Journal of Business and Psychology
https://doi.org/10.1007/s10869-020-09707-2

ORIGINAL PAPER

Insufficient Effort Responding as a Potential Confound between Survey Measures and Objective Tests

Jason L. Huang[1] · Justin A. DeSimone[2]

Abstract

Following research that demonstrates insufficient effort responding (IER) may confound survey measures and inflate observed correlations (Huang, Liu, & Bowling, 2015c), a question emerges as to *whether* and *when* IER can act as a confound between objective tests and surveys. Using data ($N = 243$) originally designed to examine training and transfer, study 1 demonstrates that (a) IER is negatively related to performance on tests, and (b) IER's influence on surveys depends on the sample means of these measures. As a result, IER could inflate a test's association with other tests and surveys. Study 2 investigates the impact of two parameters—within-person consistency of IER and percentage of IER cases in the sample—by randomly replacing bootstrapped attentive responses (10,000 bootstrapped samples of 200 cases identified from study 1). When predicting the confounding effects of IER, within-person consistency has positive linear and quadratic effects, percentage of IER cases has a positive linear effect, and consistency and percentage have a positive interactive effect. Research and practical implications for the design and evaluation of surveys and tests are discussed.

Keywords Insufficient effort responding · Careless responding · Random responding · Response effort · Measurement

Research in management and organizational psychology often relies on respondents to provide data, be it reports of internal states, perceptions, and experiences in surveys or performance on objective tests and tasks. When utilizing surveys, researchers have emphasized the need to screen for insufficient effort responding (IER; see DeSimone, Harms & DeSimone, 2015; Johnson, 2005; Kam & Meyer, 2015; Meade & Craig, 2012; Maniaci & Rogge, 2014), which occurs when respondents provide careless or random responses to survey items due to low motivation to comply with survey instructions (Huang, Curran, Keeney, Poposki, & DeShon, 2012). This increased awareness has led researchers to propose methods to indicate whether a participant engaged in IER on a survey (e.g., Curran, 2016; Huang, Bowling, Liu, & Li, 2015b; Oppenheimer, Meyvis, & Davidenko, 2009; Wood, Harms, Lowman, & DeSimone, 2017).

Researchers long believed that IER has an attenuating effect on observed associations (McGrath, Mitchell, Kim, & Hough, 2010). However, recent research identified conditions under which IER can inflate correlations between survey measures, confounding estimates of association. Specifically, when attentive respondents score very high or very low on a substantive measure, the tendency for IER scores to congregate around the midpoint of the response scale will result in a nonzero correlation between IER behavior and observed scores on the substantive measure (Huang, Liu, & Bowling, 2015c). Regardless of whether IER attenuates or inflates correlations between survey measures, the potential for biased findings makes it important to detect and remove IER cases prior to data analysis (Huang, Liu, & Bowling, 2015c; McGonagle, Huang, & Walsh, 2016).

Despite the recent surge of research on IER, two gaps loom large in the current literature. First, the investigation of IER has primarily focused on survey data, leaving unexamined other measurement approaches where response effort can play a critical role. Unlike survey measures that typically contain multiple response options corresponding to different levels of

✉ Jason L. Huang
huangjl@msu.edu

Justin A. DeSimone
jadesimone@cba.ua.edu

[1] School of Human Resources and Labor Relations, Michigan State University, 368 Farm Lane, 437 SKH, East Lansing, MI 48824, USA

[2] Culverhouse College of Business, University of Alabama, Tuscaloosa, AL, USA

Published online: 22 August 2020

the focal trait (Guttman, 1944), objective tests typically distinguish between correct and incorrect responses (see Longstaff & Porter, 1928) that are modeled dichotomously (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). As performance on objective tests requires response effort (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011), IER on objective tests and tasks should result in lower performance than attentive responding. Thus, to the extent that IER occurs, scores on objective tests can become confounded with IER. More importantly, given the potential confounding role of IER in survey measures identified in Huang, Liu, and Bowling (2015c), the presence of IER can confound the measurement of constructs obtained using surveys *and* tests, two distinct methods that appear less likely to share a common methodological confound than the reliance on objective tests or survey measures alone (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). This confounding effect, if demonstrated, will help researchers identify conditions under which inflated associations are likely to occur due to the presence of IER, thus preventing unsuspecting researchers from drawing erroneous conclusions.

Second, although simulation studies have been instrumental in demonstrating IER's potential confounding influence in survey measures (Credé, 2010; DeSimone, DeSimone, Harms, & Wood, 2018; Huang, Liu, & Bowling, 2015c), two IER parameters have yet to receive close examination in the literature: (a) within-person consistency of IER and (b) percentage of IER in a sample. First, to gauge the potential influence of IER, existing simulation studies have typically relied on a simple dichotomy of response effort, generating data that are either consistently attentive or consistently inattentive. However, instead of constantly producing random data, actual respondents' IER behavior can vary in the degree of *within-person consistency of IER behavior*, which is reflected in the percentage of a participant's responses in which they do not respond effortfully or attentively. There is no reason to believe a given participant exerts consistent effort across an entire survey, so IER may range from an occasional streak of careless responses to a deliberate attempt at generating random answers.[1] Attending to within-person consistency will cover the underlying continuum of IER behavior (Huang et al., 2012) and thus more accurately depict the potential impact of IER. Second, *the percentage of IER in a sample* may not only influence the impact of IER but also interact with within-person consistency. Together, examining the impact of within-person consistency of IER behavior in conjunction with the percentage of IER in a sample can help researchers identify the scenarios under which IER is most damaging to data quality and thus direct their efforts accordingly to minimize the impact of IER on study results.

Addressing these two gaps can result in insights for practice and research in management and organizational psychology. As an example, in a selection setting, validation of selection tools is often conducted using current employees, some of whom may not be as motivated to provide attentive responses as applicants (Arvey, Strickland, Drauden, & Martin, 1990). If IER behavior in a concurrent validation sample can result in an artificially inflated correlation between a survey measure (e.g., openness to experience) and an objective test (e.g., general cognitive ability), an unsuspecting analyst may decide against using one or both measures in the selection battery due to the strong empirical overlap.

Additionally, research on transfer of training has indicated that the association between a predictor and transfer can be susceptible to inflation when both variables are assessed at the same time using the same method (Blume, Ford, Baldwin, & Huang, 2010). However, researchers are less likely to be concerned about potential inflation when different measurement methods are used (e.g., between self-reported openness to experience and objective transfer test scores; Podsakoff et al., 2003). Understanding the joint impact of within-person consistency of IER behavior and the percentage of IER in a given sample can enable scholars to implement methods to detect and potentially deter IER behavior, thus reducing the potential for type I errors in identifying key predictors of transfer.

The goal of the current paper is to address these two gaps. After introducing the rationale behind the confounding effect of IER between objective tests and surveys, a laboratory study (study 1) on training and transfer that included both objective tests and survey measures is used to demonstrate the extent to which IER confounds observed scores in a realistic research setting. A follow-up simulation (study 2) serves as the basis to examine the joint influence of within-person consistency and percentage of IER on IER's confounding effect.

The confounding effect of IER

Until recently, researchers generally focused on IER in terms of its psychometric impact on survey results. When perceived as a source of random measurement error (Huang et al., 2012), IER behavior is expected to attenuate the expected relationship between two variables, such as a predictor and a criterion (McGrath et al., 2010). However, Huang, Liu, and Bowling (2015c) identified conditions under which IER can introduce a systematic source of variance in survey data, thereby inflating observed relationships. Specifically, Huang, Liu, and Bowling (2015c) noted that IER behavior as a whole, in the absence of other response sets (e.g., socially desirable responding), will

[1] We focus on within-person consistency as the percentage of items in a given data collection that a participant responds to with insufficient effort. This narrower focus is distinct from rank-order stability, which is reflected in a correlation coefficient between IER measures obtained from two different survey administrations (see Bowling et al., 2016). Rank-order stability captures the consistency of individuals' relative standings on their IER behavior across time and situations.

resemble a random variable that has a mean at the midpoint of the response scale. Huang, Liu, et al. (2015, p. 830) described three mechanisms that cause IER scores to average near a scale midpoint. First, truly random IER data are uniformly distributed with a mean on the midpoint of the response scale. Second, when respondents provide patterned IER responses, such as a long string of the same response options, the average score *across different respondents* tends to be around the scale midpoint when patterned response selection is uniformly distributed. Third, when attentive responses average away from the midpoint, there is a higher probability for occasional inattentive responses to average near the scale midpoint than otherwise.[2] For instance, if attentive respondents on average score 4 (agree) on a five-point Likert scale (1 = strongly disagree; 5 = strongly agree), there are more response categories (1, 2, and 3) for an occasional error to land below the attentive scores as opposed to above (5). Likewise, if attentive respondents average lower than the scale midpoint (e.g., 2 out of 5), there are more response categories for the occasional errors to occur higher than this expected score of 2. Thus, the expected score of 4 (or 2) is likely to be pulled lower (or higher) toward the scale midpoint in the presence of error.

Given the expectation that IER scores tend to average near the scale midpoint, when attentive respondents score lower (or higher) than the midpoint on a substantive variable, expected IER scores will be higher (or lower) than the attentive scores. The mean difference between attentive and IER scores will result in a positive (or negative) correlation between IER behavior and the observed scores. Following the mechanism described above, Huang, Liu et al. (2015) demonstrated that (a) students identified to have engaged in IER behavior tended to have means near the midpoint on a number of personality scales; (b) the addition of 10% random responses introduced spurious correlations among otherwise uncorrelated variables in a simulated dataset; and (c) partialing out IER in an employee survey reduced the magnitude of correlations between substantive variables having high or low means. Further, McGonagle et al. (2016) replicated Huang, Liu, and Bowling (2015c) findings with two employee surveys focusing on work and occupational health.

Although the revelation that IER can confound estimates of relationships between study variables is certainly important, it is limited due to the exclusive focus on survey data. In fact, the majority of IER research to date has focused exclusively on the effects of IER on survey measures, neglecting other types of tests widely used in management and organizational psychology. However, IER can also affect scores on objective tests. For instance, Wise and colleagues (Wise, 2006; Wise & DeMars, 2006; Wise & Kong, 2005) examined IER in low-stakes educational testing context, where some students do not put in sufficient effort to respond to achievement tests. When respondents engage in IER on objective test items, they are unlikely to correctly answer these items, so their expected scores will be lower than those respondents who engage in attentive responding. As a result, the association between higher IER scores and lower objective test scores should translate into a negative correlation between IER and test scores. Thus:

Hypothesis 1: *IER will be negatively correlated with scores on objective tests.*

As IER behavior stems from respondents' personality traits and exhibits rank-order stability across surveys (Bowling et al., 2016; DeSimone, Davison, Schoen, & Bing, 2020), individuals who engage in higher levels of IER on one test or survey may tend to engage in higher levels of IER on other tests or surveys. For instance, Bowling et al. (2016, study 1) measured 166 employees' IER on two identical surveys administered 13 months apart and found high temporal consistency, with $r = .67$ between two standardized IER measures. Thus, respondents exhibiting IER behavior will be expected to score lower across different tests than their attentive counterparts. As a result, the presence of somewhat stable IER behavior across different tests will serve as a common confound for scores on different objective tests, thus artificially inflating their associations. At first blush, the notion that IER can confound objective tests seems incompatible with classical test theory, which states that *random errors* simply add noise to measurements, reduce reliability, and thus attenuate the association between two tests. However, "measurement error can be in the form of either a systematic bias or random errors" (Nunnally, 1978, p. 190). In the present context, these errors are systematic instead of random because they are caused by IER. Put another way, if we (overly) simplify the underlying IER continuum into a dichotomized grouping variable (IER = 1 or 0), two different tests will share systematic variance due to respondents' group membership, and partialing out this systematic variance will reveal the unbiased association between the tests. Thus,

Hypothesis 2: *Partialing out IER will reduce the magnitude of correlation between objective tests.*

The potential confounding effect of IER should not be limited to bivariate relations obtained with the same measurement format, be it between two survey measures (Huang, Liu, & Bowling, 2015c) or between two objective tests (Hypothesis 2 above). Indeed, IER can act as a confound between a survey measure and a test. Unlike on an objective test, where IER is consistently expected to result in lower observed scores, IER's confounding effect on a survey measure is contingent on the attentive respondents' average score. If a survey measure's

[2] We should note that when attentive responses have an average near the scale midpoint, positive and negative errors tend to cancel each other out, resulting in IER scores near the scale midpoint as well.

average response is higher or lower than its scale midpoint, which frequently occurs in survey research (Huang, Liu, & Bowling, 2015c), IER may affect the relationship of that survey with other measures. Consider the core self-evaluations (CSE; Judge, Erez, Bono, & Thoresen, 2003) measure as an example. The average score of attentive respondents tends to be higher than the scale midpoint and hence higher than inattentive respondents as a whole (Huang, Liu et al., 2015), so IER behavior should have a negative correlation with observed CSE scores. This negative correlation, when coupled with a negative correlation between IER and objective test scores (see Hypothesis 1 above), renders IER a common confound between a survey measure and an objective test.

The confounding influence of IER on a survey measure should depend on where attentive respondents score as a whole relative to the scale midpoint. Whether attentive respondents, on average, score higher or lower than the midpoint of a scale is unknown before survey administration, though it can often be predicted from characteristics such as item wording/direction and social desirability or estimated using response distributions from previous studies. When average attentive scores are higher than the scale midpoint, as in the aforementioned CSE measure, we expect IER to have a negative correlation with scale scores. However, if average attentive scores are lower than the scale midpoint, IER should be positively correlated with scale scores. As a result, the direction of association between IER and a survey measure, which influences the association between the survey measure and objective measures, can plausibly be either negative or positive. In the presence of IER, such associations should typically be nonzero, but the expected direction depends on scale characteristics (see Huang, Liu, & Bowling, 2015c). Therefore, in Hypothesis 3, we propose a general expectation of change without specifying the directionality of the change.

Hypothesis 3: *Partialing out IER will change the correlation between an objective test and a survey measure that has a mean away from its scale midpoint.*

Study 1: Laboratory study investigating the confounding effect of IER

A laboratory study originally designed to examine the role of CSE on learning and transfer served as the basis to test for Hypotheses 1–3. Figure 1 summarizes the research model. The central hypothesis focused on how CSE, through its effects on trainees' state mastery orientation and self-regulatory mechanisms, influences learning outcomes and transfer. It is worth noting that the research model resembles the model proposed and tested by Stanhope, Pond, and Surface (2013), who examined CSE as a distal predictor of learning in a field study of military personnel. The goal of the present study is not to replicate their work, but rather to use existing data to test the role of IER as a methodological confound. Also note that the possibility of investigating IER as a methodological confound only emerged after data collection, making this study a post hoc evaluation of the potential role of IER in an observational setting. Throughout the data collection, the research team paid close attention to collecting high-quality data.

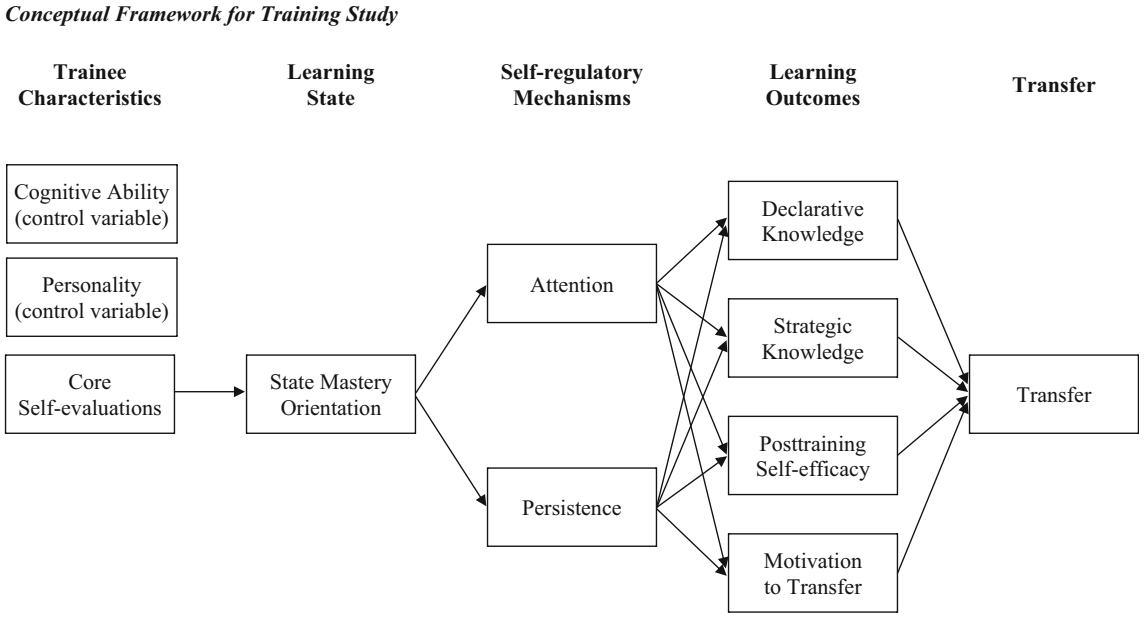

Fig. 1 Conceptual framework for training study

Study 1 method

Participants and procedure

Two hundred forty-three students enrolled in undergraduate psychology courses at a public university in the Midwestern United States completed the research study on computer-based learning. Participants had an average age of 21 years ($SD = 5$), and 65% were female.

Participants were instructed that the study consisted of a pretraining phase and a laboratory session involving computer-based learning and assessment. Participants were also informed that their participation was completely voluntary and they would receive extra course credit in exchange for their time. Consenting respondents were first directed to an online survey comprising a CSE scale, a personality inventory, and a verbal ability test (see "Measures" below). At the end of the online survey, participants received a link and a password to an online test of general cognitive ability. Upon the completion of the pretraining phase, they signed up for and subsequently attended a two-hour laboratory session.

Upon arrival in the laboratory, participants were instructed to complete a computer-based training program (Huang & Bramble, 2016), which presented rules and strategies for the game of Mahjong, a four-player tile game where each player attempts to win a game by absorbing new tiles and discarding existing tiles. The training program, delivered through MediaLab (Jarvis, 2008), progressed from declarative knowledge (e.g., names of tiles, names of specific tile combinations, general rules, how to win, etc.) to strategic knowledge (e.g., general progression of the game, how to improve one's tiles, various decisions, etc.). Learners were told to use as much time as they would need, and they were afforded control over several aspects of the learning environment (Kraiger & Jerden, 2007), including choosing the pace and sequence of materials and deciding whether to utilize optional quizzes, feedback, and review. Training outcomes were assessed upon completion of the training, followed by a transfer task where participants played the game against three computer players for forty-five minutes.

Measures

Pretraining measures The following substantive measures were assessed before participants started the laboratory training sessions: (a) CSE, (b) personality, (c) verbal ability, and (d) general cognitive ability.

Core self-evaluations CSE was assessed with the 12-item scale ($\alpha = .83$) by Judge et al. (2003). A sample item is "I am confident I get the success I deserve in life." The response scale ranged from 1 (*strongly disagree*) to 7 (*strongly agree*).

Five-factor model personality traits A personality inventory was obtained from the international personality item pool (IPIP; Goldberg, 1999), with 20-items per scale measuring *openness* ($\alpha = .85$), *conscientiousness* ($\alpha = .86$), *extraversion* ($\alpha = .88$), *agreeableness* ($\alpha = .87$), and *emotional stability* (the reverse of neuroticism; $\alpha = .87$). Participants indicated whether each item described their typical behavior with a 7-point scale (1 = *very inaccurate*; 7 = *very accurate*).

Verbal ability The 10-item vocabulary test from the General Social Survey (Cor, Haertel, Krosnick, & Malhotra, 2012) was adapted ($\alpha = .72$) to measure verbal ability (Caplan & Miller, 2010). For each of the 10 items, respondents were given a focal word in capital letters and were asked to choose, among five other words, one that closely matches the focal word in meaning. As an example, the focal word BEAST had the following five response options: (a) afraid; (b) words; (c) large; (d) animal; and (e) separate, with (d) being the correct response.

General cognitive ability General cognitive ability was assessed with Wonderlic Personnel Test–Quicktest (WPT-Q), an eight-minute online unproctored test consisting of 30 verbal, numeric, and logic questions. WPT-Q scores have been shown to correlate ($r = .86$) with scores from the full-length proctored Wonderlic Personnel Test in over 50,000 test takers (Callans, 2012). Due to technical problems, 10 participants had missing data on WPT-Q scores, which could not be linked to their responses on other parts of the study.

Training process measures Three training-related processes were measured: (a) state mastery goal orientation; (b) attention; and (c) persistence. Training process measures were adapted for the context of the present learning task (i.e., learning Majong). The assessments of mastery-oriented learning state and self-regulatory mechanisms were separated in time: Participants responded to mastery goal orientation items halfway through the training and then completed attention and persistence scales near the end of the training. Participants indicated their agreement with each item on a 5-point scale (1 = *strongly disagree*; 5 = *strongly agree*).

State mastery goal orientation *State mastery goal orientation* ($\alpha = .78$) was operationalized with the four-item scale from Bell and Kozlowki (2008). A sample item is "The opportunity to learn new things about Mahjong is important to me."

Attention To capture the degree to which trainees focus their cognitive resources on the learning task (Zimmerman, 2000), *attention* ($\alpha = .86$) was measured using seven items from Kanfer and Ackerman (1989) and Kanfer, Ackerman, Murtha, Dugdale, and Nelson (1994). A sample item is "I

did not focus my total attention on learning the Mahjong material" (reverse-scored).

Persistence To measure trainees' continued effort toward learning in the face of difficulty and boredom (Elliot, McGregor, & Gable, 1999), *persistence* (α = .88) was assessed with the four-item scale from Elliot et al. (1999). A sample item is "Regardless of whether or not I liked the material on Mahjong, I worked my hardest to learn it."

Learning outcomes and transfer measures Upon the completion of the training session, trainees reported their posttraining self-efficacy and motivation to transfer first, before completing tests on declarative knowledge and strategic knowledge. These four measures mapped on Kraiger, Ford, and Salas's (1993) taxonomy of cognitive (declarative knowledge), skill-based (strategic knowledge), and affective (posttraining self-efficacy and motivation to transfer) learning outcomes. Finally, participants were asked to put their newly acquired knowledge and skills to use in performing the transfer task.

Posttraining self-efficacy To assess the degree to which participants felt confident in following the rules to play Mahjong, *Posttraining self-efficacy* (α = .91) was measured with five items adapted from Ford, Smith, Weissbein, Gully, and Salas (1998). An example item is "I can deal with the decisions surrounding the game of Mahjong." Participants indicated their responses on a 5-point scale (1 = *strongly disagree*; 5 = *strongly agree*).

Motivation to transfer To assess the degree to which trainees were motivated to attempt and apply the newly acquired knowledge and skills (Noe, 1986), *motivation to transfer* (α = .88) was assessed with four items (again adapted for context) from Stevens and Gist (1997) and Warr, Allan, and Birdi (1999). An example item is "I am motivated to apply what I just learned to playing Mahjong." Participants indicated their agreement to each item on a 5-point scale (1 = *strongly disagree*; 5 = *strongly agree*).

Declarative knowledge To assess the degree to which trainees can recognize and recall the key learning points, the test for *declarative knowledge* (α = .84) was extracted directly from the learning material (Huang & Bramble, 2016). The test contained 22 items varying in formats (true-or-false, multiple-choice, and open recall). We dichotomously scored the items (1 = correct; 0 = incorrect) and placed the final score on a 100-point scale for ease of interpretation.

Strategic knowledge Intended to measure whether trainees could adopt the correct behavioral strategy in a given situation, *strategic knowledge* was assessed with 10 scenarios (Huang & Bramble, 2016). In each scenario, trainees were presented with a combination of tiles and were asked to decide on the best strategy to improve their tiles. Identifying the correct strategies required more than remembering and recognizing the learned materials: Trainees needed to integrate the knowledge and decision-making rules to arrive at a decision. Each scenario was dichotomously scored (1 = correct; 0 = incorrect), and the final score was rescaled onto a 100-point scale.

Transfer After completing the learning outcome measures, each trainee played the computer-based game Mahjong against three computer players for 45 min. Trainees were explicitly told that they should perform as well as they could in the transfer session and their scores in the game would be recorded. *Transfer* was operationalized with the proportion of games each trainee won, with an arc sine square root transformation to stabilize the variance of the proportions (Kutner, Nachtsheim, Neter, & Li, 2004). Due to computer program crashes during transfer sessions, transfer scores were not available for six trainees.

IER measures and proxies The fact that trainees were measured at different time points made it possible to assess IER at each occasion. Specifically, the pretraining survey presented an opportunity to obtain multiple indices for IER. Although IER measures could not be obtained beyond the pretraining survey, behaviors and outcomes associated with IER were identified for the other time points: (a) during training; (b) during learning outcome assessment; and (c) during the transfer task. As we conceptualize IER along a continuum (see Huang, Bowling, et al., 2015b), we retained the original scores on the IER measures instead of dichotomizing respondents into attentive versus inattentive categories. In addition, due to our focal interest in the potential effects of IER, we retained all respondents in this study regardless of their levels of IER.

Pretraining overall IER Following Huang et al. (2012) and Meade and Craig (2012), four specific IER indices were calculated: (a) a three-item infrequency scale (α = .80; sample item "I eat cement occasionally") from Huang, Bowling, et al. (2015b) scattered within the pretraining survey; (b) a psychometric antonym index; (c) a psychometric synonym index; and (d) individual reliability.[3] We conducted a principal axis factor analysis on these four IER indices, and all decision rules we examined, including parallel analysis (Hayton, Allen, & Scarpello, 2004), the scree plot, and the Kaiser criterion, indicated a single factor solution, with the single factor accounting for 63.36% of observed variance.

[3] See Huang et al. (2012) and Meade and Craig (2012) for procedures to calculate the psychometric antonym index, psychometric synonym index, and individual reliability. Details of all IER indices are available from the first author.

The four IER indices had an average loading of .72 on the latent factor, with a minimum loading of .67 from the 3-item infrequency scale. Given the factor analytic results, we first standardized these four IER indices and then averaged the four standardized indices into *pretraining overall IER* ($\alpha = .81$).

IER during training As IER was not the focus of the original training study, the training period did not include specific measures of IER. The total duration of the learner-controlled training session (*time for training*, in minutes) recorded by MediaLab served as a proxy for IER during training (see Huang, Bowling, et al., 2015b).

IER during learning outcome assessment Similar to IER during training, IER during learning outcome assessment was assessed with a proxy, namely the duration of the assessment (*time for assessment*, in minutes) recorded by MediaLab. The use of time for assessment to capture IER is consistent with Wise and Kong (2005), who identified quick response time as an indicator of unmotivated solution behavior on tests.

IER behavior during transfer task IER behavior during transfer task was captured with two sources of observational data: (a) experimenters' notes; and (b) screen recordings of transfer task. First, since the study was originally designed to understand learning and transfer, experimenters were trained to take notes about any questionable or aberrant behaviors throughout experimental sessions. Second, each trainee's screen was recorded while he/she played the Mahjong game as the transfer task.

Two research assistants independently analyzed the experimenter notes and screen recordings to identify clear indications of *aberrant transfer behavior*. They flagged 16 participants who manifested aberrant transfer behavior according to experimenters' notes (e.g., dozing off; using cell phone despite earlier instruction not to do so) or screen recordings (e.g., letting the game progress and failing to act for a prolonged period of time). The interrater agreement was 100%. It should be noted that the flagged aberrant behaviors did not include all possible forms of IER behavior during transfer (e.g., randomly selecting possible moves without thinking about strategies), making the detection percentage potentially lower than the actual IER occurrence rate (see Meade & Craig, 2012). Thus, the current aberrant transfer behavior measure should be viewed as a conservative measure of the actual IER behavior during transfer.

Study 1 results

Table 1 presents descriptive statistics and intercorrelations for study 1 variables. An initial inspection of the correlations between substantive variables did not yield any alarming relationships, as evidenced by three observations. First, measures of second-order constructs were strongly related to measures of corresponding lower-order components. CSE was purported to be a higher-order construct that included emotional stability as a component (Judge et al., 2003), and a positive association between them was found as expected ($r = .61$, $p < .001$). Similarly, a positive association was anticipated and observed between WPT-Q and verbal ability ($r = .49$, $p < .001$) because the former measures general cognitive ability that subsumes the latter. Second, conceptually overlapping variables shared positive associations, as shown in the positive correlations between (a) the two self-regulatory mechanisms (attention and persistence; $r = .45$, $p < .001$); (b) the two knowledge measures (declarative and strategic knowledge; $r = .45$, $p < .001$); and (c) the two affective learning outcomes (posttraining self-efficacy and motivation to transfer; $r = .50$, $p < .001$). Finally, the results were largely consistent with the expected role of CSE in the learning and transfer processes (see Stanhope, Pond, & Surface, 2013). Despite a nonsignificant association with state mastery orientation ($r = .11$, $p = .100$), CSE had significantly positive correlations with attention and persistence (rs = .19 and .18, ps < .001), and significantly positive associations with all four learning outcomes (rs ranging from .23 to .45, ps < .001) and transfer ($r = .21$, $p < .001$).

Given the observations above, an unsuspecting researcher might not find anything suspicious about the bivariate correlations between substantive measures. However, an examination of the relationships between IER indices and substantive variables revealed otherwise — that several substantive variables had been contaminated by IER. For instance, pretraining overall IER was negatively correlated with CSE ($r = -.30$, $p < .001$), openness to experience ($r = -.43$, $p < .001$), and conscientiousness ($r = -.26$, $p < .001$).[4] Thus, some trainees who reported having lower CSE, openness to experience, or conscientiousness might have also paid limited attention to the content of the survey items. Echoing the findings from Huang, Liu, and Bowling (2015c) and McGonagle et al. (2016), albeit with a smaller number of scales, the associations between pretraining overall IER and the six pretraining substantive survey measures (i.e., CSE and personality traits) depended on the mean of each survey measure ($r = -.84$, $N = 6$, $p = .035$). Furthermore, IER behavior was somewhat stable across different measurement formats and occasions, the correlations between IER indices were all significant. That is, respondents who engaged in higher levels of pretraining overall IER tended to spend shorter amounts of time during

[4] Although it may be tempting to interpret these correlations as if they indicate substantive associations (e.g., respondents with low CSE tended to engage in IER), we encourage readers to interpret these relationships with caution. We do not intend to establish associations between IER and substantive constructs in this paper because respondents who engage in IER produce scores that may not validly indicate their standing on these substantive measures. As a result, these correlations are partially a function of the methodological confound of IER (see Huang, Liu et al., 2015; McGonagle et al., 2016).

Table 1 Descriptives and intercorrelations for study 1 variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. Core self-evaluations																				
2. Openness	.28																			
3. Conscientiousness	.52	.20																		
4. Extraversion	.41	.39	.49																	
5. Agreeableness	.28	.45	.28	.34																
6. Emotional stability	.61	.26	.37	.34	.30															
7. Verbal ability	.10	.44	.06	.16	.36	.03														
8. WPT-Q	.07	.32	-.08	.03	.22	.07	.49													
9. State mastery orientation	.11	.29	.09	.19	.25	.08	.13	.05												
10. Attention	.19	.25	.07	.08	.17	.16	.19	.14	.32											
11. Persistence	.18	.29	.18	.26	.29	.11	.19	.12	.49	.45										
12. Declarative knowledge	.09	.31	-.09	.03	.27	.11	.45	.54	.24	.37	.41									
13. Strategic knowledge	-.05	.16	-.16	-.14	.03	.01	.23	.21	.18	.23	.26	.45								
14. Posttraining self-efficacy	.22	.27	.15	.20	.14	.22	.17	.07	.32	.26	.50	.30	.18							
15. Motivation to transfer	.21	.36	.23	.25	.36	.16	.23	.16	.57	.45	.60	.39	.22	.50						
16. Transfer	.02	.09	-.03	.02	.05	.06	.18	.25	.11	.21	.19	.33	.34	.15	.16					
17. Pretraining overall IER	-.30	-.43	-.26	-.36	-.59	-.32	-.48	-.38	-.22	-.19	-.24	-.40	-.18	-.20	-.29	-.24				
18. Time for training	.12	.20	.05	.07	.26	.10	.23	.08	.18	.27	.33	.50	.30	.09	.26	.12	-.31			
19. Time for assessment	.17	.30	.07	.05	.29	.22	.35	.15	.20	.34	.41	.52	.41	.25	.30	.17	-.33	.68		
20. Aberrant transfer behavior	-.05	-.13	-.05	-.08	-.16	-.02	-.18	-.16	-.10	-.18	-.16	-.12	-.10	.05	-.12	.02	.21	-.13	-.13	
M	4.71	4.90	4.69	4.93	5.29	4.41	5.63	22.14	3.65	3.07	3.57	53.97	41.89	2.94	3.80	0.22	0.00	22.99	9.16	0.07
SD	0.83	0.75	0.80	0.82	0.80	0.87	2.18	4.13	0.77	0.92	1.03	22.92	17.45	0.93	0.91	0.25	0.80	7.61	4.33	0.25

Note. $N = 243$, except for correlations involving (a) WPT-Q, where $N = 233$; and (b) Transfer, where $N = 237$.

When $|r| > .12$, $p < .05$; when $|r| > .16$, $p < .01$; when $|r| > .21$, $p < .001$.

training and assessment ($r = -.31$ and $-.33$, $ps < .001$), and were more likely to display observable aberrant transfer behavior ($r = .21$, $p < .001$).

Hypothesis 1 stated that IER would be negatively correlated with scores on objective tests. Indeed, participants' pretraining overall IER, assessed when they responded to survey items, had negative associations with their subsequent verbal ability and WPT-Q scores ($r = -.48$ and $-.38$, $ps < .001$), declarative and procedural knowledge test scores ($r = -.40$, $p < .001$ and $r = -.18$, $p = .006$), and transfer scores ($r = -.24$, $p < .001$).

Hypothesis 2 predicted that partialing out IER would result in decreased magnitude of association between two objective tests. We utilized partial correlations controlling for IER to allow direct comparison against the corresponding confounded zero-order correlations (Olkin & Finn, 1995). Table 2 (upper panel) presents first-order partial correlation between any objective tests, controlling for pretraining overall IER, as well as change from zero-order correlation to first-order partial correlation ($r_{xy} - r_{xy.IER}$). Each of the 10 relationships decreased from zero-order correlation to first-order partial correlation, with ($r_{xy} - r_{xy.IER}$) values ranging from .03 to .11 (average change = .07). We conducted a nonparametric sign test (Dixon & Mood, 1946) to assess whether the directionality of changes from zero-order correlation to first-order partial correlation might have occurred by chance. The sign test indicated a statistically significant reduction (10 successes out of a total of 10 trials, $p = .002$) as a result of partialing out IER, thus lending support to Hypothesis 2. Similar to Huang, Liu, and Bowling (2015c), the significance of change to individual correlations was also calculated using the formula from Olkin and Finn (1995, Model C, p. 160). Consistent with our hypothesis, partialing out pretraining overall IER resulted in significantly smaller correlations for six correlations. In contrast, all four correlations involving strategic knowledge did not decrease significantly after controlling for pretraining overall IER scores.

Hypothesis 3 stated that partialing out IER would change the magnitude of correlation between a test and a survey measure that has a mean away from its scale midpoint. An initial inspection of the correlation between IER and substantive measures revealed that IER was negatively correlated with each survey measure (rs ranging from −.19 to −.59) and each test (rs ranging from −.18 to −.48). Therefore, the correlation between any survey measure and any test was expected to become more negative after partialing out IER as a confound. Consistent with this expectation, all 50 first-order partial correlations were more negative than their zero-order counterparts (see Table 2, lower panel), with ($r_{xy} - r_{xy.IER}$) ranging from .03 to .25 (average change = .10). Again, the nonparametric sign test indicated a statistically significant change in the negative direction (50 successes out of 50 trials, $p < .001$),

Table 2 Changes in observed correlations after partialling out IER

	1. Verbal ability	2. WPT-Q	3. Declarative knowledge	4. Strategic Knowledge	5. Transfer
Objective tests					
1. Verbal ability					
2. WPT-Q	.38, .11***				
3. Declarative knowledge	.32, .13***	.46, .08***			
4. Strategic knowledge	.17, .06†	.16, .05†	.42, .03†		
5. Transfer	.08, .10**	.19, .07**	.26, .07**	.32, .03†	
Survey measures					
6. Core self-evaluations	−.05, .15***	−.05, .11***	−.03, .12***	−.11, .06**	−.06, .08**
7. Openness	.30, .14***	.19, .13***	.16, .14***	.09, .07*	.00, .10**
8. Conscientiousness	−.08, .14***	−.19, .12***	−.22, .13***	−.22, .06**	−.10, .06**
9. Extraversion	−.02, .17***	−.12, .15***	−.12, .16***	−.22, .07**	−.06, .08**
10. Agreeableness	.11, .25***	−.01, .23***	.05, .22***	−.10, .12**	−.11, .16***
11. Emotional stability	−.15, .18***	−.06, .12***	−.02, .13***	−.05, .06**	−.02, .08**
12. State mastery orientation	.03, .10**	−.03, .08**	.17, .07**	.14, .03*	.06, .05*
13. Attention	.11, .08*	.07, .06*	.33, .04*	.21, .03†	.17, .04*
14. Persistence	.09, .10**	.04, .08**	.35, .06**	.23, .03†	.14, .05*
15. Posttraining self-efficacy	.09, .08*	.00, .07*	.25, .06*	.15, .03†	.12, .04*
16. Motivation to transfer	.11, .12***	.06, .10***	.31, .08***	.18, .04*	.10, .06**

First-order partial correlations, controlling for pretraining overall IER, are displayed on the left of the comma. Change from zero-order correlation to first-order partial correlation, computed as ($r_{xy} - r_{xy.IER}$), are displayed on the right of the comma, with the significance of change indicated by asterisks

† $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

thus providing support for Hypothesis 3. In terms of statistical significance of individual correlations, 47 of the 50 changes were significantly negative, with the three exceptions again involving strategic knowledge.[5]

Study 1 discussion

Using data from a learning and transfer task, Study 1 demonstrated that IER is negatively associated with performance on objective tests, thus introducing a confound in estimating relationships between objective tests and survey measures. Specifically, removing variance associated with IER resulted in weaker associations between tests and more negative associations between objective tests and survey measures. In addition, study 1 also supported the assumption that individuals' IER behavior is somewhat stable across time and measurement formats in that IER on the pretraining survey was correlated with survey measure and objective test scores in the subsequent laboratory study. This finding is concordant with Bowling et al.'s (2016) finding that rank-order stability in IER behavior is associated with respondents' personality traits (see also DeSimone et al., 2020). Finally, the pattern of associations between IER and survey scores replicated recent findings about the confounding mechanism of IER, depending on survey measures' mean location (Huang, Liu, & Bowling, 2015c; McGonagle et al., 2016).

Consistent with previous literature (Credé, 2010; DeSimone & Harms, 2018; Huang, Liu, & Bowling, 2015c), study 1 serves as proof-of-concept that the presence of IER can influence study results. The current results extend this literature by (1) specifying the direction of this influence; (2) elucidating the reasons for this influence; and (3) demonstrating that measuring and partialing out IER can help control the potential confounding effect introduced by IER. Given the potential impact of IER, it is important to better understand the characteristics of IER that serve to influence empirical relationships. Previous research has demonstrated that study results are more strongly influenced when the prevalence of IER in a sample is high (Credé, 2010; DeSimone et al., 2018). However, as noted above, most existing IER simulation research treats IER dichotomously, simulating fully attentive or fully inattentive respondents. We suspect that participant IER behavior is unlikely to be perfectly consistent throughout the course of a survey (Berry et al., 1992; Clark, Gironda, & Young, 2003). Study 2 extends study 1's findings by examining the confounding effects of IER on study results as a function of consistency and prevalence of IER.

Study 2: simulating parameters behind the confounding effect

The finding that pretraining IER was associated with survey measures and objective tests across different phases of study 1 indicates the presence of some level of within-person consistency in IER behavior. Past research suggests that IER behavior is more likely to be sporadic than completely consistent (Berry et al., 1992). The question remains whether the confounding effect found in study 1 is disproportionally caused by consistent intentional random responding to most items or by occasional careless responding to a small number of items. Prior research has demonstrated that even partial or intermittent random responding can influence psychometric estimates and empirical relationships (Clark et al., 2003; DeSimone & Harms, 2018). However, no study to date has empirically explored the relationship between within-person consistency of IER and the deleterious impact of IER on correlations.

Simulations of IER have typically assumed perfect within-person consistency to demonstrate the confounding effect — i.e., dichotomizing simulated responses as either fully attentive or random (DeSimone et al., 2018 [study 1]; Huang, Liu, & Bowling, 2015c). Observational studies, on the other hand, cannot accurately assess within-person consistency, and thus lack the ability to conclude when and how within-person consistency of IER contributes to IER's confounding effect. A closer examination of within-person consistency would require treating IER as *a process* in which a participant responds to each individual item either attentively or inattentively.

A second parameter that requires examination in conjunction with within-person consistency is the percentage of IER cases in a sample. Differing IER rates have been reported in the literature — e.g., as low as 3.5% (Johnson, 2005); 10–12% (e.g., Meade & Craig, 2012); as high as 73% of very mild form of IER (Baer, Ballenger, Berry, & Wetter, 1997). Despite initial evidence indicating a greater percentage of perfectly consistent IER cases can introduce a larger confounding effect in surveys (Credé, 2010; Huang, Liu, & Bowling, 2015c), the joint influence of the percentage of IER cases and within-person consistency on IER's confounding effects remains unknown.

Exploring the joint impact of within-person consistency and the percentage of IER can further explicate conditions under which IER inflates observed relationships. Further, examining these two parameters can provide practical input for researchers using survey measures and objective tests, allowing them to determine where to devote resources in preventing IER (e.g., trying to limit IER behavior vs. minimizing the number of inattentive respondents).

[5] The lack of support for Hypotheses 2 and 3 involving strategic knowledge might be attributable to the high difficulty level of the strategic knowledge test. On a difficult test, attentive respondents cannot outperform inattentive respondents by much — a floor effect. As a result, IER will share a small correlation with observed test score. In the case of strategic knowledge, the test was indeed quite difficult (M = 41.89 on a 100-point scale) and its correlation with IER was rather weak (r = −.18).

Within-person consistency (*C*) of IER

Despite the conceptual ease of dichotomizing study responses into attentive and inattentive or careless ones, researchers have recognized that response effort (or the lack thereof) varies in degree: Some respondents pay attention throughout the study, some commit occasional inattentive errors, and some exhibit more consistent IER (Huang et al., 2012; Meade & Craig, 2012). Indeed, IER reflects a response process that occurs at the item level, and a respondent can be attentive or inattentive in his/her response to each item. Within-person consistency of IER (noted as *C* hereafter), denotes the percentage (ranging from 0 to 100%) of items to which a person gives inattentive responses. Thus, a *C* of 0% indicates the respondent is consistently attentive, whereas a *C* of 100% indicates the respondent is completely inattentive.

It is worth noting that all IER indices should be related to *C* but most are too coarse to capture IER processes at the item level across an entire survey. For instance, infrequency items focus on responses to particular items, inconsistency indices rely on a specific set of items, and response time measures capture fast IER, usually at either the page or survey level. Thus, understanding *C* is necessary to investigate how within-person consistency of IER relates to IER's confounding effect.

Since *C* is difficult to directly measure or manipulate, a simulation remains the optimal method of investigating the effects of *C* on IER indices and the confounding effect of IER on study results. As noted above, most extant simulation studies have treated individuals as either perfectly attentive ($C = 0\%$, or 0% IER) or perfectly inattentive ($C = 100\%$, or 100% IER). Allowing *C* to assume values between 0 and 100% gives rise to different, more ecologically valid possibilities. For instance, when *C* is small, IER behavior is only sporadic and may be limited to a small number of items or scales. Thus, a small *C* may contribute little to confounding interrelations between different variables. In contrast, when *C* is high, IER behavior is common and spread across different items and scales, thus compounding the confounding effect across different variables.

Hypothesis 4: *C will be positively related to the confounding effect of IER.*

Compared with this hypothesized linear association between *C* and IER's confounding effect, there is reason to suspect *C* may have nonlinear influence on the confounding effect of IER. Individual response effort may fluctuate over the course of a survey due to lapses in attention or distractions. Additionally, participant interest may fluctuate based on the content or style of survey measures. As the level of interest, distraction, and attention may influence IER (Meade & Craig, 2012), differences in response effort spread out across different objective tests and survey measures may cause occasional bias between some variables. In contrast, as IER behavior becomes more and more consistent, the confounding effect of IER may become increasingly pronounced. Thus, the positive influence of *C* on the confounding effect of IER may become stronger at higher levels of *C*. From a practical standpoint, addressing the nonlinear effect of *C* can help scholars answer the question: "Is consistent IER behavior exponentially problematic in causing type I errors?"

Research Question 1: *Does C have a positive quadratic influence on the confounding effect of IER, in addition to its hypothesized positive linear effect?*

Percentage (*P*) of IER cases

The influence of the percentage of IER cases (noted as *P* hereafter) may be intuitive: When there are no IER cases (i.e., $P = 0$), there can be no confounding influence due to IER. More technically, though, the confounding effect of IER stems from the weighted mean difference between attentive and IER participants (Huang, Liu, & Bowling, 2015c). Thus, as long as IER cases do not outnumber attentive ones (i.e., $P \leq 50\%$; a reasonable assumption in most studies), the larger the *P*, the stronger the confounding effect. Indeed, Credé (2010) introduced seven levels of random responses—with *P*s ranging from 0% to 30%—to attentive data and found an increasing trend of confounding effect as *P* increased. Similarly, adding 10% as opposed to 5% of IER cases appeared to create a stronger confounding effect (Huang, Liu, & Bowling, 2015c). Thus,

Hypothesis 5: *P will be positively related to the confounding effect of IER.*

Notwithstanding the expectation of a positive association between *P* and IER's confounding effect, the potential quadratic effect of *P* remains unclear. On the one hand, as *P* approaches .50, the weighted mean difference between attentive and inattentive respondents can be monotonically increasing, translating into a nonsignificant quadratic effect. On the other hand, it is likely when *P* reaches a certain level, the confounding influence will plateau, with additional IER cases adding limited incremental impact. Accordingly, it is possible that the positive association between *P* and the magnitude of the confounding effect will be stronger for low levels of *P*, but approach an asymptotic maximum at mid-to-high levels of *P*.

Research Question 2: *Does P have a quadratic influence on the confounding effect of IER, in addition to its hypothesized positive linear effect?*

Finally, within-person consistency of IER and percentage of IER warrant joint examination: *C* and *P* are unlikely to contribute to IER's confounding effect when either one of them is close to zero. When *C* is close to zero and *P* is high, occasional careless responding from a large percentage of respondents will take on the form of random measurement error. When *C* is high and *P* is close to zero, the presence of a few highly consistent IER cases may not have enough weight to sway the observed result. Thus,

Hypothesis 6: *C and P will interact such that the effect of C is stronger when P is larger.*

A simulation study is most appropriate to address Hypotheses 4–6 and Research Questions 1–2 because it is unfeasible to manipulate or measure various combinations of *C* and *P* in empirical studies. Prior simulations have generated attentive responses from predefined parameters (e.g., Credé, 2010; Huang, Liu, & Bowling, 2015c; Meade & Craig, 2012; Schmitt & Stults, 1985). Using this approach to simulation, researchers would need to make assumptions about parameters of the population from which simulated samples are created (Carsey & Harden, 2014). For example, Meade and Craig (2012) generated attentive responses to five personality scales based on correlations from observed data. Resampling is an alternative approach to simulate multiple samples of data without making assumptions about the population parameters or the sampling process (Carsey & Harden, 2014). Using the resampling approach, researchers start with an observed sample and simulate multiple new samples by drawing, with replacement, from the observed sample (Carsey & Harden, 2014). According to Carsey and Harden (2014), the resampling approach assumes that "all information about the data-generating process contained in the original sample of data is also contained in the distribution of these simulated samples" (p. 201). This approach allows the evaluation of predictions as well as the discovery of potential interactions under realistic situations (Harrison, Lin, Carroll, & Carley, 2007). Specifically, we generated resampled data for study 2 using attentive responses from study 1, thus ensuring the simulated data would closely resemble actual respondent behavior while also allowing for direct comparisons with results from study 1. We began by identifying 200 attentive respondents from study 1 as the starting point for the simulation and resampled them into 10,000 bootstrapped samples. Using the same 10,000 bootstrapped samples, the simulation randomly replaced attentive responses at the item level with random responses for each of the 20 levels of *C* (5% to 100%, in 5% increments) crossed by 20 levels of *P* (2% to 40%, in 2% increments; see "Study 2 method" below for details).

We adopted the 40% upper limit for *P* because we expect most of the studies in organizational research to encounter less than 40% of IER. This expectation is grounded in recent research on IER. Meade and Craig (2012) estimated that 10–12% of students completing a lengthy personality survey engaged in IER, and they subsequently capped their simulated IER cases at 20% of the sample. Other studies have reported similar detection rates, such as 2.5–11.2% (Ran, Liu, Marchiondo, & Huang, 2015), 9.5% (Harms & DeSimone, 2015), and 15–20% (Fleischer, Mead, & Huang, 2015). Thus, we decided to use an upper bound of 40% for *P*. The 20 × 20 factorial design afforded the precision to probe quadratic and interactive effects associated with *C* and *P*. The focal outcome was the difference in correlations between attentive data and the IER-infused data.

The large number of variables in study 1 made it difficult to conduct a computationally intensive simulation on all available variables. Thus, we included the following 10 variables in the simulation: CSE, verbal ability, WPT-Q, state mastery orientation, persistence, attention, declarative knowledge, strategic knowledge, posttraining self-efficacy, and motivation to transfer. This selection of variables represented a reasonable mix of tests and self-report measures that might be examined in a training context (see Colquitt, LePine, & Noe, 2000; Stanhope et al., 2013; Huang, Blume, Ford, & Baldwin, 2015a). Moreover, we expect generalizable findings from these 10 variables, as the magnitudes of the correlations among these variables covered the wide distribution of effect sizes commonly observed in the literature (see Bosco, Aguinis, Singh, Field, & Pierce, 2015).

It should be noted that the means of two study variables, attention and posttraining self-efficacy, were near the scale midpoint, making them unlikely to be confounded by IER (see Huang, Liu, & Bowling, 2015c). Instead, correlations involving attention and posttraining self-efficacy were expected to be attenuated due to the presence of IER (Credé, 2010; Huang, Liu, & Bowling, 2015c). Thus, we chose to include these two survey measures as a contrast to the other measures and did not involve them in formal examinations of Hypotheses 4–6 and Research Questions 1–2.

Study 2 method

Attentive responses

From the 243 respondents in study 1, 200 attentive respondents (82%) were drawn to form the basis for this simulation. The sample size of 200 made it easy to simulate different levels of P without rounding (e.g., 5% of 200 was 10 cases). The exclusion of the 18% least attentive cases was slightly higher than Meade and Craig's (2012) estimated 10–12% of careless respondents in undergraduate populations. We used this conservative screening percentage to ensure that the 200 retained cases were sufficiently attentive in their responses to the surveys and tests.

Table 3 Descriptives and intercorrelations for attentive subsample in study 2

	1	2	3	4	5	6	7	8	9	10
1. Core self-evaluations										
2. Verbal ability	.00									
3. WPT-Q	.03	.44								
4. State mastery orientation	.05	.09	.02							
5. Persistence	.17	.13	.07	.47						
6. Attention	.20	.15	.11	.30	.46					
7. Declarative knowledge	.00	.35	.52	.20	.38	.34				
8. Strategic knowledge	− .07	.24	.26	.22	.23	.22	.44			
9. Posttraining self-efficacy	.22	.12	.09	.31	.52	.26	.29	.22		
10. Motivation to transfer	.21	.10	.09	.54	.58	.42	.30	.20	.46	
M	4.77	6.03	22.68	3.71	3.70	3.17	58.09	43.80	3.00	3.93
SD	0.84	1.93	3.81	0.72	1.01	0.93	22.18	17.18	0.92	0.83

$N = 200$, except for correlations involving WPT-Q, where $N = 190$

Of the 43 excluded suspect IER cases, 16 respondents were first removed for displaying ostensible aberrant response behavior. An additional 27 respondents were removed for having the highest scores on two variables: (a) pretraining overall IER scores; or (b) a composite time measure based on standardized time for training and time for assessment ($\alpha = .81$). The final attentive sample had an average age of 21 years old ($SD = 5$), and 65% were female. These demographic variables were not statistically different from the full sample in study 1. Descriptive statistics and intercorrelations of this sample are presented in Table 3.

The focus of the simulation was to examine the influences of *C* and *P* in the *population* of attentive responses from which the sample of 200 were drawn. Thus, 10,000 bootstrapped samples were created to mimic the population of interest (Carsey & Harden, 2014). Specifically, an SPSS syntax was used to randomly and independently sample respondents from the attentive sample with replacement to form the 10,000 bootstrapped samples ($N = 200$ each). The same 10,000 bootstrapped samples provided the basis for the subsequent 400 simulation conditions (20 $C \times 20$ P).

Correlations for the 10 study variables were estimated for each of the 10,000 bootstrapped samples. For the relationship between any two variables, the median of the 10,000 correlations closely resembled the correlation from the raw data from the attentive sample (see Table 3), with negligible differences ranging from −.002 to .002. Thus, the median correlations from the bootstrapped samples offered a reasonable target for observing any biasing effects due to simulated IER.

Simulated IER

Attentive responses were replaced with random responses to simulate IER behavior. At low levels of *C* (e.g., 5% or 10%), there was a low probability for an attentive participant's responses to be replaced with random responses, representing a reasonable approximation for sporadic careless responding. In contrast, at high levels of *C* (e.g., 80% or 90%), most of the participant's responses would be replaced with random responses, thus simulating deliberate random responding. Although *C* and *P* were manipulated orthogonally and independently, the simulations of *P* and *C* are presented below in sequence for ease of description. All simulations were implemented in SPSS[6].

P parameter For each level of *P* (ranging from 2% to 40% in 2% increments), the specified percentage of respondents in each of the 10,000 samples were randomly selected to exhibit IER behavior (see simulation of *C* below). For example, when *P* was 10%, 20 respondents in each sample were identified as potential IER cases. To implement the random selection, a random number was generated from a uniform distribution for each case, and the lowest *P* percentage of cases from each sample was identified as IER cases.

C parameter For each identified IER participant (see simulation of *P* above), *C* defined the probability that each of his/her responses would be randomly replaced with random responses, ranging from 5% to 100% in 5% increments. For instance, when *C* was 20%, each attentive response had a 20% chance of being substituted by a random response. To determine whether a response would be replaced, a single trial was made from a binomial distribution with a population success rate equaling *C*.

Generating random responses The generation of random responses depended on item types: (a) Likert-type response; (b) true-or-false and multiple-choice; (c) open-ended questions; and

[6] Syntax available at https://osf.io/cwt7z/

(d) WPT-Q. For Likert-type survey items, uniform distributions with equal probabilities for each response option were used to generate random responses. For true-or-false and multiple-choice items, the probability of answering each question correctly was determined by the number of response options available (e.g., .50 for a true-of-false question and .20 for a five-option multiple-choice question). The response process for each true-or-false or multiple-choice item was simulated with a single trial from a binomial distribution with a population success rate equaling the guessing probability. For open-ended recall questions, a random response was simulated with a score of 0 because putting no effort into responding these questions would result in no scores at all. Finally, the random replacement scheme was modified for WPT-Q because item-level data were not made available by the publisher. Using a random number generator, a research assistant completed 50 WPT-Q tests in a completely random fashion and obtained an average score of 12.44 ($SD = 2.16$) for subsequent simulation input. When a WPT-Q score was randomly selected to be replaced by random responses, a standard normal distribution ($M = 12.44$, $SD = 2.16$) was then used to generate a WPT-Q IER score.

Simulation output The focal output of the simulation was the intercorrelations for the 10 study variables. After data generation for each C by P combination, variable intercorrelations were computed for each of the 10,000 bootstrapped samples. The median of the 10,000 correlation coefficients between any two variables was then retained under each condition for subsequent analysis. Therefore, each of the 400 $C \times P$ combinations resulted in a 10-variable correlation matrix.

Study 2 results

For descriptive purposes, Table 4 presents the average median correlations for the 10 study variables across the 400 simulated conditions. Recall that IER was expected to introduce a confounding effect in all but two of the study variables (i.e., attention and self-efficacy). Consistent with this expectation, visual inspection of the median correlations revealed that the introduction of simulated IER data tended to inflate the associations for the eight variables with hypothesized confounding effect, compared with the original correlation matrix (Table 3). In contrast, simulated IER data resulted in an overall decrease in associations that involved attention or self-efficacy—an attenuating effect. To prepare for analysis, change in correlation due to IER was calculated by subtracting the corresponding raw correlation from each observed median correlation. Thus, a positive change would indicate a confounding (i.e., inflating) influence, whereas a negative change would indicate an attenuating influence.

To evaluate the hypothesized effects of C and P on observed correlations, we first considered the nonindependence of observations, as the 45 correlations in the 10-variable correlation matrix were repeated across the 400 $C \times P$ combinations. Thus, we adopted random coefficient modeling to account for such nonindependence. We conducted the analysis with the Multilevel package (Bliese, 2016) in *R*, with C and P mean-centered prior to computing their quadratic and interactive terms. First, where relations were hypothesized to be inflated by IER, changes in correlations among the eight focal variables (CSE, verbal aptitude, WPT-Q, state mastery goal orientation, persistence, declarative knowledge, strategic knowledge, and motivation to transfer) were modeled as outcome variables ($n = 28$) for each condition. In contrast, where relations were expected to be attenuated, changes in any correlations involving attention or posttraining self-efficacy were modeled in a separate random coefficient regression model ($n = 17$ for each condition).

Table 5 presents the results of random coefficient modeling. The unstandardized coefficients (Bs) were obtained

Table 4 Average median correlations across all 400 simulated conditions

	1	2	3	4	5	6	7	8	9
1. Core self-evaluations									
2. Verbal ability	*.09*								
3. WPT-Q	*.11*	*.50*							
4. State mastery orientation	*.09*	*.17*	*.10*						
5. Persistence	*.19*	*.18*	*.12*	*.44*					
6. Attention	.19	.14	.10	.27	.43				
7. Declarative knowledge	*.08*	*.42*	*.55*	*.24*	*.39*	.32			
8. Strategic knowledge	− *.01*	*.29*	*.30*	*.23*	*.24*	.21	*.45*		
9. Posttraining self-efficacy	.20	.10	.07	.27	.47	.24	.26	.20	
10. Motivation to transfer	*.24*	*.18*	*.17*	*.51*	*.54*	.39	*.34*	*.23*	.41

Correlations marked in italic are hypothesized inflated relations, i.e., bivariate relations that did not involve (a) attention or (b) posttraining self-efficacy
Correlations not in italic are expected attenuated relations, i.e., bivariate relations that involve (a) attention and/or (b) posttraining self-efficacy

Table 5 Random coefficient modeling on bivariate relations

	Hypothesized inflated relations			Expected attenuated relations		
	B	Pseudo R^2	Residual variance	*B*	Pseudo R^2	Residual variance
Intercept	.03***		.0038	− .03***		.0005
Step 1: *C*	.15***	.42	.0018	− .04***	.26	.0004
Step 2: *P*	.13***	.57	.0016	− .11***	.60	.0002
Step 3: C^2	.20***	.63	.0014	.02***	.60	.0002
Step 4: P^2	− .41***	.64	.0014	− .01	.60	.0002
Step 5: $C \times P$	.51***	.71	.0011	− .18***	.67	.0002

Hypothesized inflated relations: bivariate relations that did not involve (a) attention or (b) posttraining self-efficacy; $N = 11{,}200$ (i.e., 28×400)

Expected attenuated relations: bivariate relations that involve (a) attention and/or (b) posttraining self-efficacy; $N = 6800$ (i.e., 17×400)

*** $p < .001$

from the final regression model with all five predictor terms. In contrast, to explicate the unique influence of each predictor, pseudo R^2 values based on the intercepts-only model (Kreft & de Leeuw, 1998; Singer, 1998) and residual variance were obtained as each predictor was entered into the regression model at each step. First, the left panel of Table 5 reports hypothesized inflations due to IER. Recall that Hypotheses 4 and 5 predicted that *C* and *P* would have positive influences on bivariate associations. The results supported these hypotheses. Specifically, the observed correlations became more positive with the increase of *C* ($B = .15$, $p < .001$) and *P* ($B = .13$, $p < .001$). Thus, as *C* or *P* increased, IER had a stronger confounding effect on observed relations. Hypothesis 6 predicted a $C \times P$ interactive effect would amplify the confounding effect. The observed positive interactive effect ($B = .51$, $p < .001$) also supported this hypothesis: As *P* increased, the confounding effect of IER due to *C* became stronger. The results also revealed significant quadratic effects for *C* and *P* that addressed Research Questions 1 and 2. Specifically, C^2 had a positive effect ($B = .20$, $p < .001$), indicating that the increase in the positive influence of *C* accelerated as *C* became larger. In contrast, P^2 had a negative effect ($B = -.41$, $p < .001$), suggesting the increase in the positive influence of *P* decelerated as *P* became larger. It is worth noting that the pseudo R^2 contributed by P^2 was 1%, indicating the effect of P^2 might not be practically meaningful. Together, these five predictors accounted for 71% of residual variance from the intercept only model.

The right panel of Table 5 reports the effects of *C* and *P* on expected attenuation due to IER. Both *C* ($B = -.04$, $p < .001$) and *P* ($B = -.11$, $p < .001$) had negative impact on observed correlations. Thus, as *C* or *P* increased, an observed correlation became weaker. Next, entering C^2 ($B = .02$, $p < .001$) and P^2 ($B = -.01$, $p = .648$) explained negligible amount of residual variance, suggesting their roles were quite limited. Finally, a significant $C \times P$ interaction was found ($B = -.18$, $p < .001$), indicating the attenuating effect of *C* became stronger as *P* increased. The presence and nature of the interaction between *C* and *P* further supports Hypothesis 6.

The random coefficient models reported above provided overall predictions across 28 and 17 bivariate associations, respectively. Conceptually, one can think of the modeling results as an aggregate of 28 (or 17) multiple regression models, each for a bivariate association. For illustration, the same five predictors (*C*, *P*, C^2, P^2, and $C \times P$) were entered into multiple regression models to predict two observed associations: (a) between CSE and declarative knowledge and (b) between CSE and posttraining self-efficacy (see Table 6). The former was one of the 28 bivariate associations inflated by the presence of IER, while the latter was one of the 17 bivariate associations attenuated by the presence of IER. As shown in the upper panel of Fig. 2, the association between CSE and declarative knowledge, which was .00 in the attentive sample, increased as *C* and *P* increased. The inflation (i.e., confound introduced by IER) was more troubling as IER became more consistent within-person and more prevalent in the sample. In the lower panel of Fig. 2, however, the association between CSE and posttraining self-efficacy, which was .22 in the attentive sample, decreased steadily as *C* and *P* increased. Thus, the attenuation due to IER became stronger as IER became more consistent within-person and more prevalent in the sample.

Study 2 discussion

Results from study 2 provided support for our hypotheses about the roles of within-person consistency of IER (*C* parameter) and the percentage of IER (*P* parameter).

Table 6 Multiple Regression Analyses Illustrating IER's Inflating and Attenuating Effects

	CSE and declarative knowledge			CSE and posttraining self-efficacy		
	B	*β*	R^2	*B*	*β*	R^2
Intercept	.07***			.20***		
Step 1: *C*	.21***	.80	.64	− .04***	− .66	.43
Step 2: *P*	.30***	.45	.84	− .10***	− .66	.86
Step 3: C^2	.15***	.15	.86	.01***	.02	.87
Step 4: P^2	− .39***	− .06	.87	− .02*	.01	.87
Step 5: $C \times P$	.80***	.35	.99	− .19***	− .35	.99

$N = 400$. * $p < .05$; *** $p < .001$

As *C* or *P* increased, the confounding effect of IER increased. Further, *C* and *P* interacted such that the confounding effect of IER peaked when IER was both consistent within respondents and frequent within a sample. Thus, as expected, the influence of IER on study results may be strongest when a large percentage of respondents engages in consistent careless response behaviors. Addressing the research question about the quadratic effect of *C*, study 2 revealed that IER's confounding effect became increasingly strong when IER responses were more consistent within respondents. In contrast, the quadratic effect of *P* was not practically significant.

Study 2 also revealed interesting roles of *C* and *P* on relationships expected to be attenuated by IER. *C* and *P* influenced attenuation similarly to how they influenced

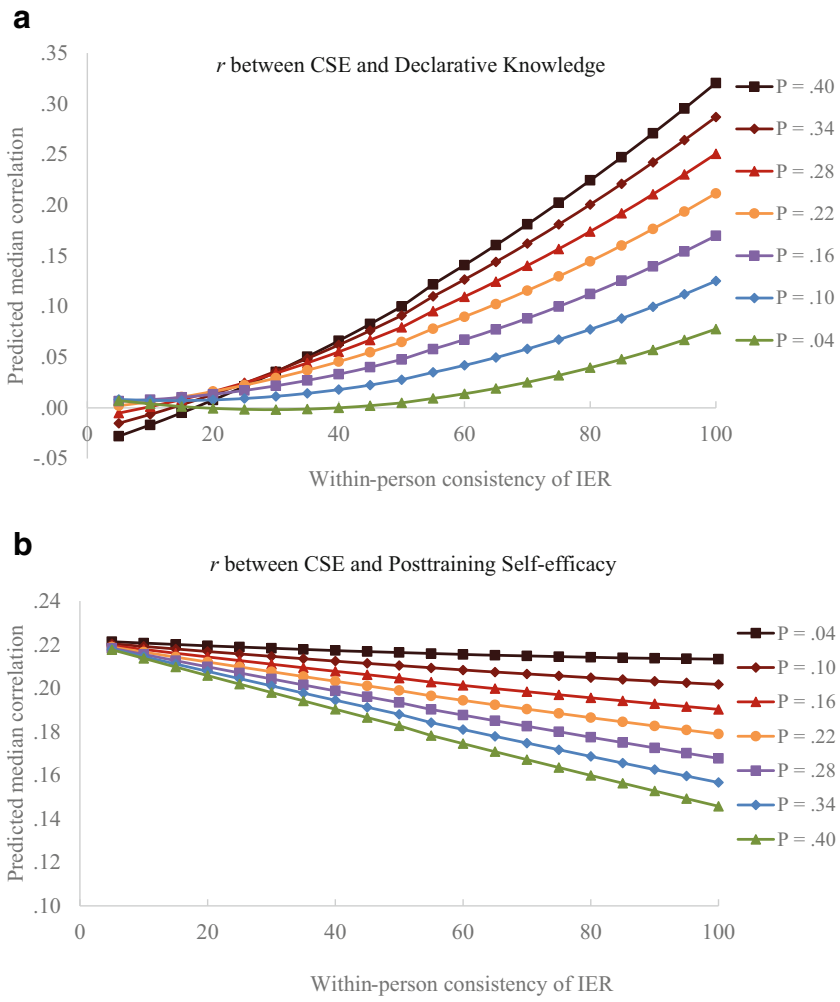

Fig. 2 Correlation between CSE and **a** declarative knowledge and **b** posttraining self-efficacy as a function of C and P

confounding. Specifically, bivariate associations were attenuated when within-person consistency was high or when the percentage of IER cases within a sample was high. Within-person consistency of IER had a stronger attenuating effect when the percentage of IER cases was high. However, neither *C* nor *P* showed a practically meaningful quadratic effect.

General discussion

The present paper makes two timely contributions to the literature. First, combining the confounding mechanism of IER on survey measures with the expected poorer performance of IER on objective tests, the current studies offer evidence that the presence of IER can serve as a confound across measurement methods (i.e., between survey measures and objective tests). These findings demonstrate that the confounding effects of IER are not limited to survey measures. Study 1 demonstrates the potential for IER to influence survey measures, objective tests, and relationships between all combinations of these measures. Study 2 extends these findings by elucidating the role of within-person consistency in IER and the percentage of IER in inflating and attenuating observed bivariate associations. The combined results of these two studies suggest important advice for researchers and practitioners interested in curbing the confounding or attenuating influence of IER.

Research implications

The current studies may inform research in management and organizational psychology in three major ways. First, reviews of the common method variance literature have revealed a strong emphasis on survey measures (Podsakoff et al., 2003; Williams, Hartman, & Cavazotte, 2010), where various stable and transient factors associated with the use of the survey method for data collection can inflate observed bivariate relations. The confounding effect of IER, as examined in the present paper, demonstrates that there may be construct-irrelevant factors influencing inflated bivariate correlations beyond common method bias. Indeed, one can anticipate inflated associations across survey measures and objective tests when: (a) the sample's mean on a survey measure does not fall on the midpoint of the response scale, which frequently occurs in our literature (Huang, Liu, & Bowling, 2015c); and (b) some respondents consistently engage in IER behavior. Given that researchers often rely on participants to provide effortful responses on survey measures and objective tests, response effort on the part of study participants becomes an important feature for researchers to assess and monitor throughout the research process.

As a case in point, Chang, Ferris, Johnson, Rosen, and Tan (2012) urged researchers to consider how common method bias contributes to the intercorrelations among the four lower-order CSE traits. They also posited that "general cognitive ability … may inflate the shared variance among the CSE traits" (p. 111). Given the current results, however, one should also consider IER as a potential factor that confounds CSE. Specifically, since the present studies demonstrate the potential for IER to influence estimates of association between two different methods (i.e., survey measures and objective tests), it is unlikely that IER is simply a subset of or driving force behind common method bias. Instead, it is possible that IER and common method bias exert additive or interactive influence on observed correlations. We encourage future research to explore the unique and overlapping effects of IER and common method bias.

Second, a closer examination of IER as a potential confound may help reduce variability of the same bivariate relation across different studies. Such variability may manifest as a wide credibility interval in a meta-analysis, adding to the uncertainty surrounding effect sizes in the research literature. Moreover, researchers should be more cognizant of the potential roles IER can play in replication efforts, as the research community pays increasing attention to replicability of research results (e.g., Kepes & McDaniel, 2013; Open Science Collaboration, 2015). When IER serves as a confounding factor, a replication study where IER behavior is curtailed may correctly fail to find a significant association that was inflated due to IER (i.e., type I error) in the original study. The reverse can occur as well. When IER acts as an attenuating factor, a replication study with higher percentage and severity of IER behavior may report an attenuated association and thus incorrectly fail to support an earlier finding.

Based on the present results, we endorse the use of IER indices and data screening techniques in an effort to more accurately estimate study relationships. IER, like other statistical artifacts (e.g., attenuation due to unreliability, range restriction), may influence observed effects. However, while some of these artifacts follow known mathematical rules (Pearson, 1903; Spearman, 1904; Thorndike, 1949), Study 2 demonstrates that the confounding or attenuating effects of IER depend on variable factors such as consistency and prevalence. Accordingly, although it may not be possible for meta-analysts to directly adjust effect sizes for IER, data screening practices may serve as a valuable meta-analytic moderating effect (Cortina, 2003). To facilitate this practice, and in light of the aforementioned implications for replicability and meta-analysis, we echo previous recommendations (Curran, 2016; DeSimone, Harms, & DeSimone, 2015) for researchers to transparently report their use of IER indices as well as all screening decisions and cutoffs.

Third, as the percentage and within-person consistency of IER are expected to increase later in the assessment process (Galesic & Bosnjak, 2009; Huang et al., 2012), researchers may guard against the potential influence of IER by separating

items for the same construct into early versus later sections. For instance, a researcher interested in the relationship between CSE and declarative knowledge may administer the 12 CSE items in two 6-item batches, with the first batch measured early and the second batch late in the pretraining survey. Controlling for other factors (e.g., randomizing item assignments), a greater rate of IER late in the pretraining survey will result in the second half-scale being more confounded by IER than the first half-scale. Thus, the researcher may observe that declarative knowledge has a stronger association with the second half-scale of CSE than with the first half-scale, and further conclude that IER may confound the relationship if left untreated.

Practical implications

Following our findings as well as recent research on IER (e.g., DeSimone et al., 2015; Huang, Liu, & Bowling, 2015c; Meade & Craig, 2012), we identify the following issues for researchers and practitioners to consider, preferably before data collection (Aguinis & Vandenberg, 2014), to guard against IER's influence on observed associations. Specifically, one should consider measurement context, measure types, and cutoff scores in planning for IER.

Measurement context

Researchers should first consider whether there are important personal consequences associated with responses—not just the act of responding—for the focal participants. In high-stakes measurement contexts, the responses can lead to important personal consequences. As a result, most of the participants will demonstrate sufficient response effort, minimizing concerns over IER. In contrast, in low-stakes survey and testing situations (Huang et al., 2012; Liu, Bowling, Huang, & Kent, 2013), it is reasonable to assume IER is likely to occur, because participants may be motivated to respond, but not motivated enough to respond attentively throughout the study. Typical low-stakes measurement contexts include students participating in a study for course credit (as in our study 1); online respondents such as Mechanical Turk (see Cheung, Burns, Sinclair, & Sliter, 2017) completing a study for small monetary reward; and employees filling out an anonymous organizational survey (e.g., study 1 from Bowling et al., 2016). A common theme is the perceived need to complete a study in exchange for a desirable outcome, be it tangible such as monetary reward or intangible such as recognition from one's supervisor. In these low-stakes measurement contexts, researchers should consider ways to encourage attentive responding (such as building rapport with participants) and deter IER (such as embedding a benign warning message against IER).

Given the nearly ubiquitous focus of previous IER research on survey measures, our results may have important implications for validation studies. Many validation efforts focus on establishing links between desirable organizational outcomes (e.g., safety, performance) and constructs typically assessed using objective tests (e.g., intelligence, ability, aptitude). As noted above, when validation studies are conducted using incumbent employees, respondents may not be as motivated to respond attentively as they would in a high-stakes situation (e.g., selection). Thus, measures of the ostensibly desirable knowledge, skills, and abilities of incumbents may be confounded by IER, which may lead to erroneous conclusions regarding the validity and utility of these measures. The present results underscore the importance of assessing IER when validating objective tests, as doing so may curtail some of the issues that can emerge with respect to the conclusions drawn in those validation studies.

Importantly, the detection of IER in objective tests may differ considerably from IER detection in survey measures. Specifically, the use of consistency indices (e.g., psychometric synonyms or antonyms, personal reliability) or norm-based indices (e.g., Mahalanobis' D) make less sense in the context of an objective test. However, it would still be possible to examine response patterns (e.g., longstring, response time) and direct measures of IER (e.g., self-reported effort or instructed items) in this context. Additionally, alternative indices of IER may emerge as options depending on the nature of the task at hand. For example, in a training context, respondent inactivity or multitasking may be indicative of insufficient effort. Just as online survey administration can facilitate the measurement of response time, researchers and practitioners may need to turn to new technology to capture time spent on-task versus off-task.

Measure types

Our studies indicate that the effects of IER on observed scores depend on whether the focal measure is an objective test or a survey measure. For objective tests, the effect of IER is known: IER will exert a downward bias on scores. We suggest that IER may have a particularly strong influence on tests that require consistent effort to perform well (e.g., simple yet repetitive tasks).

For survey measures, the effects of IER are contingent on scale means in the population. Thus, researchers need to first consider the nature of the survey measure. Survey measures that have skewed means in the population (e.g., dark personality traits; counterproductive work behavior) are likely confounded by IER. Furthermore, researchers may benefit from understanding the sample at hand, because a sample's average standing on a survey measure will determine how much confounding (or attenuating) effect IER can exert. For instance, a self-efficacy measure will be confounded by (i.e., negatively

correlated with) IER when the attentive respondents feel efficacious in general.

As noted above, Podsakoff et al. (2003) address the potential benefits of using multiple types of questions in an effort to mitigate common method bias. We suspect that this practice will also reduce the monotony of repetitively responding to survey items, which may have the added benefit of maintaining respondents' attention. When a study requires respondents to provide responses to many items, occasional changes to response types (e.g., alternating between surveys and tests) may reduce IER rates, potentially decreasing both *C* and *P* in a sample, and leading to more accurate and trustworthy results.

Cutoff scores

The present findings also provide input for setting cutoff scores for various IER indices in practice. Results from Study 2 suggest there is merit in considering Huang et al.'s (2012) proposition to adopt a lenient cutoff, because isolated incidents of IER exert minimal influence on study results that may be tolerable. In contrast, high within-person consistency in IER (e.g., a deliberate attempt to randomly respond) is particularly impactful in confounding some associations and attenuating others. From a practical perspective, it would be unrealistic to expect all participants to respond attentively to all items or to screen out 73% of the sample because they answered a single item carelessly (see Baer et al., 1997). To be clear, minor and isolated incidents of IER may still exert some influence on study results, but the benefits of eliminating any and all flagged IER from a study must be weighed against practical concerns such as sample recruitment costs, statistical power, and false positive rates for flagging IER. Of course, researchers and practitioners will also benefit from taking proactive approaches to encourage high response effort from all participants, as decreasing the percentage of IER can also reduce the confounding or attenuating influence of IER.

Cutoffs for IER detection in survey measures are addressed at length elsewhere (e.g., DeSimone & Harms, 2018), and much more research is required before researchers can begin to address cutoffs for objective tests. Our study can shed some light on the obfuscating effects of levels of *C* and *P*. As shown in Table 5 and depicted in Fig. 2, although any level of *C* or *P* can influence correlations, higher levels of *C* or *P* are clearly more problematic, especially given the effects are interactive such that high *C* is most problematic at high levels of *P*. Hence, our simulation demonstrates that when it comes to *C* and *P*, "any is bad, and more is worse." While the effects of *C* and *P* seem somewhat small at lower levels (e.g., < 20% for *C*, < 10% for *P*), the effects become more pronounced at higher levels (e.g., > 40% for *C*, > 20% for *P*). Researchers should pay particular attention to egregious forms of consistent IER, which, just like a bad apple spoiling a whole barrel, can lead to biased findings. Meanwhile, even typical levels of *P* (10 to 20%) can be meaningfully problematic if careless responders consistently exhibit inattentiveness on as little as one-third to one-half of the questions.

As a caveat, while the previous paragraph generally addresses problematic levels of *C* and *P*, we hesitate to endorse specific cutoff values for either parameter. Reasonable cutoff values for *C* or *P* need to take into account both measurement context and measurement type. Further complicating matters, the levels of *C* or *P* identified in a study may depend on the IER indices employed and the sentivity levels associated with the selected cutoff values used. We call on future research to extend the present results in an effort to determine whether these parameters are differentially important when considering various types of assessment tools (e.g., survey measures, objective tests), or other factors associated with the data collection process (e.g., number and variety of questions, types of constructs, measurement context).

Limitations and future research directions

Limitations of the present studies should be noted. First, the participants from study 1 were recruited from students who may not have a vested interest to remain attentive throughout the study. However, samples from college students have continued to drive important discoveries in management research, including training research that focused on the psychological processes underlying learning and transfer (e.g., Bell & Kozlowski, 2008; Brown, 2005; Chen, Thomas, & Wallace, 2005; Cianci, Klein, & Seijts, 2010; Orvis, Fisher, & Wasserman, 2009). More importantly, the focal research question concerned IER, which may occur in low-stakes measurement contexts regardless of participant types. Indeed, issues regarding IER can also surface when conducting organizational surveys of employees (e.g., Green & Stutzman, 1986; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Kotrba, Nieminen, Denison, & Carter, 2014). Second, the simulation in Study 2 utilized actual human responses as attentive data, resulting in the examination of a few selected variables, as opposed to simulating attentive responses that may allow the study of various variable characteristics (e.g., different means and different distributional properties). Indeed, future simulations can be performed on a wider range of correlations across a larger number of variables to evaluate whether the effects of IER may depend on the magnitude of correlations—for instance, it is possible that the confounding effect of IER may be constrained by a ceiling effect when two substantive variables have near perfect correlation. Despite this limitation, the simulation afforded the opportunity to examine the impact of IER in a realistic research context, with variables highly relevant to training studies (see Stanhope et al., 2013). Additionally, prior research has demonstrated that random responding and straightlining may exert different effects on study relationships (DeSimone et al., 2018).

The current studies suggest several directions for future investigations. First, the impact of within-person consistency of IER demonstrated through our simulation calls for more fine-grained measures of IER. The IER indices available in the current literature (e.g., Huang et al., 2012; Meade & Craig, 2012; Wood et al., 2017) can only serve as coarse proxies for within-person consistency of IER, because some indices such as infrequency, inconsistency, and outlier indices rely on responses to *a limited sample* of survey items, while other indices can only capture specific types of IER, such as speeding through measures and long string response patterns or "straightlining" (Schonlau & Toepoel, 2015). To closely monitor IER behavior throughout assessment, researchers may wish to include physiological measures such as eyetracking and advanced statistical techniques such as item response theory.

Second, researchers need to come to a better understanding of factors that drive IER. Bearing in mind that behaviors are in general under the interactive influence of the situation and the person (Endler & Magnusson, 1976), one may start examining the main effects of the situation and the person as a first step. The situation may be viewed through the lens of situational strength, which is defined as "implicit or explicit cues provided by external entities regarding the desirability of potential behaviors" (Meyer, Dalal, & Hermida, 2010, p. 122). Meyer et al. (2010) discussed four facets of a strong situation: (a) clarity—cues about expected behaviors are present; (b) consistency—cues about expected behaviors are congruent with each other; (c) constraints—choice of behaviors is limited by external forces; and (d) consequences—behaviors will result in positive or negative implications. Future research on IER may examine how variation on these four facets of situational strength in low-stakes questionnaire administration may impact IER and subsequent survey/test results. For example, collecting data in person as opposed to online will increase the constraints on respondent behavior, and mentioning techniques to screen for inattentive responses (e.g., Ward & Pond, 2015) may increase the perceived consequence.

As for characteristics of the person that influence IER, recent research by Bowling et al. (2016) indicates that acquaintance-rated conscientiousness, agreeableness, extraversion, and emotional stability were negatively related to undergraduate students' IER on a survey measure. However, the associations were quite weak. Future research may expand beyond the Big-Five personality framework to examine whether specific traits (e.g., Machiavellianism) can predict IER behavior. For example, DeSimone et al. (2020) found that implicit aggression was associated with IER. Understanding the characteristics of the person that drive IER behavior raises another interesting question for future research. If a conglomerate of personality variables reliably drives IER behavior, then the observed statistical confounding effect due to IER could be partially attributed to the combination of these variables. For instance, if low conscientiousness and high Machiavellianism consistently predict high IER, then a variable contaminated by IER (e.g., general cognitive ability) in a low-stakes test could be alternatively considered to be contaminated by conscientiousness and Machiavellianism.

Conclusion

Extending research that identified the conditions under which IER can confound survey measures, the current paper demonstrates that IER can also confound estimated relationships between objective tests and survey measures in a training study. Furthermore, a simulation extends these findings by revealing how within-person consistency and percentage of IER influence IER's confounding or attenuating effects. The current discoveries add to the emerging stream of research on the potential impact of IER in management research, providing practical guidelines for research design as well as pointing to interesting future investigations.

References

Aguinis, H., & Vandenberg, R. J. (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior, 1*, 569–595.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695–716.

Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment, 68*, 139–151.

Bell, B. S., & Kozlowski, S. W. J. (2008). Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology, 93*, 296–316.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment, 4*, 340–345.

Bliese, P. (2016). Package 'multilevel' (version 2.6). Retrieved April 2017 from: https://cran.r-project.org/web/packages/multilevel/.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*, 1065–1105.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*, 431–449.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*, 218–229.

Brown, K. G. (2005). An examination of the structure and nomological network of trainee reactions: A closer look at "smile sheets". *Journal of Applied Psychology, 90*, 991–1001.

Callans, M. (2012, December 28). Non-proctored vs. proctored assessments. Retrieved from http://blog.wonderlic.com/nonproctoredvsproctoredassessments

Caplan, B., & Miller, S. C. (2010). Intelligence makes people think like economists: Evidence from the General Social Survey. *Intelligence, 38*, 636–647.

Carsey, T. M., & Harden, J. J. (2014). *Monte Carlo simulation and resampling methods for social science*. Thousand Oaks, CA: Sage.

Chang, C. H., Ferris, D. L., Johnson, R. E., Rosen, C. C., & Tan, J. A. (2012). Core self-evaluations: A review and evaluation of the literature. *Journal of Management, 38*, 81–128.

Chen, G., Thomas, B., & Wallace, J. (2005). A multilevel examination of the relationships among training outcomes, mediating regulatory processes, and adaptive performance. *Journal of Applied Psychology, 90*, 827–841.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523–562.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology, 32*, 347–361.

Cianci, A. M., Klein, H. J., & Seijts, G. H. (2010). The effect of negative feedback on tension and subsequent performance: The main and interactive effects of goal content and conscientiousness. *Journal of Applied Psychology, 95*, 618–630.

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back random responding: Effectiveness of MMPI-2 and personality assessment inventory validity indices. *Psychological Assessment, 15*, 223–234.

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85*, 678–707.

Cor, M. K., Haertel, E., Krosnick, J. A., & Malhotra, N. (2012). Improving ability measurement in surveys by following the principles of IRT: The Wordsum vocabulary test in the General Social Survey. *Social Science Research, 41*, 1003–1016.

Cortina, J. M. (2003). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods, 6*, 415–439.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*, 596–612.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19.

DeSimone, J. A., Davison, H. K., Schoen, J. L., & Bing, M. N. (2020). Insufficient effort responding as a partial function of implicit aggression. *Organizational Research Methods, 23*, 154–180.

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology: An International Review, 67*, 309–338.

DeSimone, J. A., & Harms, P. D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology, 33*, 559–577.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36*, 171–181.

Dixon, W. J., & Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association, 41*, 557–566.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences, 108*, 7716–7720.

Elliot, A. J., McGregor, H. A., & Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. *Journal of Educational Psychology, 91*, 549–563.

Endler, N. S., & Magnusson, D. (1976). Toward an interactional psychology of personality. *Psychological Bulletin, 83*, 956–974.

Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive responding in MTurk and other online samples. *Industrial and Organizational Psychology, 8*, 196–202.

Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. *Journal of Applied Psychology, 83*, 218–233.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*, 349–360.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. D. Fruyt & F. Ostendorf (Eds.), Personality psychology in Europe (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Green, S. B., & Stutzman, T. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology, 39*, 543–564.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.

Harms, P. D., & DeSimone, J. A. (2015). Caution! MTurk workers ahead—fines doubled. *Industrial and Organizational Psychology, 8*, 183–190.

Harrison, J. R., Lin, Z., Carroll, G. R., & Carley, K. M. (2007). Simulation modeling in organizational and management research. *Academy of Management Review, 32*, 1229–1245.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*, 191–205.

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

Huang, J. L., Blume, B. D., Ford, J. K., & Baldwin, T. T. (2015a). A tale of two transfers: Disentangling maximum and typical transfer and their respective predictors. *Journal of Business and Psychology, 30*, 709–732.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015b). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30*, 299–311.

Huang, J. L., & Bramble, R. J. (2016). Trait, state, and task-contingent conscientiousness: Influence on learning and transfer. *Personality and Individual Differences, 92*, 180–185.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*, 99–114.

Huang, J. L., Liu, M., & Bowling, N. A. (2015c). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*, 828–845.

Jarvis, B. (2008). *MediaLab (version 2008) [computer software]*. New York: Empirisoft.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2003). The core self-evaluations scale: Development of a measure. *Personnel Psychology, 56*, 303–331.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods, 18*, 512–541.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*, 657–690.

Kanfer, R., Ackerman, P. L., Murtha, T. C., Dugdale, B., & Nelson, L. (1994). Goal setting, conditions of practice, and task performance: A resource allocation perspective. *Journal of Applied Psychology, 79*, 826–835.

Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 6*, 252–268.

Kotrba, L. M., Nieminen, L., Denison, D., & Carter, N. T. (April, 2014). Respondent versus response screening: Looking beyond the class clowns. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology: Honolulu, HI.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*, 311–328.

Kraiger, K., & Jerden, E. (2007). A meta-analytic investigation of learner control: Old findings and new directions. In S. M. Fiore & E. Salas (Eds.), *Toward a science of distributed learning* (pp. 65–90). Washington, DC: APA.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied linear statistical models* (5th ed.). New York: McGraw-Hill/Irwin.

Liu, M., Bowling, N. A., Huang, J. L., & Kent, T. A. (2013). Insufficient effort responding to surveys as a threat to validity: The perceptions and practices of SIOP members. *The Industrial-Organizational Psychologist, 51*, 32–38.

Longstaff, H. P., & Porter, J. P. (1928). Speed and accuracy as factors in objective tests in general psychology. *Journal of Applied Psychology, 12*, 636–642.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83.

McGonagle, A. K., Huang, J. L., & Walsh, B. M. (2016). Insufficient effort survey responding: An under-appreciated problem in work and organizational health psychology research. *Applied Psychology: An International Review, 65*, 287–321.

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.

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management, 36*, 121–140.

Noe, R. A. (1986). Trainees' attributes and attitudes: Neglected influences on training effectiveness. *Academy of Management Review, 11*, 736–749.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872.

Pearson, K. (1903). Mathematical contributions to the theory of evolution-XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, London, Series A, 200*, 1–66.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879–903.

Ran, S., Liu, M., Marchiondo, L. A., & Huang, J. L. (2015). Difference in response effort across sample types: Perception or reality? *Industrial and Organizational Psychology, 8*, 202–208.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9*, 367–373.

Schonlau, M., & Toepoel, V. (2015). Straightlining in Web survey panels over time. *Survey Research Methods, 9*, 125–137.

Singer, J. D. (1998). Using SAS PROCMIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*, 323–355.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.

Stanhope, D. S., Pond III, S. B., & Surface, E. A. (2013). Core self-evaluations and training effectiveness: Prediction through motivational intervening mechanisms. *Journal of Applied Psychology, 98*, 820–831.

Stevens, C. K., & Gist, M. E. (1997). Effects of self-efficacy and goal-orientation training on negotiation skill maintenance: What are the mechanisms? *Personnel Psychology, 50*, 955–978.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.

Ward, M. K., & Pond III, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior, 48*, 554–568.

Warr, P., Allan, C., & Birdi, K. (1999). Predicting three levels of training outcome. *Journal of Occupational and Organizational Psychology, 72*, 351–375.

Williams, L. J., Hartman, N., & Cavazotte, F. (2010). Method variance and marker variables: A review and comprehensive CFA marker technique. *Organizational Research Methods, 13*, 477–514.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*, 95–114.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*, 19–38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163–183.

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science, 8*, 454–464.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). San Diego, CA: Academic Press.