult to modify performance during a task, given the relatively strong personal relevance and disparate evocative valence of the individuals for each participant. The expected pattern of stimulus control for this strategy was exhibited during standard administrations; for each IRAP in each condition that was administered without instructions for faking the task, the average overall D-score was substantially biased in the expected pro-friend/anti-enemy direction. In contrast, for both faking conditions, average D-scores dramatically reversed after faking instructions were provided, showing substantial pro-enemy/anti-friend biases. No evidence indicated that faking occurred spontaneously during any of the non-faking IRAP administrations; rather, it appeared only in conjunction with a request to fake the measure and instructions in how to do so. Furthermore, faking was maintained for a second faked IRAP in the real-fake-fake condition, suggesting that respondents are able to maintain a faking strategy for successive IRAPs. The current findings do not support the contention that idiographic stimuli might protect the validity of the measure when a faking strategy is known. These results appear to replicate the common finding among IAT faking studies that straightforward instructions facilitate faked IAT performance.

Despite clear success at faking, there remained a dissociation between IRAP performance and self-reports of faking performance. The correlations between actual and estimated performance were not significant, a finding that resembles poor self-reported estimations of faked performance among certain IAT studies (Kim, 2003; McDaniel et al., 2009). Even though participants were consciously attempting to fake performance of the task, they may not have been aware of how well they were doing. It seems more likely, however, that respondents were generally aware that they were successful at faking their performance; the average self-reported estimation of success was relatively high ($M = 7.51$ on a scale of 0 to

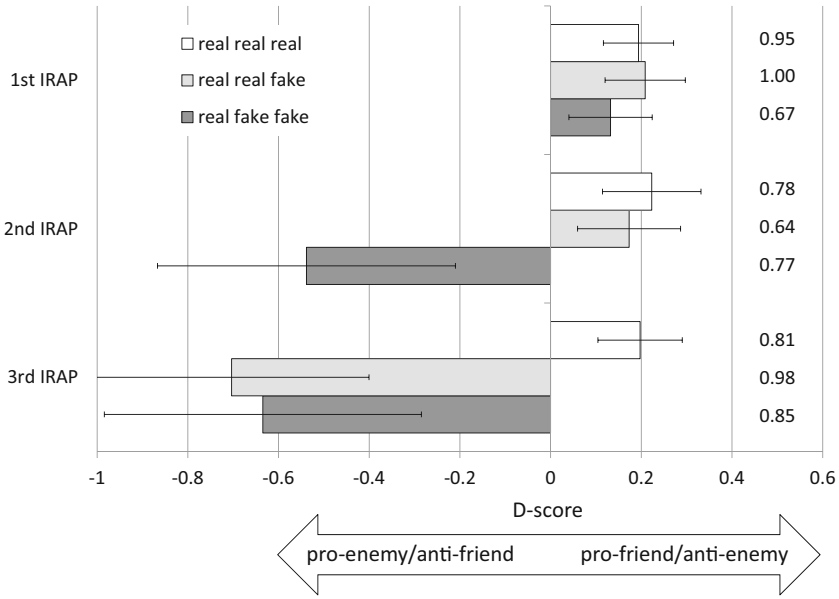

Fig. 2 Average IRAP D-scores, 95 % confidence intervals, and Cohen's *d* effect sizes by administration and condition (effect sizes are listed in the column on the right)

10). The non-significant correlations between the FQQ and the IRAP may instead reflect failure of the FQQ to discriminate participants who were *very* successful from those who were *mildly* successful at faking the task. The amount of IRAP experience appears to have been a factor in these estimates, as participants who first engaged in two real IRAPs reported less success at faking than those who engaged in only one before attempting to fake. In hindsight, a potentially informative condition could have involved the provision of faking instructions before any IRAP was administered. Some IAT faking studies have utilized this approach, with mixed results (e.g., Agosta et al., 2011; Fiedler & Bluemke, 2005). A future IRAP study in which faking instructions are provided to a sample lacking any IRAP experience might address the utility of the IRAP in natural assessment settings where faking motives might be anticipated. Given the relative complexity of the IRAP compared to the IAT, perhaps the IRAP would be less fakable among those lacking experience with the task.

A detailed inspection of the current data suggests that faking the IRAP may be detectable at the level of group comparisons and potentially among individual performance. Faking was revealed by more extreme average D-scores and greater variability in these averages. These extreme averages suggest that participants were quite successful at prolonging their response latencies during the pro-friend/anti-enemy blocks beyond their latencies during the pro-enemy/anti-friend blocks. Analyses of the average latencies for this block-type generally support this idea; longer latency averages were obtained for all faked IRAP administrations, although not all analyses obtained significant differences with non-faked administrations. Greater variability in performance within certain faked administrations may have attenuated the magnitude of these effects. D-scores appeared to be less influenced by this variability, revealing comparable effects across all faked IRAP administrations. Additionally, there was evidence that participants were able to respond more quickly to the pro-enemy/anti-friend blocks over time, which may have been due to familiarity with the procedure, rehearsal of the required answers when following faking instructions for this block-type, or both. In any case, prolonged latencies for one block-type and truncated latencies for the other would increase the absolute value of a D-score, especially if variability remained comparable or decreased. With respect to the idiographic paradigm for IRAP content with the current study, if an evaluator were blind to the self-reported nature of the relationships entered as samples, the evaluator might infer faking based on the extremeness of the overall D-score (e.g., scores greater than an absolute value of 0.5 to 0.7) for an individual performance. For comparisons of known groups (e.g., a forensic population vs. a non-forensic control), conclusions could be based on the extremeness of the average D-score as well as on the variability of the average. Other strategies could also be developed, such as empirically derived indices based on comparisons of latencies for each block-type among known fakers and non-fakers, similar to the strategy embraced by Agosta and colleagues (2011).

If an IRAP administrator wanted to prevent, or at least minimize, faking among respondents who might be motivated to do so, and/or among respondents who were aware of a strategy for faking it, one approach might be to reduce the latency requirement to a stricter standard. Given that previous IRAP research has shown that reduction of the latency requirement resulted in increased biases, internal reliability, and validity of the task (Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010), it seems conceivable that further stringency might protect the integrity of the task when faking

motives are anticipated (e.g., in a forensic setting). One might even take an idiographic approach to the latency requirement: After a respondent's ability with respect to latency has been established during practice blocks, the latency criterion could be strategically reduced to a more challenging standard for test blocks. Whether this would be effective in disrupting faking attempts, or the means of selecting a latency criterion that would do so, remains an empirical matter.

The current study also provides needed data regarding the reliability of the IRAP, especially with respect to test–retest reliability. In general, split-half reliability was acceptable and was comparable to that of other IRAP studies utilizing a latency criterion of 2000 ms (Golijani-Moghaddam et al., 2013). It is not clear whether the idiographic nature of the IRAP in the current study was a factor in these findings, as no nomothetic condition was administered to provide a basis for comparison. Also, two of the nine analyses for split-half reliability were not significant (although one was marginally significant), suggesting that this psychometric property may vary unpredictably across conditions or studies. The results for the real-real-real condition may provide the first published evidence in support of test–retest reliability among IRAPs, absent any experimental intervention designed to modify IRAP performance between administrations. Correlations for test–retest comparisons of IRAPs 1 and 2 and IRAPs 2 and 3 were significant, and were comparable to a compilation of test–retest results for the IAT (Nosek, Greenwald, & Banaji, 2007). However, IRAPs 1 and 3 in this condition were not significantly correlated. IRAPs 1 and 2 in the real-real-fake condition also did not exhibit acceptable test–retest reliability, although the analysis was likely limited by the relatively low split-half reliability results for these administrations. As with split-half reliability, the results for test–retest reliability provide a mixture of findings, suggestive of unpredictable variability in successive administrations of the IRAP across conditions or studies. Of note, both types of reliability were impressive among faking conditions, bearing correlations resembling and even exceeding those for many well-established self-report measures. It seems likely that these results were due to two quite disparate clusters of latencies for each block-type (i.e., relatively long for consistent blocks and short for inconsistent blocks), at least in comparison to the latencies for each block-type in non-faked administrations. The descriptive statistics for latency do indicate larger differences between block-types among faked IRAPs, in some cases with lower variability. The faking instructions may have promoted a relatively narrow and consistent class of response patterns compared to those obtained with standard IRAP instructions. Perhaps the degree of stimulus control exerted by the rules diminished the heterogeneity of responding that would occur with a variety of trials under normal, non-faked conditions. This data may suggest an additional means of detecting faked performance, at least at the level of group comparisons—extremely high split-half reliability may be evidence of faking. In any case, while this study provides some encouraging data regarding the reliability of the IRAP, further investigation of variables that might enhance reliability seems to be warranted.

In sum, it would seem that the IRAP, like other implicit measures, offers a means of minimizing potential demand effects, yet is also not entirely immune to them. Nevertheless, the IRAP appears to hold much promise as an approach to the assessment of complex human behavior. There are many content domains in which fakability may be largely irrelevant, and the means of faking performance on the IRAP may not be obvious to respondents. Rather than evidence of the shortcomings of the task, the current study may serve as a step toward a better understanding of the nature of instructional control over IRAP performance, as well as an improved understanding of factors that enhance the psychometric properties of the measure.

Compliance with Ethical Standards

Conflict of Interest Chad Drake declares that he has no conflict of interest. Kail Seymour declares that he has no conflict of interest. Reza Habib declares that he has no conflict of interest.

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

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

References

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2011). Detecting fakers of the autobiographical IAT. *Applied Cognitive Psychology, 25*(2), 299–306.

Allport, G. W. (1935). Attitudes. In C. Murchinson (Ed.), *A handbook of social psychology* (pp. 798–844). Worcester: Clark University Press.

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the implicit relational assessment procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist, 32*(7), 169–177.

Barnes-Holmes, D., Murphy, A., Barnes-Holmes, Y., & Stewart, I. (2010). The implicit relational assessment procedure: exploring the impact of private versus public contexts and the response latency criterion on pro-white and anti-black stereotyping among white Irish individuals. *The Psychological Record, 60*(1), 57–80.

Barnes-Holmes, D., Murtagh, L., Barnes-Holmes, Y., & Stewart, I. (2010). Using the implicit association test and the implicit relational assessment procedure to measure attitudes toward meat and vegetables in vegetarians and meat-eaters. *Psychological Record, 60*(2), 287–305.

Carpenter, K. M., Martinez, D., Vadhan, N. P., Barnes-Holmes, D., & Nunes, E. V. (2012). Measures of attentional bias and relational

responding are associated with behavioral treatment outcome for cocaine dependence. *The American Journal of Drug and Alcohol Abuse, 38*, 146–154. doi:10.3109/00952990.2011.643986.

Cullen, C., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). The implicit relational assessment procedure (IRAP) and the malleability of ageist attitudes. *The Psychological Record, 59*, 591–620.

De Houwer, J. (2006). What are implicit measures and why are we using them? In R. W. Wiers & A. W. Stacy (Eds.), *Handbook of implicit cognition and addiction* (pp. 11–28). Thousand Oaks: Sage Publications, Inc.. doi:10.4135/9781412976237.n2.

De Houwer, J., Beckers, T., & Moors, A. (2007). Novel attitudes can be faked on the implicit association test. *Journal of Experimental Social Psychology, 43*(6), 972–978.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology, 27*(4), 307–316.

Golijani-Moghaddam, N., Hart, A., & Dawson, D. L. (2013). The implicit relational assessment procedure: emerging reliability and validity data. *Journal of Contextual Behavioral Science, 2*(3), 105–119.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for Black and White patients. *Journal of General Internal Medicine, 22*(9), 1231–1238. doi:10.1007/s11606-007-0258-5.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. doi:10.1037/0033-295X.102.1.4.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*(1), 3–25. doi:10.1037/0033-295X.109.1.3.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. Springer Science & Business Media.

Hughes, S. J., & 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 S. Dymond & B. Roche (Eds.), *Advances in relational frame theory: Research and application* (pp. 97–125). Oakland: Context Press/New Harbinger Publications.

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: propositional and behavioral alternatives. *The Psychological Record, 61*(3), 465–496.

Hussey, I., & Barnes-Holmes, D. (2012). The implicit relational assessment procedure as a measure of implicit depression and the role of psychological flexibility. *Cognitive and Behavioral Practice, 19*(4), 573–582. doi:10.1016/j.cbpra.2012.03.002.

Kim, D. (2003). Voluntary controllability of the implicit association test (IAT). *Social Psychology Quarterly, 66*(1), 83–96. doi:10.2307/3090143.

McDaniel, M. J., Beier, M. E., Perkins, A. W., Goggin, S., & Frankel, B. (2009). An assessment of the fakeability of self-report and implicit personality measures. *Journal of Research in Personality, 43*(4), 682–685.

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*(2), 253–268.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. doi:10.1037/0033-295X.84.3.231.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The implicit association test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). New York: Psychology Press.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Sciences, 15*(4), 152–159.

Paulhus, D. L. (1989). Socially desirable responding: Some new solutions to old problems. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 201–209). New York: Springer-Verlag.

Roddy, S., Stewart, I., & Barnes-Holmes, D. (2010). Anti-fat, pro-slim, or both? Using two reaction-time based measures to assess implicit attitudes to the slim and overweight. *Journal of Health Psychology, 15*(3), 416–425.

Röhner, J., Schröder-Abé, M., & Schütz, A. (2011). Exaggeration is harder than understatement, but practice makes perfect! Faking success in the IAT. *Experimental Psychology, 58*(6), 464–472. doi:10.1027/1618-3169/a000114.

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*(4), 330–338.

Steffens, M. C. (2004). Is the implicit association test immune to faking? *Experimental Psychology, 51*(3), 165–179. doi:10.1027/1618-3169.51.3.165.

Steffens, M. C. (2005). Implicit and explicit attitudes towards lesbians and gay men. *Journal of Homosexuality, 49*(2), 39–66. doi:10.1300/J082v49n02_03.

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*(1), 219–230. doi:10.2466/03.09.22.28.PR0.109.4.219-230.

Tulbure, B. T. (2006). Dissimulating anxiety in front of the implicit association test (IAT). *Cognition, Brain, Behavior, 10*(4), 559–579.

Vecchione, M., Dentale, F., Alessandri, G., & Barbaranelli, C. (2014). Fakability of implicit and explicit measures of the Big five: research findings from organizational settings. *International Journal of Selection and Assessment, 22*(2), 211–218.