Article

Organizational Research Methods
1-28

DOI: 10.1177/1094428117703685
journals.sagepub.com/home/orm

Not Aggressive or Just Faking It? Examining Faking and Faking Detection on the Conditional Reasoning Test of Aggression

Nathan E. Wiita[1], Rustin D. Meyer[2], Elnora D. Kelly[2], and Brian J. Collins[3]

Abstract

Substantial research has been dedicated to examining and combating respondent misrepresentation (i.e., "faking") on personality assessments. Two approaches to combat faking that have garnered particular attention include: (a) designing systems to identify likely fakers and (b) developing difficult-to-fake measures. Consistent with suggestions to combine these strategies, the present article examines a new faking detection system specifically designed for a difficult-to-fake measure (i.e., the Conditional Reasoning Test for Aggression; CRT-A). Four studies (a) help elucidate the conditions under which the CRT-A is fakeable, (b) provide initial construct validity evidence for the faking detection system developed here, (c) examine the effects of faking and faking detection on the CRT-A's criterion-oriented validity, and (d) show that participants identify CRT-based faking detection items at worse-than-chance levels even when they are fully informed about how these items work. Taken together, these studies reinforce the importance of maintaining the indirect nature of CRTs but also show that the faking detection system developed here represents a promising method of identifying those who may have used inside information to manipulate their scores.

Keywords

criterion and predictive validity strategies, reliability and validity, test-retest, exploratory research design

[1]RHR International, Atlanta, GA, USA
[2]Georgia Institute of Technology, Atlanta, GA, USA
[3]University of Southern Mississippi, Hattiesburg, MS, USA

Corresponding Author:
Nathan E. Wiita, RHR International, Ten 10th Street NW, Suite 390, Atlanta, GA 30309, USA.
Email: nwiita@rhrinternational.com

Personality assessment has experienced a resurgence of research attention in the organizational sciences in the past several decades (e.g., Morgeson et al., 2007a; Nicholson et al., 1997). Indeed, various personality tests and constructs have been shown to predict a host of valued outcomes, including job performance (Barrick & Mount, 1991; Hurtz & Donovan, 2000), academic performance (Chamorro-Premuzic & Furnham, 2003), and counterproductive work behavior (Salgado, 2002). Further, personality tests remain popular among scientists and practitioners due to a host of other positive characteristics, including a general lack of adverse impact (Hough, 1998), ease of use and administration (Bing, LeBreton, Davison, Migetz, & James, 2007), and face validity among test takers (Bornstein, Rossner, Hill, & Stepanian, 1994).

Despite the popularity of personality assessments, critics argue that participants can too easily detect their intent (e.g., Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; McFarland & Ryan, 2000; Morgeson et al., 2007a, 2007b; Schwarz, 1999), thereby enabling test takers to manipulate their responses for self-serving purposes (Ziegler, MacCann, & Roberts, 2011). This phenomenon is most commonly referred to as "faking good" or simply "faking." In an effort to approach this issue from multiple angles, the present article examines (a) the conditions under which faking good is possible on a difficult-to-fake personality assessment tool (i.e., the Conditional Reasoning Test for Aggression; CRT-A), (b) the efficacy and effects of a new system designed to detect faking on the CRT-A, and (c) the extent to which participants can identify CRT faking detection items after being taught how they work and what they look like. Such research is important because multipronged approaches to combat faking are especially likely to be effective, but researchers have not yet begun to examine them (Rothstein & Goffin, 2006). Before describing the present efforts to do so, however, it is first important to understand the processes underlying faking and the effects of faking on several scientific and practical issues.

Faking on Self-Report Personality Tests

According to one highly cited model, faking requires that participants (a) have the ability to fake and (b) be motivated to do so (Snell, Sydell, & Lueke, 1999). According to McFarland and Ryan (2000), ability requires that participants know what is being measured and are able to discern which options reflect a low versus high standing on this construct. Given the transparent nature of most traditional self-report measures, many (if not most) respondents should naturally have the ability to fake (Viswesvaran & Ones, 1999). Motivation to fake, on the other hand, can come from a variety of sources, including (but not necessarily limited to) a dishonest personality, a high-stakes testing situation, or a strong experimental manipulation designed to encourage faking. Research suggests that the faking base rate is likely somewhere between 15% (for completely fabricating information) and 62% (for deemphasizing negative attributes) (Donovan, Dwight, & Hurtz, 2003). This state of affairs is potentially problematic because faking serves as a deleterious source of systematic error (Hough, 1998), which may threaten the validity and/or utility of widely used personality instruments (Douglas, McDaniel, & Snell, 1996; Hough, 1998; Mueller-Hanson, Heggestad, & Thornton, 2003; Rossé, Stecher, Miller, & Levin, 1998; Schmit & Ryan, 1993). It should be acknowledged, however, that some researchers (e.g., Ones & Viswesvaran, 1998) argue that in practice, the impact of faking has been overestimated and does not adversely affect construct validity (Smith & Ellington, 2002) while others (e.g., Hogan, 2005) argue that response distortion can be a meaningful reflection of one's personality.

Dealing With Faking

Given the aforementioned faking base rates and the fact that one of the primary goals of valid measurement is to minimize as many extraneous sources of variance as possible, it is not surprising

that dealing with faking has been the focus of much research in the past several decades (for a recent discussion of relevant treatments, see Ziegler et al., 2011). First, however, it is important to determine which type of faking one is most likely to encounter given the tests and testing conditions in question. One highly cited model suggests that faking generally comes in two forms: self-deception and impression management (Paulhus, 1984). Self-deception exists when a participant possesses an inflated sense of self-worth but does not recognize that this perception is inaccurate. As such, the respondent selects overly favorable response options, but this person truly believes that the selected responses are accurate. Impression management, on the other hand, occurs when a respondent knows that he or she possesses one or more traits that may be viewed negatively but attempts to conceal them (or attempts to overstate positive traits) by responding in a self-serving way.

Traditional faking detection scales are primarily intended to detect impression management because the desire to look good creates opportunities for researchers to capitalize on response patterns suggesting that respondents have likely used intentional techniques to fake. One popular method involves the use of "faking scales," which are designed to detect response patterns that are indicative of socially desirable responding (e.g., endorsing virtuous-sounding items that are unlikely to describe honest respondents). A classic example is "I have never taken something that did not belong to me." Endorsing this item suggests that the respondent in question may be attempting to present himself or herself in an overly positive light by selecting a response option that appears to assess a socially desirable trait that accurately describes very few respondents. Thus, the higher number of faking items endorsed by a respondent, the higher the probability that he or she has attempted to fake good. In practice, however, utilizing these scales to "correct" personality scale scores seldom results in increased criterion-oriented validities (Ellingson, Sackett, & Hough, 1999), and they are often limited by ambiguous construct validity (Christiansen, Goffin, Johnston, & Rothstein, 1994; McGrath, Mitchell, Kim, & Hough, 2010).

Faking prevention strategies, on the other hand, typically focus on self-deception because this form of faking is difficult to detect (indeed, even the respondent is not aware that his or her self-perceptions are inaccurate). One such approach is to use indirect measures of relevant individual differences (e.g., Conditional Reasoning Tests: James, 1998; the Implicit Association Test: Greenwald & Banaji, 1995). The idea underlying this category of tests is that various personality traits, attitudes, and so on are manifested through subtle distinctions in the way people think, feel, and behave. As such, participants who are naïve to the nature and intent of these tests generally lack the ability to fake them even if they have the motivation to do so (Alliger, Lilienfeld, & Mitchell, 1996; Bowler, Bowler, & Cope, 2013; LeBreton, Barksdale, Robin, & James, 2007). The Conditional Reasoning Test for Aggression is one such system, but questions remain about its fakeability when its implicit nature is violated and what, if applicable, can be done to detect such faking. Thus, before discussing these issues in depth, it is first necessary to better understand this measurement system.

Conditional Reasoning Tests

CRTs are implicit personality instruments designed to quantify unrecognized biases in reasoning, which are part of a fundamentally different system of personality than what is typically measured via traditional self-reports (James & LeBreton, 2012). Specifically, CRTs are designed to assess what Stanovich and West (2000) called "System 1," which consists of cognitive structures and processes (e.g., motives, defense mechanisms) that exist beyond the reach of introspection (Bing, Kluemper, Davison, Taylor, & Novicevic, 2007; Kihlstrom, 1999; Winter, John, Stewart, Klohnen, & Duncan, 1998). Traditional self-report measures, on the other hand, are typically designed to assess "System 2," which consists of explicit self-perceptions, self-ascribed traits, and are typically assessed via introspective reports about explicit cognitions and motives that exist within one's conscious awareness (Greenwald & Banaji, 1995; James & Rentsch, 2004).

The CRT-A works by capitalizing on the notion that implicitly aggressive individuals must fulfill two motives simultaneously: (a) act in accordance with their implicit motive to aggress and (b) rationalize this behavior in a way that allows them to believe that it complies with societal norms regarding civil behavior. More specifically, James et al. (2005) argue that self-protective implicit biases known as *justification mechanisms* (JMs) "implicitly shape reasoning so as to enhance the rational appeal of behaving aggressively" (p. 73). Thus, over the course of a lifetime, those individuals who possess an implicit motive to aggress develop ways to rationalize aggressive behavior via seemingly plausible explanations. As such, JMs provide the basis for the conditional reasoning measurement system in that aggressive individuals are likely to perceive hostile JMs as more logical than prosocial responses. Conversely, responses associated with hostile JMs will be less likely to appeal to nonaggressive individuals, who will instead tend to favor prosocial JMs.

Measurement of JMs is made possible through the construction of what appear to be inductive reasoning items. Each item ostensibly asks participants to weigh the various response options for their logical merits, but in actuality, each scenario represents an evocative stimulus that permits the expression of one or more JMs. In this system, each scenario is typically followed by four response options: one designed to appeal to participants characterized by a specific JM (e.g., aggressive individuals who use the hostile attribution bias to justify their harmful behaviors), one designed to appeal to participants who are not characterized by that JM (in this case, nonaggressive individuals), and two illogical distractors. The following represents an example of one such Conditional Reasoning item:

> The old saying, "an eye for an eye," means that if someone hurts you, then you should hurt that person back. If you are hit, then you should hit back. If someone burns your house, then you should burn that person's house.
>
> Which of the following is the biggest problem with the "eye for an eye" plan?
>
> a. It tells people to "turn the other cheek."
> b. It offers no way to settle the conflict in a friendly manner.
> c. It can be used only at certain times of the year.
> d. People have to wait until they are attacked before they can strike.

In this particular item, options a and c are illogical distracters and are typically ignored by roughly 95% of participants (LeBreton et al., 2007). Option b represents a prosocial interpretation of the stimulus (i.e., the response that nonaggressive individuals should prefer), and option d exemplifies the "retribution bias" (i.e., the response that individuals with a motive to justify aggression should prefer).

CRTs have been shown to predict diverse outcomes in both workplace and academic settings. For example, the CRT-A has demonstrated significant positive correlations with work absences, dishonesty, employee turnover, theft, and student conduct violations as well as a significant negative correlation with supervisory ratings of overall performance (Frost, Ko, & James, 2007; James et al., 2005; James & LeBreton, 2012). Further, the indirect nature of CRTs makes them generally fake-resistant under standard testing instructions in that—unlike what is traditionally found for self-reports—job applicant mean scores on the CRT-A are *not* significantly different from test norms, job incumbents, or students (e.g., LeBreton et al., 2007; Wiita, Schnure, & James, 2010).

Faking on the CRT-A

Despite their generally faking-resistant nature, contemporary best practices stress the importance of maintaining the implicit nature of CRTs (i.e., ensuring that participants believe that they are taking

an inductive reasoning test) and closely following standard administrative protocols (Bowler et al., 2013; James & LeBreton, 2012). Maintaining the indirect nature of these tests is necessary because recent research suggests that participants are able to artificially manipulate their observed CRT-A scores when provided with inside information about these tests (LeBreton et al., 2007). The LeBreton et al. (2007) study, however, examined participants' ability to "fake bad" under those conditions wherein indirect measurement conditions were violated by demonstrating how the tests functioned and instructing participants to fake bad. Faking good was only examined under conditions wherein their indirect nature was maintained with a field sample of job applicants. As such, we argue that additional research is warranted regarding response distortion on CRTs when the indirect nature is violated.

This line of research is important because as conditional reasoning tests continue to gain popularity, it is possible that information about the nature and structure of these tests will leak to the general public (e.g., through webpages dedicated to teaching respondents how to appear driven by prosocial motives). It is important to note that this sort of cottage industry is not without precedent. In fact, the history of high-stakes testing is, at least in part, a history of how to beat high-stakes tests (Madaus, Russell, & Higgins, 2009). Perhaps the earliest example of someone with inside information helping test takers develop insights and strategies dates back to the mid-13th century, where "a master in the Arts Faculty at the University of Paris between 1230 and 1240 . . . listed the questions most frequently asked during oral examinations and succinct answers to them" (Madaus et al., p. 142) and "during the seventeenth century at Oxford, books containing previous exam questions became available to students." More recently, a survey of educators found that "forty per cent of teachers reported that colleagues found ways to raise the state-mandated test scores without actually improving student learning" (Madaus et al., p. 155), and in the past decade, at least two major international efforts dedicated to helping students cheat on the GRE have been uncovered (Tyre, 2016).

With respect to CRTs, information about how these tests work has been featured in popular media outlets such as National Public Radio (NPR), *Maxim* magazine, and the *New York Times*. To the extent that inside information about CRTs continues to leak, motivated test takers may someday be able to acquire enough inside information about CRTs to artificially alter their scores. Such pages already exist for self-report personality and intelligence measures (e.g., Chamorro-Premuzic, 2015).

Specifically, we posit here that respondents will be able to fake CRTs if they know that (a) they are tests of personality, not tests of inductive reasoning; (b) two items are illogical distractors; and (c) the two remaining items assess opposite poles of the personality trait in question (aggression, in the case of the CRT-A). When these three pieces of information are in place, participants are likely to meet McFarland and Ryan's (2000) criteria for having the ability to fake (i.e., knowing what construct is being measured and accurately determining which response options reflect a low vs. high standing on this construct), thereby only additionally requiring the motivation to fake in order to successfully do so.

A CRT-Based Faking Detection System

We also argue, however, that participants who have both the ability and motivation to fake CRTs will use a qualitatively different response strategy than those who respond under typical testing conditions: Namely, fakers will focus their efforts on eliminating from consideration the two illogical responses and comparing the two remaining options with respect to the characteristic in question. This shift from implicit reasoning to explicit comparisons, however, also creates an opportunity for those who are interested in detecting faking on CRTs. Namely, it suggests that the key to detecting faking on CRTs lies in developing special items that contain response options that

differentially appeal to those who are responding under normal conditions and those who have the ability and motivation to fake.

The system developed here attempts to achieve this goal by using items that mimic the appearance of traditional CRTs but contain one aggressive sounding response option that is obviously logically superior to an ostensibly prosocial alternative. Thus, participants who are not faking should select the aggressive sounding response as it is designed to appear to be the most logical choice. When respondents know how to fake and are motivated to do so, however, they should endorse the less realistic but prosocial sounding response option because their item selection strategy will focus not on judging how realistic each option is but on determining whether each of the two logical response options appears either aggressive or prosocial. The following is an actual CRT-A faking detection item tested in the present article:

> War often mobilizes large groups of people toward one effort: victory. This includes people on the home front, who work to support soldiers, as well as the soldiers themselves. History tells us that those who are not fully committed to their cause will not be victorious. As a consequence, soldiers are trained to act violently on the battlefield to stay alive and achieve victory. Indeed, it has been famously remarked that "war is hell."
>
> Which of the following is the most logical conclusion based on the above?
>
> a. As a soldier in violent combat, you sometimes must kill or be killed.
> b. Soldiers would be more compassionate if they were never trained to be violent.
> c. The older soldiers get, the more medical assistance they need.
> d. Soldiers behave more violently in mountainous areas.

First—and consistent with actual CRT-A item structuring—two response options, in this case options c and d, are the distractor response choices. They are intended to be illogical, irrelevant, and/or otherwise not following the statements from which one is to draw a conclusion. This leaves two potential logical response choices, a and b. For respondents who believe they are solving an inductive reasoning problem (i.e., those who do not have the knowledge necessary to fake), option a is designed to be logically superior to option b, given human nature, the reality of war, and the reality world in which we live. For respondents who are attempting to fake (i.e., those presumably seeking the most prosocially worded response option), b is designed to be more appealing. In other words, although option a contains relatively harsher verbiage, it is designed to be logically superior to option b, which is instead designed to appear more prosocially worded but generally unrealistic. Responses like option b in this item are termed *honeypots* because they are similar to a phenomenon of the same name in the computer security industry wherein hackers exploit a seemingly unsecured network and in so doing unwittingly expose their ill intent to those who have a vested interest in thwarting their efforts.

The Present Article

The four studies presented in the present article utilize the foundational issues described thus far to examine several issues associated with faking the CRT-A as well as the detection thereof. To provide a feasibility test for a CRT-A faking detection scale, we follow the initial steps outlined by Hinkin (1998). More specifically, Study 1 uses a between-subjects design to examine both faking and faking detection across three instructional sets. Study 2 uses a within-subjects design to examine the efficacy of correctly identifying participants who successfully learned how to fake. Study 3 examines differences in the criterion-oriented validity of the CRT-A under standard conditions compared to faking conditions. Finally, Study 4 examines whether respondents can identify our faking detection items after being told how they work.

Study 1: Tests of Between-Persons Effects

As outlined previously, the effects of learning how CRTs work on faking have only been tested under those conditions wherein participants were asked to fake bad. Although one could argue that if participants can fake bad they can surely fake good, this may not necessarily be the case because aggression's low base rates (typically 5%-10% of participants) may create floor effects that prevent significant levels of faking. Despite this potentially limiting practical issue, we posit here that participants who are motivated and able to fake good on the CRT-A will be able to do so.

Hypothesis 1: Participants who have the motivation and ability to fake the CRT-A will endorse significantly fewer aggressive responses than those who lack either the ability or motivation to fake this test.

Further, given the CRT-based theory of faking and faking detection outlined previously, we also predict the following:

Hypothesis 2: Participants who have the ability and motivation to fake the CRT-A will endorse significantly more honeypot responses than those who lack either the ability or motivation to fake this test.

Study 1 Method

Participants

Study 1 utilized a total of 162 undergraduate participants from a midsized technical institution in the Southeastern United States. This sample was 40.7% female, with an average age of 20 years old (range = 18-32). With respect to ethnicity, this sample was 59.9% Caucasian, 3.7% African American, 6.2% Hispanic, 27.2% Asian, and 3.1% "mixed/other."

Materials

Conditional reasoning test for aggression. The CRT-A is a 22-item implicit measure of aggression, which is scored by summing the number of hostile responses endorsed by each participant. Higher scores indicate higher implicit readiness to justify aggression. Scores of 8 or greater are used as a "nonarbitrary cutting point for distinguishing aggressive from nonaggressive individuals on the CRT-A" (James & LeBreton, 2012, p. 151). Additionally, three inductive reasoning items are included to mask the intent of the measure. James, McIntyre, Glisson, Bowler, and Mitchell (2004) reported a .76 internal consistency estimate of reliability and a .82 alternative forms estimate of reliability. As per standard recommendations (James & McIntyre, 2000), participants who endorsed five or more illogical options were eliminated from the data set prior to conducting any substantive analyses.

CRT faking items. Eleven faking detection items were developed for the present article based on the framework outlined previously. Higher scores indicate a higher probability of attempting to fake good on the CRT-A (i.e., one point was given for each honeypot response option that was endorsed; no points were given if any other response option was selected). They were developed in accordance with Hinkin (1998) utilizing the deductive approach to item generation noted as his first step in scale development.

Personality research form-E. The Personality Research Form-E (PRF-E) is a self-report measure of several dimensions of personality. Five of its 22 subscales were used here. First, the Aggression

subscale and the Social Desirability (SDR) scale (designed to detect faking) were used as a manipulation check to gauge the efficacy of the faking instructions described subsequently. Second, the Infrequency subscale was used to eliminate those participants who appeared to be responding carelessly. Lastly, the Impulsivity subscale was used to help examine the present faking detection system's nomological network. Each of these subscales has 16 true-false items. Jackson (1974) reported test-retest reliabilities for Aggression (.85), SDR (.81), Impulsivity (.81), and Infrequency (.46).

Procedure

The CRT-A (with the 11 additional response distortion items randomly embedded) and the five aforementioned PRF-E subscales were administered under three conditions. These conditions are intended to permit direct tests of the fakeability of the CRT-A under various levels of information and provide an initial feasibility test of the faking detection method developed here.

In Condition 1 (i.e., control), participants were given standard instructions to complete all measures. In Condition 2 (i.e., fake good), participants were instructed to complete all measures as though they were applying for a job they really wanted (i.e., "put your best foot forward"). Specific instructions were consistent with those used in previous research utilizing this methodology (e.g., Martin, Bowen, & Hunt, 2002; Mueller-Hanson et al., 2006; Peeters & Lievens, 2005). In Condition 3 (i.e., disclose/fake good), instructions were the same as the fake good condition with the exception that participants were provided with inside information about the CRT-A. Specifically, participants were informed that although the CRT-A looks like a test of inductive reasoning, it is actually a personality test that measures aggression by capitalizing on unrecognized information processing biases. They were also informed that out of the four possible response alternatives, two are illogical distractors, one is intended to appeal to aggressive individuals, and one is intended to appeal to nonaggressive individuals. Further, three hypothetical CRT-A sample items were provided to participants where—as a group—they were shown which response options represented the two distractors, which represented the aggressive option, and which represented the nonaggressive option.

Study 1 Results

Manipulation Check

Consistent with the intended manipulation, self-reported aggression scores were significantly lower when participants were instructed to fake good (Condition 2: $M = 2.50$, $SD = 1.95$) compared to when participants were asked to respond honestly (Condition 1, $M = 4.13$, $SD = 1.52$), $F(1, 107) = 23.66$, $p < .001$ (partial $\eta^2 = .18$). This change (i.e., Cohen's $d = .94$) lies within the average upper limit of the 90% meta-analytic confidence intervals reported by Viswesvaran and Ones (1999) for between-person, fake good manipulations of self-report personality scales. This pattern of results suggests that any resistance to faking shown by the CRT-A in Condition 2 is likely due to this measure's difficult-to-fake nature (LeBreton et al., 2007) as opposed to an impotent manipulation (see Table 1 for descriptive statistics).

Also consistent with the intended manipulation, scores on the PRF-SDR scale were significantly higher when participants were instructed to fake good (Condition 2: $M = 7.70$, $SD = .76$) compared to when participants were asked to respond honestly (Condition 1: $M = 7.17$, $SD = 1.01$), $F(1, 108) = 9.83$, $p < .01$ (partial $\eta^2 = .08$). Again, these results suggest that subsequent analyses are likely to provide a meaningful test of the fakeability of the CRT-A and the detection thereof.

Table 1. Study 1 Descriptive Statistics.

	Condition 1 (Control)						Condition 2 (Fake Good)						Condition 3 (Disclose/Fake Good)					
	N	*M*	*SD*	Skew	Kurt	α	*N*	*M*	*SD*	Skew	Kurt	α	*N*	*M*	*SD*	Skew	Kurt	α
CRT-A	52	4.90	2.64	.35	–.67	.81	55	4.53	2.64	.45	–.30	.81	51	1.29	1.29	1.24	1.0	.86
CRT-F	53	1.51	1.77	2.54	9.57	.92	56	1.95	1.42	.53	–.08	.78	52	7.06	2.82	–.68	–.23	.92
PRF-SDR	53	7.17	1.01	–.47	–.19	.32	57	7.70	.76	–3.05	10.46	.71	52	7.83	.64	–4.31	24.2	.67
PRF-Agg	53	4.13	1.52	–.33	–.32	.40	56	2.50	1.95	.82	–.13	.69	52	1.42	1.42	.74	–.48	.57
PRF-Impuls	52	3.40	2.09	.05	–.69	.69	57	1.44	1.97	1.86	3.51	.82	51	.76	1.24	1.97	3.37	.63

Note: CRT-A = Conditional Reasoning Test of Aggression; CRT-F = Conditional Reasoning Test faking detection scale; PRF-SDR = Personality Research Form Socially Desirable Responding scale; PRF-Agg = Personality Research Form Aggression scale; PRF-Impuls = Personality Research Form Impulsivity scale.

Hypothesis Tests

Consistent with Hypothesis 1, the CRT-A showed significantly lower values when participants were asked to fake good and were taught how this test works (Condition 3: $M = 1.29$, $SD = 1.29$) compared to when participants were asked to fake good but not taught how this test works (Condition 2: $M = 4.53$, $SD = 2.64$), $F(1, 104) = 62.77$, $p < .001$ (partial $\eta^2 = .38$) and compared to the control condition (Condition 1: $M = 4.90$, SD = 2.64). As further expected, results also showed that participants required inside information about how the CRT-A works in order to significantly fake good in that the fake good mean (Condition 2: $M = 4.53$, SD = 2.64) was not significantly less than the control mean (Condition 1: $M = 4.90$, $SD = 2.64$), $F(1, 105) = .54$, $p > .05$ (partial $\eta^2 = .01$).

Consistent with Hypothesis 2, significantly more honeypot response options were endorsed when participants were not only asked to fake but were also taught how the CRT-A works (Condition 3: $M = 7.06$, $SD = 2.82$) compared to when participants were asked to fake but not taught how this test works (Condition 2: $M = 1.95$, $SD = 1.42$), $F(1, 106) = 144.80$, $p < .001$ (partial $\eta^2 = .58$) and compared to the control condition (Condition 1: $M = 1.51$, $SD = 1.77$), $F(1, 103) = 146.6$, $p < .001$ (partial $\eta^2 = .59$). There was no significant difference between the number of honeypot responses selected when participants were asked to fake but not taught how the test works (Condition 2: $M = 1.95$, $SD = 1.42$) and the number of such options selected by participants in the control condition (Condition 1: $M = 1.51$, $SD = 1.78$), $F(1, 105) = 2.03$, $p > .05$ (partial $\eta^2 = .01$).

Supplementary Analyses

It is also important to examine a host of additional psychometric considerations that are sometimes associated with faking and faking detection methods (Griffith & Peterson, 2008). Thus, this section provides several supplementary analyses designed to better understand two broad secondary themes. The first theme focuses on the present faking system's nomological network given that one of the primary criticisms of self-report faking detection scales is that they sometimes conflate faking with substantive personality characteristics (McCrae & Costa, 1983). The second theme pertains to the effects of faking on the CRT-A itself given that previous research (e.g., Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; LeBreton et al., 2007; Viswesvaran & Ones, 1999; Zickar & Robie, 1999) suggests that faking has four primary psychometric effects on the faked scale: (a) mean shifts (see Hypothesis 1), (b) range restriction, (c) increased skew, and (d) increased reliability.

Nomological network. The first nomological issue examined here is the correlation between CRT-based faking detection items and self-report faking detection items. Consistent with the numerous substantive differences between these two systems, results suggest that scores on the CRT-A faking detection system developed here and the PRF-SDR were not significantly correlated, $r(51) = -.17$, $p > .05$, under standard testing conditions (Condition 1). Scores were positively and significantly correlated with each other, however, when respondents were instructed to fake good (Condition 2), $r(54) = .32$, $p < .05$, but returned to being uncorrelated when participants were asked to fake good and were informed how the CRT-A operates (Condition 3), $r(50) = -.09$, $p > .05$. See Tables 2 and 3 for a summary correlation table. The Condition 1 and Condition 2 correlations were significantly different from each other, $Z = -2.55$, $p < .01$, as were the same correlations for Condition 3 and Condition 2, $Z = -2.13$, $p < .05$. In contrast, the correlations of the CRT-A faking detection system and the PRF-SDR did not differ significantly between Condition 1 and Condition 3, $Z = -.041$, $p < .34$.

The second nomological issue examined here is the CRT-based faking detection system's relationships with various substantive personal characteristics, which was investigated by examining

Table 2. Study 1 Exploratory Tests of the CRT-Based Faking Detection System's Nomological Network (Conditions 1 and 2).

	1	2	3	4	5	6	7	8	9	10	11
1. Age	—	0.16	–0.03	–0.24	0.07	0.18	–0.01	0.15	–0.17	–0.03	–0.14
2. Gender	0.19	—	0.25	–0.38**	–0.58**	0.00	0.24	0.04	0.01	0.19	–0.27*
3. SAT-Math	0.26	0.29	—	0.16	0.05	–0.05	0.10	–0.08	–0.06	0.04	–0.06
4. SAT-Verbal	–0.01	0.06	–0.02	—	0.59**	0.00	–0.18	–0.07	–0.10	–0.29	–0.03
5. ACT	–0.12	0.25	0.64**	0.13	—	–0.18	–0.40	0.36	–0.37	–0.17	0.31
6. GPA	–0.41**	0.00	–0.03	0.05	0.21	—	–0.24	0.14	–0.14	–0.10	–0.05
7. PRF-Agg	0.00	0.26	–0.06	0.15	–0.25	–0.04	—	–0.45***	0.47***	0.45***	–0.35**
8. PRF-SDR	–0.15	–0.04	–0.14	0.26	–0.18	0.13	0.30*	—	–0.61***	–0.21	0.32*
9. PRF-Impuls	–0.20	0.06	0.29	0.10	0.04	–0.06	0.27	0.09	—	0.10	–0.07
10. CRT-A	0.05	0.23	–0.03	–0.20	–0.03	–0.03	0.08	–0.14	0.12	—	–0.34*
11. CRT-F	–0.13	0.05	–0.04	–0.24	0.09	0.24	0.07	–0.17	0.02	–0.14	—

Note: Gender coded as 1 = female, 2 = male. Values below diagonal represent correlations from Condition 1 (i.e., control; $N = 53$). Values above diagonal represent correlations from Condition 2 (i.e., fake good; $N = 57$). CRT-A = Conditional Reasoning Test of Aggression; CRT-F = Conditional Reasoning Test faking detection scale; PRF-SDR = Personality Research Form Socially Desirable Responding scale; PRF-Agg = Personality Research Form Aggression scale; PRF-Impuls = Personality Research Form Impulsivity scale.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 3. Study 1 Exploratory Test of the CRT-Based Faking Detection System's Nomological Network (Condition 3: Disclose/Fake Good).

	1	2	3	4	5	6	7	8	9	10	11
1. Age	—										
2. Gender	–0.06	—									
3. SAT-Math	–0.27	–0.02	—								
4. SAT-Verbal	–0.13	–0.02	0.33*	—							
5. ACT	–0.24	0.26	0.68**	0.48	—						
6. GPA	–0.46**	–0.01	0.11	–0.14	0.04	—					
7. PRF-Agg	0.11	–0.19	0.09	–0.14	–0.24	0.12	**0.57**				
8. PRF-SDR	–0.02	0.06	–0.16	0.1	0.04	–0.13	–0.22	**0.67**			
9. PRF-Impuls	–0.13	–0.11	0.23	–0.30	–0.25	0.24	0.55***	–0.08	**0.63**		
10. CRT-A	–0.21	0.05	0.22	–0.02	0.10	0.17	0.21	–0.03	0.32*	**0.86**	
11. CRT-F	0.13	0.08	–0.24	0.18	–0.28	–0.23	–0.31*	–0.09	–0.39**	–0.48***	**0.92**

Note: $N = 52$. Scale reliability values reported in bold on the diagonal where applicable. Gender coded as 1 = female, 2 = male. CRT-A = Conditional Reasoning Test of Aggression; CRT-F = Conditional Reasoning Test faking detection scale; PRF-SDR = Personality Research Form Socially Desirable Responding scale; PRF-Agg = Personality Research Form Aggression scale; PRF-Impuls = Personality Research Form Impulsivity scale.
*$p < .05$. **$p < .01$. ***$p < .001$.

bivariate correlations between CRT faking scores and various individual differences under the three conditions examined here. This is an important issue because to operate in a truly convincing fashion, scores on faking detection systems should not correlate significantly with either surface- or deep-level test-taker characteristics. Consistent with this perspective, CRT faking scores in the control condition showed no significant correlations with any of the other variables included in this study. In the fake good condition, however, CRT faking scores showed significant negative correlations with gender ($r = –.27, p < .05$) and PRF-Aggression ($r = –.35, p < .01$). Further, in the disclose/fake good condition, the significant negative correlation with PRF-Aggression was maintained

($r = -.31$, $p < .05$), and a significant negative correlation emerged with PRF-Impulsivity ($r = -.39$, $p < .01$) (see Tables 2 and 3 for all relevant correlations). It is also important to note there were no statistically significant correlations between proxies of GMA and either CRT-A or CRT faking scale scores.

Psychometric effects. The expected pattern of results would be an increase in range restriction when faking instructions are administered, presumably due to the experimental manipulation (e.g., "respond as though you greatly desire the job") rather shifting individual differences. Regarding range restriction, Levene's test of homogeneity suggests that significantly less variance exists in the CRT-A when participants have both the motivation and ability to fake (Condition 3: $s^2_{\text{CRT-A}} = 1.66$) compared to when participants have the motivation but not the ability to do so (Condition 2: $s^2_{\text{CRT-A}} = 6.97$), $F(1, 104) = 23.69$, $p < .001$. When the same analysis is conducted on those who theoretically have neither the motivation to fake nor the ability to do so, the effects are essentially identical (Condition 1: $s^2_{\text{CRT-A}} = 7.00$), $F(1, 101) = 23.29$, $p < .001$.

Regarding differences in skew, the present study provided support for the prediction that this distributional characteristic would be greater than that expected due to chance factors alone in those conditions wherein participants were able to successfully fake. Specifically, observed skew values on the CRT-A for each condition were: Condition 1 = .35 ($SE = .33$), $z = 1.05$, $p > .05$; Condition 2 = .45 ($SE = .32$), $z = 1.35$, $p > .05$; Condition 3 = 1.24 ($SE = .33$), $z = 3.76$ $p < .001$.

Regarding differences in reliability, the present results are consistent with previous research in that internal consistency estimates for the CRT-A *increased* after participants learned how to fake this test. Specifically, the Condition 1 reliability estimate was .81, the Condition 2 estimate was .81, and the Condition 3 estimate was .86. The PRF-E Aggression scale demonstrated a similar pattern in that the Condition 1 Cronbach's alpha estimate was .40, the Condition 2 estimate was .69, and the Condition 3 estimate was .57.

Study 1 Discussion

Study 1 provided support for the notion that participants are able to artificially manipulate their observed CRT-A score in a manner that reduces perceived aggressiveness as long as they are motivated to do so and also possess inside information about the nature and structure of CRTs. This finding supplements previous research showing that participants can *inflate* their CRT-A scores under conditions of full disclosure (LeBreton et al., 2007) by demonstrating that participants are also able to *deflate* their CRT-A scores under similar conditions. This finding is important because it suggests that efforts to teach test takers how to conceal their aggressiveness when responding to the CRT-A are likely to be successful despite the potential for substantial floor effects caused by implicit aggression's naturally low base rate (James & LeBreton, 2012).

Study 1 also provided preliminary support for the efficacy of the CRT-based faking detection system. As described previously, this system consists of items that were designed to mimic standard CRT items but contain honeypot response options intended to only appeal to participants who were motivated and able to fake the CRT-A. Results suggest that this system worked as planned in that participants who were asked to fake the CRT-A and taught how to do so (Condition 3) endorsed significantly more honeypot response options than those who were either asked to respond honestly (Condition 1) or asked to fake but were naïve to the nature and structure of CRTs (Condition 2). Additionally, honeypot response options were generally unappealing to participants who were naïve to the nature and structure of CRTs. Despite these generally promising trends, however, more fine-grained data and analyses are needed to fully test the efficacy of the CRT-A faking detection system.

Study 2: Tests of Within-Person Effects

Although the between-person effects found in Study 1 represent an important test of the fakeability of the CRT-A as well as a key initial examination into the general merit of the CRT-A faking detection method, it is also important to recognize that within-person analyses are a crucial form of evidence to report when judging whether a given faking detection system performs as intended (Griffith & Peterson, 2008). Indeed, some have argued that within-person designs generally provide superior tests of faking and its effects because between-groups designs can be adversely affected if sampling error results in nonequivalent groups, whereas this issue is not of concern for within-person designs (Viswesvaran & Ones, 1999).

One of the other main benefits of within-person designs is that they permit examinations into the extent to which specific participants' response patterns change as a function of their ability and motivation to fake. Given the difficult-to-fake nature of the CRT-A (LeBreton et al., 2007) and the fact that many participants will naturally have low CRT-A scores (James & LeBreton, 2012), it is important to not assume that the significant mean effects observed in Study 1 necessarily represent evidence that the faking detection system developed here will correctly classify both fakers and nonfakers. For instance, a significant mean decrease may be due to a small number of participants greatly deflating their CRT-A scores (thereby suggesting that these tests are rather difficult for the average test taker to fake) or a large number of participants deflating their CRT-A scores by a rather small amount (thereby suggesting that faking is unlikely to have important practical effects). As such, we test the following research questions:

> *Research Question 1a:* What proportion of test takers will be able to successfully fake the CRT-A after being provided with inside information about how this test works?
> *Research Question 1b:* By how much will the average test taker be able to successfully fake the CRT-A after being provided with inside information about how this test works?

Another benefit of within-person designs is that they permit the calculation of hit rate analyses (i.e., the percentage of participants who were identified correctly as fakers and nonfakers). Four categories are key here: (a) *True positives* are those participants who significantly deflate their CRT-A score after learning how this tests work and are correctly flagged as likely fakers using our faking detection system, (b) *true negatives* are those participants who do not successfully deflate their aggression score and are not flagged as fakers, (c) *false negatives* are those participants whose successful faking efforts go undetected, and (d) *false positives* are those who do not fake but are flagged as if they did. With these definitions in place, the following research question is examined in an effort to address the overall accuracy of the present faking detection system:

> *Research Question 2:* What percentage of participants will be correctly categorized as true positives, true negatives, false positives, and false negatives using the CRT-based faking detection method developed here?

Study 2 Method

Participants

Forty-six (46) employed master's of business administration (MBA) students from a university in the Southeastern United States (different from that used in Study 1) participated in a within-person test of the 11 CRT-A faking detection items. This sample was approximately 42.9% female, with an average age of 30 years old (range = 23-49, SD = 4.9). Participants worked for their current employers for an average of approximately 6 years. One participant was dropped from subsequent

Table 4. Study 2 Descriptive Statistics.

	Time 1 (Control)					Time 2 (Disclose/Fake Good)				
	M	*SD*	Skew	Kurt	α	*M*	*SD*	Skew	Kurt	α
CRT-A	3.67	1.81	0.38	0.14	0.56	1.29	1.73	1.94	4.38	0.96
CRT-F	1.11	1.11	0.71	–0.41	0.76	6.87	2.32	–0.52	–0.68	0.87

Note: $N = 45$. CRT-A = Conditional Reasoning Test of Aggression; CRT-F = Conditional Reasoning Test faking detection scale.

analyses due to extreme item skipping in both conditions, thereby yielding an initial functional *N* of 45 (although, as described subsequently, this *N* is further reduced by the fact that some participants' very low original CRT-A scores did not allow them to fake).

Materials

Conditional reasoning test for aggression. The same measure used in Study 1, including the 11 CRT-A faking detection items, was used in this study. No participants endorsed five or more illogical response options, so all participants' data were retained for subsequent analyses.

Procedure

Participants at Time 1 took the CRT-A under standard testing conditions (i.e., control). One week later (Time 2), the same participants were taught how the CRT-A worked and asked to fake good (i.e., disclose/fake good). All participant instructions were identical to those used in Study 1, Condition 1 and Study 1, Condition 3 (respectively). It was not possible to counterbalance these conditions while still obtaining a valid within-person test because if participants were first informed of the intent of the CRT-A, they would then be unable to take it under truly naïve conditions due to this inside knowledge.

Study 2 Results

Manipulation Check

Consistent with the intended manipulation (as well as the results of Study 1), participants endorsed significantly more honeypot response options at Time 2 ($M = 6.87$, $SD = 2.32$) compared to Time 1 ($M = 1.11$, SD = 1.11), $F(1, 44) = 227.50$, $p < .001$ (partial $\eta^2 = .84$). Said differently, when participants were naïve to the nature and structure of the CRT-A, they endorsed an average of roughly one honeypot response option, whereas after being provided with inside information about the CRT-A and asked to fake good, they endorsed an average of nearly seven of these options. Further, the presence versus absence of this background information (plus any variance attributed to threats to validity such as history or maturation) explained nearly 84% of the variance in one's CRT faking score. See Table 4 for relevant descriptive statistics.

Tests of Research Questions

Research Question 1a focused on estimating the proportion of test takers who were able to successfully fake the CRT-A. In order to test this question, however, it is first necessary to determine how much participants must deflate their CRT-A scores to constitute faking. The most liberal estimate would be to simply say that any reduction in one's CRT-A score from Time 1 to Time 2 constitutes

faking, but such a cut-score would run the risk of capitalizing on chance-based fluctuations. A more empirically defensible approach is to use the standard error of measurement (SEM) to quantify the amount of fluctuation that can be expected based on chance factors alone (Harvill, 1991). SEM values for the CRT-A in those conditions wherein participants were naïve to the nature and structure of the CRT-A were 1.15 (Study 1, Condition 1), 1.15 (Study 1, Condition 2), and 1.20 (Study 2, Time 1). Due to the facts that (a) change scores of 1.0 are within the SEM (i.e., CRT-A scores can be expected to fluctuate by more than 1 point due to chance factors alone) and (b) CRT-A scores can only be whole numbers, we had to "round up" by categorizing those participants who were able to reduce their CRT-A score by 2 or more points as "fakers."

With this standard set, it is now possible to directly examine Research Question 1a. The first step in this process is to eliminate from consideration any participants whose CRT-A score at Time 1 was zero or one ($N = 5$) because it would be functionally impossible for these participants to fake (i.e., they cannot deflate their CRT-A score by two or more points). After eliminating these participants from consideration, frequency analyses suggest that 77.5% of participants ($N = 31$ of 40) were able to significantly reduce their CRT-A score after learning how this test works. With respect to Research Question 1b, the present analyses suggest that participants were able to reduce their CRT-A scores by nearly 65.8% on average, from a Time 1 mean of 4.03 ($SD = 1.58$) to a Time 2 mean of 1.38 ($SD = 1.81$), $F(1, 39) = 69.73$, $p < .001$ (partial $\eta^2 = .64$).

Research Question 2 focused on the accuracy of the faking detection system developed here. Again, participants who were able to deflate their CRT-A score by two or more points from Time 1 to Time 2 were categorized as fakers. These findings raise the question: What is the best approach to setting a CRT-F cut-score that can be used to signal that a given respondent has endorsed a suspicious number of honeypot response items? Absent sufficient data to employ more sophisticated techniques, this "suspicion" cut-score was created by flagging participants who selected honeypot response options at a rate that was three standard deviations higher than that of the average participant under naïve conditions. Consequently, participants were suspected of being fakers if they endorsed six or more honeypot response options. Using this standard, 75% of participants were correctly classified, whereas 25% were incorrectly classified. Specifically, 62.5% ($N = 25$) were categorized as true positives, 12.5% ($N = 5$) were true negatives, 15% ($N = 6$) were false negatives, and 10% ($N = 4$) were false positives.

Study 2: Discussion

The present results indicate that once test takers have inside information about how CRTs work, more than 75% were able to fake good at rates that may cause problems for researchers and/or practitioners (addressed more directly in Study 3). In terms of most accurately identifying these likely fakers, results suggest that a CRT-F cutoff of five should be favored. When this standard is applied, 75% of participants were correctly classified, and only 25% were erroneously identified as likely fakers. This last value, however, should be tempered by the fact that some participants were likely motivated to fake but were not able to do so. Thus, in real-world testing situations, such participants would still be viewed as undesirable.

To put these success rates in perspective, a similar feasibility test of a novel method of faking detection (Kuncel & Borneman, 2007) was able to detect between 62% and 78% of fakers, with false positive rates between 14% and 21%. Using different cut-scores, these authors were able to reduce their false positive rate to 1%, but their correct classification rate dropped to between 20% and 37%. In the clinical realm, however, efforts to detect faking good in the Millon Multiaxial Inventory-II have been successful at rates of 72% (Bagby, Gillis, Toner, & Goldberg, 1991), although this

estimate was based on a between-subjects design wherein participants were allowed to self-select into honest, fake bad, or fake good conditions, thereby providing a substantially different test than that performed in the present study. Taken together, the 75% overall success rate found here provides reason for initial optimism, though it is also important to examine the effects of faking and faking detection on more applied issues.

Study 3: Criterion-Oriented Validity

Studies 1 and 2 showed that the CRT-A is susceptible to faking when respondents know how this test works but that the CRT faking detection system developed here can be used to identify participants who have attempted to distort their scores. Past faking research has also demonstrated that intentional response distortion can (though does not always) have a deleterious effect on the criterion-oriented validity of relevant measures (Douglas et al., 1996). Thus, it is important to examine the effects of faking on the criterion-oriented validity of the CRT-A.

As discussed previously, the CRT-A predicts various forms of aggressive behavior, ranging from relatively subtle acts primarily involving deception and dishonesty to more overt forms such as physical altercations (Frost et al., 2007; James & LeBreton, 2012). Examples of specific criteria that the CRT-A (or iterations/derivations thereof) has been shown to predict include: lying about experimental participation ($r = .49$, $p < .05$), infractions in intramural athletics ($r = .38$, $p < .05$), lying and cheating in an Internet-based simulation ($r = .40$, $p < .05$), and student conduct violations ($r = .55$, $p < .05$) (see James & LeBreton, 2012, Table 4.6 for a summary). Consistent with these findings, the present study examines the CRT-A's correlation with two forms of maladaptive behaviors (i.e., dishonesty and physical sports violations) among college students.

These two behaviors were combined for the purposes of the present study using a formative measurement model wherein measures are combined to form an index that is given meaning by the nature of its component parts (MacKenzie, Podsakoff, & Jarvis, 2005). This model differs from the more common latent construct model wherein the subordinate dimensions represent shared variance that emanates from the superordinate construct. Thus, the index used here is necessarily construct-deficient (i.e., we do not purport to capture all relevant manifestations of maladaptive behavior). As mentioned previously, however, past studies have found statistically significant correlations between CRT-A scores and behaviors that are directly relevant to both component parts. As such, we offer the following hypothesis:

Hypothesis 3: The CRT-A will demonstrate significant criterion-oriented validity with maladaptive behavior in the control condition but not in the disclose/fake good condition.

Study 3 Method

Participants

Participants were 71 undergraduate students (63.4% male, 36.6% female) at a midsized technical institution in the Southeastern United States, recruited from the school's psychology subject pool. Participants were either required to participate in experiments for their Introductory Psychology course (or do an alternate assignment) or were given the opportunity to participate for extra credit in one or more other psychology courses. In order to qualify for this study, participants were explicitly told that they needed to have completed at least one full season of any intramural sport at the institution in question. This qualification was necessary to create the criterion of interest (described in greater detail subsequently).

Materials

Conditional reasoning test for aggression. The same measure used in the previous two studies, including the same 11 CRT-A faking detection items, was used here. No participants endorsed five or more illogical response options, so all participants' data were retained for subsequent analyses.

Maladaptive behavior. Intramural sports infractions data were obtained from the institution's intramural sports office. Infractions include (but are not necessarily limited to) yellow or red cards in soccer, technical fouls in basketball, physical altercations, and/or ejections in any sport. Dishonesty was quantified by identifying participants who claimed to have completed at least one season of intramural sports but were not included in official institute records. It is important to point out that the requirement to have been officially enrolled in this institution's intramural sports program was specifically stated on all recruitment materials, at the beginning of each data collection session, and participants had to specifically state *which* intramural sport they had played. As such, it is highly unlikely that participants were simply not aware of this precondition.

Procedure

This study repeated the control and disclose/fake good conditions outlined previously—that is, roughly half the participants took the CRT-A under standard conditions whereas the other half were taught how this test works and were asked to fake good. This study's informed consent document, however, also stated that participants were (a) granting their consent for the researchers to obtain data about their intramural sports experiences from the campus recreation office and that this information would (b) include the specific sports they participated in, documented infractions, and the number of seasons played. After participation was complete, subjects were debriefed and informed that the true purpose of this study was to detect faking in personality testing and assess the relationship between personality and maladaptive behavior. Per Institutional Review Board requirements, after learning the purpose of the study, participants were given the opportunity to withdraw their original consent.

Study 3 Results

Manipulation Check

Consistent with the results of Studies 1 and 2, participants in the disclose/fake good condition had significantly lower CRT-A mean scores ($M = 1.29$, $SD = .96$) than those in the control condition ($M = 4.72$, $SD = 1.91$), $F(1, 69) = 91.31$, $p < .001$ (partial $\eta^2 = .57$). Similarly, participants in the disclose/fake good condition also endorsed significantly more honeypot response options ($M = 7.06$, $SD = 2.46$) than those in the control condition ($M = 1.39$, $SD = 1.32$), $F(1, 69) = 148.92$, $p < .001$ (partial $\eta^2 = .68$).

It is also important here to describe the base rates of the criterion utilized in this study. Consistent with past CRT research (e.g., James et al., 2005), the outcomes utilized here were quite rare in the present sample. Specifically, 11.2% of participants demonstrated some maladaptive behavior, with both dishonesty and intramural infractions shown by 5.6% of participants. No demographic variables showed significant relationships with our measure of maladaptive behavior or either of its component parts.

Hypothesis Tests

In support of Hypothesis 3, the correlation between scores on the CRT-A and maladaptive behavior was positive and significant under standard testing conditions, $r = .36$, $p < .05$ (two-tailed) but was

attenuated to nonsignificant levels in the disclose/fake good condition ($r = .12$, *ns*). These correlations were not significantly different, $Z = 1.03$, $p = .15$. To the extent that inside information about the CRT-A continues to leak, however, real-world data sets will likely contain a mix of respondents who respond honestly and respondents who attempt to fake good. In an effort to approximate these conditions, we recalculated the CRT-A's criterion-oriented validity after (a) combining the two conditions tested here but (b) eliminating from consideration those who were identified as likely fakers (i.e., those who endorsed six or more honeypot response options). While combining these two conditions yielded an $N = 71$, twenty-five total exclusions of likely fakers were made, yielding a final $N = 46$ for the condition used to simulate real-world conditions under which information about the CRT-A had leaked. Under these conditions, the correlation between CRT-A scores and maladaptive behavior remained close to its original value and maintained its statistical significance ($r = .32$, $p < .05$, $N = 46$). The correlations for the standard testing conditions and those under conditions intended to approximate those of the real world were not significantly different, $Z = 0.2$, $p = .42$.

Study 3 Discussion

The primary contribution of the present study is that it demonstrates the real-world effects of faking (and the detection thereof) on the CRT-A's criterion-oriented validity. Consistent with past conditional reasoning research (e.g., Frost et al., 2007; James et al., 2005), the CRT-A predicted maladaptive behavior among college students at better-than-chance levels under typical testing conditions, but its criterion-oriented validity was attenuated to nonsignificant levels after participants were taught how this test works and were encouraged to appear nonaggressive. Further, the present study also shows that the faking detection system developed here helps to maintain the validity of the CRT-A even when a large proportion of respondents know how this test works. These effects are important because previous research has shown that the effects of faking on criterion-oriented validity are oftentimes ambiguous or nonexistent (e.g., Hough et al., 1990; Schmitt & Oswald, 2006).

One potential reason why the present study showed a substantial faking-induced reduction in validity whereas some others have not is that most faking studies examine changes in relationships between essentially normally distributed predictors and criteria. Assuming that faking behavior is relatively consistent across participants, the end result will be similar to adding or subtracting a constant to participants' predictor scores, thereby having little to no effect on subsequent criterion-oriented validity. Even if we assume that those who fake most are those who have the most disadvantageous trait profile, changes in predictor scores among such a small proportion of individuals are unlikely to greatly attenuate validities given robust linear effects. The CRT-A, on the other hand, was developed to measure a low baseline phenomenon and predict low baseline outcomes (James & LeBreton, 2012). Given these considerations, it is not particularly surprising that validity would be attenuated if the small number of aggressive individuals who account for the preponderance of maladaptive behaviors are able to erroneously appear nonaggressive.

Study 4: Item Identification

The previous studies demonstrated the conditions under which faking on the CRT-A is possible, the potential for identifying those who have likely faked, and the effects of faking and faking detection on the CRT-A's criterion-oriented validity. But just as it is possible that respondents may become aware of the CRT-A's measurement intent, it is also possible that inside information about the present faking detection system could also eventually become public. If respondents are able to

successfully differentiate actual CRT items from CRT-based faking detection items, future efforts to identify respondents who are attempting to fake good on the CRT-A could be jeopardized. The present study begins the process of addressing this issue by examining the extent to which participants who are fully informed about self-reports and CRTs as well as faking detection efforts within both of these traditions are able to successfully identify each type of item. Given the relative paucity of research focused on identifying faking detection items, the following two research questions are examined here:

Research Question 3a: Can respondents identify faking detection items at better-than-chance levels?
Research Question 3b: Are respondents better at identifying self-report faking detection items or CRT-based faking detection items?

Study 4: Method

Participants

Participants were 55 advanced undergraduate students in a business management course at a mid-sized university in the Southern region of the United States. This sample was 58.3% female, with an average age of 24.5 years (range = 21-47), $SD = 4.9$.

Materials

The CRT-A (with the same 11 faking detection items embedded) and the PRF-E aggression and faking detection scales were again used in this study.

Procedure

Participants were first introduced to the concept of personality testing in employee selection. This introduction included examples of standard self-report items and how their transparency might lead to intentional response distortion in high-stakes testing situations such as personnel selection. Participants were then shown examples of items designed to catch potential fakers on self-report measures. Several examples of both standard self-report items and self-report faking detection items were then shown to participants for purposes of comparison, descriptions for each item type were provided, and participant questions were answered. Participants subsequently reviewed the 36-item CRT-A-F—which combined the full CRT-A measure with 11 CRT-A faking items—and a mixture of standard self-report and self-report faking detection items, with the goal of correctly identifying (but not responding to) each item by type. Specifically, CRT-A based items could be identified as *personality*, *faking detection*, or *reasoning* items, while self-report items could be identified as *personality* or *faking detection* items.

One week later, the same group of participants was introduced to conditional reasoning tests using the same approach utilized in Study 1, Condition 3 (disclose/fake good), including general information about the purpose and structure of CRT-A items as well as examples thereof. Participants were then taught about the CRT-A faking detection system. Specifically, they were told that this approach uses items that are designed to look like standard CRT-A items but contain only one realistic answer, which also happens to appear somewhat aggressive. The prosocial sounding option, on the other hand, was intentionally designed to appear unrealistic to participants who are naïve to the nature and structure of CRTs so that only those who know how the CRT-A works and are attempting to fake good are likely to endorse the unrealistic (but prosocial sounding) response option. The experimenter then showed participants several examples of mock CRT faking detection

Table 5. Study 3 Descriptive Statistics.

	Control						Disclose/Fake Good					
	N	*M*[a]	*SD*	Skew	Kurt.	α	*N*	*M*[a]	*SD*	Skew	Kurt.	α
CRT-A	36	4.72	1.91	0.19	–0.05	0.39	35	1.29	0.96	0.44	–0.6	0.74
CRT-F	36	1.29	1.31	0.81	0.112	0.79	35	7.02	2.43	–0.62	–0.72	0.83

Note: CRT-A = Conditional Reasoning Test of Aggression; CRT-F = Conditional Reasoning Test faking detection scale.
[a]No statistically significant differences across demographic groups according to age, gender, or ethnicity detected.

Table 6. Study 4 Descriptive Statistics

	N	*M*	*SD*	Observed % Correct	Expected % Correct[a]
PRF-Agg	50	7.4	2.56	46.25	50.00
PRF-SDR	50	6.72	2.57	42	50.00
CRT-A	54	9.69	3.86	44.02	33.33
CRT-F	54	2.17	1.71	19.69	33.33

Note: CRT-A = Conditional Reasoning Test of Aggression; CRT-F = Conditional Reasoning Test faking detection scale; PRF-SDR = Personality Research Form Socially Desirable Responding scale; PRF-Agg = Personality Research Form Aggression scale.
[a]Based on chance alone.

items and explained how each response option was intended to work. In general, every effort was made to provide participants with as much inside information as possible. Participants then read the 36-item CRT-A-F and were asked to identify, as opposed to actually respond to, the three item types on this version of the CRT-A: (a) actual CRT-A items, (b) CRT-A faking detection items, and (c) actual inductive reasoning items.

It is important to note that no attempt was made to counterbalance exposure to self-reports versus CRTs in an effort to make the present test as conservative as possible. That is, we wanted participants to use any information they learned about faking detection during the first session to *benefit* them as they attempted to identify CRT-based faking detection items in the second session.

Study 4: Results

The results of Research Question 3a show that participants identified both types of items at worse-than-chance levels. Specifically, self-report faking detection items were correctly identified in 42.00% of cases, $t(49) = -4.89$, $p < .001$, and conditional reasoning-based faking detection items were correctly identified in 19.69% of cases, $t(53) = -7.86$, $p < .001$. The former test was calculated by comparing the mean number of self-report faking detection items correctly identified by respondents (i.e., 6.73) to the number that would be expected if respondents selected response options at random (i.e., 8.50).[1] The latter test was calculated by comparing the mean number of CRT-A faking detection items correctly identified as faking items by respondents (i.e., 2.17) to the number that would be expected if respondents categorized the 11 faking detection items at random into the three categories in question (i.e., 4.00) (see Table 5 for additional statistics).[2]

Research Question 3b focused on overall identification rates for each measurement approach. Comparing mean differences, CRT-based faking detection items were identified at significantly lower levels (i.e., 19.69%) than were the self-report faking detection items (42.00%), $F(1, 102) = 51.65$, $p < .001$. See Table 6 for summary statistics.

Study 4: Discussion

The present study suggests that respondents were unable to identify faking detection items at better-than-chance levels (across measurement approaches) even after learning specific details about how both faking detection systems work. In addition to this promising general trend, it is also important to note that respondents were even less likely to correctly identify CRT-based faking detection items than they were to identify self-report faking detection items. This finding suggests that even respondents who are fully informed about the nature and structure of CRTs and the CRT-A based faking detection system developed here are unlikely to be able to simultaneously fake good and evade faking detection efforts, which is especially important given contemporary concerns over the security of standardized testing strategies/content. Further, it is important to note here that the present test is a particularly conservative one in that (a) information learned about faking in session one may have benefited participants in session two and (b) no effort was made to screen CRT-A faking detection items for quality, meaning that some of the poorer performing items may have been especially easy to detect.

General Discussion

Minimizing faking is critical to improving the quality of modern personality assessment (Murphy & Dzieweczynski, 2005; Reynolds, 2010). The present article combined two approaches to combat faking: using a difficult-to-fake test (i.e., the Conditional Reasoning Test for Aggression) and including items intentionally designed to identify individuals who have learned to fake this test. Results suggest this multipronged approach holds much promise in that (a) the CRT-A was only susceptible to faking after participants were taught how to beat it, (b) the faking detection system developed here generally performed as intended (including positive effects on criterion-oriented validity), and (c) participants were not able to identify CRT-based faking detection items at greater-than-chance levels even after learning how this faking detection system works.

Even though faking on the CRT-A is presently only a hypothetical problem, it appears to be one with a readymade solution. Specifically, it simply requires researchers to develop items that (a) look like traditional CRT items, (b) contain one clearly logical response option that seems on its surface to be somewhat aggressive (e.g., features words pertaining to violence, crime, fighting, death, dishonesty), and (c) contain a clearly illogical response option that portrays a nonaggressive perspective in a way that is unrealistically naïve, to the point that not even an extreme pacifistic would select it unless he or she was trying to intentionally appear unaggressive. Unlike traditional CRT items, then, creating these items does not require any formal comparison of underlying logic because only the aggressive sounding option is actually logical—indeed, the honeypot response option is intentionally illogical and unreasonable but in a way that appears extremely starry-eyed. Instead, testing these items can occur in a purely empirical manner by comparing base rates under varying conditions (i.e., standard vs. disclose/fake good), with ideal items having zero honeypot options endorsed under standard conditions and 100% endorsement when participants are attempting to fake good.

Further, unlike some faking detection scales wherein high scores are also common among those who are particularly virtuous, honeypot response options were almost completely inert unless participants were provided with inside knowledge about how the CRT-A operates and were instructed to fake good. Specifically, only 3 of 190 participants (1.5%) endorsed six or more honeypot response options under naïve testing conditions, the ability and motivation to fake CRTs explained an average of 67.25% of variance on participants' faking detection scores, and our faking detection system showed no significant correlations with deep or surface-level characteristics under standard testing conditions.

Perhaps the most valuable results were obtained by examining the effects of faking and faking detection on criterion-oriented validity. Specifically, the present article demonstrates that the CRT-A predicts antisocial outcomes under normal administration conditions, but this validity drops to nonsignificant levels under those conditions wherein participants have the ability and motivation to fake. That being said, when participants with suspicious faking detection scores are eliminated from the data set, observed validities return to significant levels. This line of evidence is critical because organizational scientists are ultimately interested in making inferences that will increase the probability of positive outcomes and decrease the probability of negative outcomes in the workplace—a goal that is theoretically furthered by the present findings.

Lastly, it is also important to note that participants who were fully informed about faking detection efforts within both the self-report and conditional reasoning traditions were unable to identify faking detection items at better-than-chance levels. Moreover, these participants were also significantly worse at identifying CRT-based faking detection items than self-report faking detection items. Although this finding should not necessarily be taken to mean that all of the effects of the faking detection system outlined previously will necessarily remain unchanged should participants learn about CRT-based faking detection efforts, it is a promising first step nonetheless.

Critically, all of these conclusions are based on conservative tests in which all 11 faking detection items were utilized at every step (i.e., at no point were any items edited or dropped). Thus, the results reported throughout this article likely represent low-end estimates because some of the items utilized here performed better than others. As such, the present investigation has important implications for continued faking and faking detection research but also possesses some limitations that can and should be improved on.

Limitations and Future Research

The most obvious limitation of the present study is its reliance on students. Three important points, however, should be noted here. First, the study is the only known attempt to develop and test a method of CRT-based faking detection, so student samples help to provide a critical feasibility test. Such a test is particularly relevant because learning how to beat the CRT-A is ultimately based on cognitive processing and not contingent on having relevant work experience, thereby suggesting that students can provide meaningful data for this purpose (Hinkin, 1998). Second, the use of more mature and experienced samples in Studies 2 and 4 helps to assuage some of the concerns associated with relying on traditional undergraduates. Third, extant evidence suggests the demographic nature of the sample(s) used in CRT-based research does not appear to substantively affect reported scores (LeBreton et al., 2007), so the use of students here should be viewed as a rather minor limitation. That being said, it is possible that some aspects of actual testing environments (e.g., increased cognitive load, high pressure) may influence observed response patterns. Thus, this line of research should be extended to real-world assessment contexts prior to fully implementing the faking detection system developed herein.

Regarding future research, the overarching theme of the potential additional studies outlined here is that researchers should better understand the psychological processes respondents use when responding to CRT-based items, both in standard testing conditions (i.e., wherein their indirect nature has been maintained) and under other conditions wherein respondents have varying levels of information about the nature, structure, and intent of these scales (i.e., wherein their indirect nature has been compromised in various subtle ways). Such research might include (but is not necessarily limited to) studies that utilize techniques wherein participants think aloud while completing assessments to provide insights into why they selected (or did not select) each specific response option (Goffin & Boyd, 2009; Messick, 1995) and/or research examining response-time latencies to different item types under various conditions.

This line of research would also allow scholars to better determine the extent to which providing participants with inside information about CRTs bring response processes into conscious awareness, thereby making subsequent responses more similar to the ways in which participants respond to traditional (i.e., explicit) personality scales. This is a critical issue because the fact that the CRT-A was susceptible to faking could "call into question the very ability of this measure to assess cognitions that are purportedly unconscious" (LeBreton et al., 2007, p. 6). It is argued here, however, that the present findings should not be interpreted so harshly. Namely, the faking detection system developed here is predicated on the notion that providing participants with inside information about the CRT-A's nature and structure moves response processes out of the domain of the unconscious and into the domain of the conscious, thereby suggesting that respondents across the conditions tested here respond using qualitatively different cognitive processes for qualitatively different reasons. Although the positive results of the present study provide initial support for this explanation, more direct tests of this possibility merit continued research as it speaks to an important issue about the specific constructs CRTs assess and how they go about doing so.

A second stream of research (one that is not necessarily mutually exclusive with the first) involves refining the faking detection system developed here and examining faking and faking detection within the context of CRT-based measures of other constructs (e.g., Achievement Motivation; James, 1998). Regarding the former, additional refinements of the items tested here would likely focus primarily on validating a final, optimally parsimonious faking detection scale, including setting a specific cutoff value to determine when faking has likely occurred (the cutoff used in Study 2 would obviously not apply to a shorter scale and should be viewed here as study-specific, as opposed to a generally applicable rule).

Regarding the latter, given that the faking detection items developed here were designed to appear to assess aggressiveness, it is important to explicitly state that researchers should not utilize these items for other CRTs. Specifically, the nature of the faking detection system developed here is such that items are most likely to work best to the extent that their content does not noticeably differ from that of the other items within the scale. Instead, new faking detection items will need to be developed to reflect the content of the scale(s) in question. Faking detection items developed for other CRT scales *can* however use the same structure as those used here—namely, one logical but highly implausible option that appears to reflect the more socially desirable end of the construct and one logical but more plausible option that appears to reflect the socially undesirable end of the construct. Thus, the main output of this stream of research would be a series of parsimonious CRT-based faking detection scales (one for each CRT instrument), each with its own generalizable faking detection cut-score.

We believe this research also has important implications for domains beyond CRT-focused research. First, the technique of including honeypot items for the purposes of detecting faking could be of use in other indirect measures (e.g., Implicit Association Test; Greenwald & Banaji, 1995), provided respondents are unable to identify such items. This technique may also be applicable in other domains of personnel selection, such as selection interviews. For instance, selection interviewers may better detect socially desirable interview responding (e.g., Levashina & Campion, 2007) by posing questions aimed at eliciting prosocial responses when in fact the more logical responses may sound less socially desirable. For example, a hiring manager for a job that requires high levels of autonomy (e.g., computer programmer, park ranger) may ask about their willingness and ability to collaborate with others when in reality one's ability to work independently is a more important skill.

Second, and more focused on the issue of socially desirable responding, the topic of item identification merits additional research from organizational scientists and practitioners. More specifically, it appears that test developers work under the assumption that respondents cannot or will not identify faking detection items. While the present results suggest respondents cannot identify

CRT faking items with better-than-chance accuracy, respondents were comparatively better at identifying faking detection items within self-report measures. This suggests the aforementioned assumption—namely, that respondents cannot successfully determine faking detection items on self-report measures—may be untenable. Additionally, given the contentious debate around faking on personality measures in organizational research, results from Study 3 warrant continued examination of the extent to which faking adversely affects criterion-oriented validity of personality measures used in personnel selection, including conditional reasoning measures of personality. Finally, future researchers should examine the extent to which the CRT faking scale functions as a suppressor (e.g., Bing et al., 2011; Ones, Viswesvaran, & Reiss, 1996).

Conclusion

As long as respondents are motivated and able to fake, doing so will likely continue to be a substantial area of concern for researchers and practitioners alike. Thus, efforts to prevent faking and develop the tools necessary to identify those who have faked in a particular setting represent important validity safeguards. Although indirect measurement systems like CRTs represent difficult-to-fake options, an accumulating body of evidence suggests that respondents can manipulate their scores on CRTs when provided with detailed inside knowledge about how these tests work. The faking detection system developed here, however, identifies fakers and nonfakers alike with promising levels of accuracy, thereby helping to ensure the effectiveness of conditional reasoning tests as their popularity grows and their "secret" continues to get out. Although it would be ideal to maintain the indirect nature of CRTs in perpetuity, the high-stakes nature of employment testing suggests that this secret is likely to continue to leak to the test-taking public. As such, the present research is important because it illustrates what could happen if respondents attempt to use inside information about the CRT-A to artificially deflate their measured aggressiveness scores while simultaneously providing a solution to this potential problem.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. This comparison value was calculated by multiplying the chance of correctly categorizing a given self-report item (i.e., 50%, given that there were two response options for self-report scales) by 17 (i.e., one plus the total number of items, which accounted for the fact that participants could have identified zero items correctly).
2. This comparison value was calculated by multiplying the chance of correctly categorizing a given Conditional Reasoning Test (CRT)-based item (i.e., 33%, given that there were three response options for CRT-based scales) by 12 (i.e., one plus the total number of items, which again accounted for the fact that participants could have identified zero items correctly).

References

Alliger, G. M., Lilienfeld, S. O., & Mitchell, K. E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science*, *7*, 32-39.

Bagby, R. M., Gillis, J. R., Toner, B. B., & Goldberg, J. (1991). Detecting fake-good and fake-bad responding on the Millon Clinical Multiaxial Inventory-II. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*, 496-498.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1-25.

Bing, M. N., Kluemper, D., Davison, H. K., Taylor, S., & Novicevic, M. (2011). Overclaiming as a measure of faking. *Organizational Behavior & Human Decision Processes*, *116*, 148-162.

Bing, M. N., LeBreton, J. M., Davison, H. K., Migetz, D. Z., & James, L. R. (2007). Integrating implicit and explicit social cognitions for enhanced personality assessment. *Organizational Research Methods*, *10*, 346-389.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Smith, M. A., & Brannick, M. T. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*, 317-335.

Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face validity and fakeability of objective and projective measures of dependency. *Journal of Personality Assessment*, *63*, 363-386.

Bowler, J. L., Bowler, M. C., & Cope, J. G. (2013). Measurement issues associated with conditional reasoning: An examination of faking. *Personality and Individual Differences*, *55*, 459-464.

Chamorro-Premuzic, T. (2015). *Ace the assessment*. Retrieved from https://hbr.org/2015/07/ace-the-assessment

Chamorro-Premuzic, T., & Furnham, A (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality*, *37*, 319-338.

Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology*, *47*, 847-860.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of prevalence, severity, and verifiability of entry-level faking using the randomized response technique. *Human Performance*, *16*, 81-106.

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996). *The validity of non-cognitive measures decays when applicants fake*. Cincinnati, OH: Academy of Management Proceedings.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology*, *84*, 155-166.

Frost, B. C., Ko, C. H. E., & James, L. R. (2007). Implicit and explicit personality: A test of a channeling hypothesis for aggressive behavior. *Journal of Applied Psychology*, *92*, 1299-1319.

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. *Canadian Psychology*, *50*, 151-160.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4-27.

Griffith, R. L., & Peterson, M. H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 308-311.

Harvill, L. M. (1991). An NCME instructional module on standard error of measurement. *Instructional Topics in Educational Measurement*, *10*(2), 33-41.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*, 104-121.

Hogan, R. (2005). In defense of personality measurement: New win for old whiners. *Human Performance*, *18*, 331-341.

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance*, *11*, 209-244.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effects of response distortion on those validities. *Journal of Applied Psychology*, *75*, 581-595.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology*, *85*, 869-879.

Jackson, D. N. (1974). *Personality research form manual* (2nd ed.). Port Huron, MI: Research Psychologists Press.

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods*, *1*, 131-163.

James, L. R., & LeBreton, J. M. (2012). *Assessing the implicit personality through conditional reasoning*. Washington, DC: American Psychological Association.

James, L. R., & McIntyre, M. D. (2000). *Conditional Reasoning Test of Aggression test manual*. Knoxville, TN: Innovative Assessment Technology.

James, L., R., McIntyre, M. D., Glisson, C. A., Bowler, J., & Mitchell, T. R. (2004). The conditional reasoning measurement system for aggression: An overview. *Human Performance*, *17*, 271-295.

James, L. R., McIntyre, M. D., Glisson, C. A., Green, P. D., Patton, T. W., LeBreton, J. M., . . . Williams, L. J. (2005). A conditional reasoning measure for aggression. *Organizational Research Methods*, *8*, 69-99.

James, L. R., & Rentsch, J. R. (2004). J-U-S-T-F-Y to explain the reasons why: A conditional reasoning approach to understanding motivated behavior. In B. Schneider & D. B. Smith (Eds.), *Personality and organizations* (pp. 223-250). Mahwah, NJ: Lawrence Erlbaum Associates.

Kihlstrom, J. F. (1999). The psychological unconscious. In L. Pervin & O. P. Johs (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 424-442). New York, NY: Guilford.

Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment*, *15*, 220-231.

LeBreton, J. M., Barksdale, C. D., Robin, J., & James, L. R. (2007). Measurement issues associated with Conditional Reasoning Tests: Indirect measurement and test faking. *Journal of Applied Psychology*, *92*, 1-16.

Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: Development and evaluation of an interview faking behavior scale. *Journal of Applied Psychology*, *92*, 1638-1656.

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology*, *90*, 710-730.

Madaus, G. F., Russell, M., K., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age Publishing Inc.

Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences*, *32*, 247-256.

McCrae, R. R., & Costa, P. T., Jr. (1983). Social desirability: More substance than style. *Journal of Consulting and Clinical Psychology*, *51*, 882-888.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, *85*, 812-821.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*, 450-470.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology*, *60*, 1029-1049.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*(3), 683-729.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology*, *88*, 348-355.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2006). Individual differences in impression management: An exploration of the psychological processes underlying faking. *Psychology Science*, *48*, 288-312.

Murphy, K. R., & Dzieweczynski, J. L. (2005). Why don't measures of broad dimensions of personality perform better as predictors of job performance? *Human Performance*, *18*, 343-357.

Nicholson, R. A., Mouton, G. J., Bagby, R. M., Buis, T., Peterson, S. A., & Buigas, R. A. (1997). Utility of MMPI-2 indicators of response distortion: Receiver operating characteristic analysis. *Psychological Assessment*, *9*, 471-479.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, *11*, 245-271.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*, 660-679.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*, 598-609.

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, *65*, 70-89.

Reynolds, C. R. (2010). Measurement and assessment: An editorial view. *Psychological Assessment*, *22*, 1-4.

Rossé, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, *83*, 634-644.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection. What does the current research support? *Human Resource Management Review*, *16*, 155-180.

Salgado, J. F. (2002). The Big Five personality dimensions and counterproductive work behavior. *International Journal of Selection and Assessment*, *10*, 117-125.

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, *78*, 966-974.

Schmitt, N., & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitve measures in selection settings. *Journal of Applied Psychology*, *91*, 613-621.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93-105.

Smith, D. B., & Ellington, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology*, *87*, 211-219.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review*, *9*, 219-242.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645-665.

Tyre, P. (2016, March 21). How sophisticated test scams from China are making their way into the U.S. *The Atlantic*. Retrieved from http://www.theatlantic.com/education/archive/2016/03/how-sophisticated-test-scams-from-china-are-making-their-way-into-the-us/474474/

Viswesvaran, C., & Ones, D. S. (1999). Meta-analysis of fakeability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*, 197-210.

Wiita, N. E., Schnure, K., & James, L. R. (2010, April). *A comparison of law enforcement applicant scores on the MMPI-2, PRF-E, and CRT-A*. Poster presented at the Annual Meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Winter, D. G., John, O. P., Stewart, A. J., Klohnen, E. C., & Duncan, L. E. (1998). Traits and motives: Toward an integration of two traditions in personality research. *Psychological Review*, *105*, 230-250.

Zickar, M. J., & Robie, C. (1999). Modeling faking on personality items. *Journal of Applied Psychology*, *84*, 551-563.

Ziegler, M., MacCann, C., & Roberts, R. (2011). *New perspective on faking in personality assessment*. New York, NY: Oxford University Press.

Author Biographies

Nathan Wiita is a principal and research and innovation lead at RHR International, LLP.

Elnora Kelly is a doctoral student at Georgia Institute of Technology's School of Psychology.

Rustin Meyer is an assistant professor at the Georgia Institute of Technology's School of Psychology.

Brian Collins is an assistant professor at the University of Southern Mississippi's Department of Management and International Business.