Insufficient Effort Responding on Mturk Surveys: Evidence-Based Quality Control for Organizational Research

Lee Cyr
Portland State University

Recommended Citation

10.15760/etd.6337

Let us know how access to this document benefits you.
Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds
Part of the Psychology Commons

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.
Insufficient Effort Responding on Mturk Surveys: Evidence-Based Quality Control for Organizational Research

by

Lee Cyr

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science
in
Psychology

Thesis Committee:
Liu-Qin Yang, Chair
Larry Martinez
Joel Steele

Portland State University
2018
Abstract

Each year, crowdsourcing organizational research grows increasingly popular. However, this source of sampling receives much scrutiny focused on data quality and related research methods. Specific to the present research, survey attentiveness poses a unique dilemma. Research on updated conceptualizations of attentiveness—insufficient effort responding (IER)—shows that it carries substantial concerns for data quality beyond random noise, which further warrants deleting inattentive participants. However, personal characteristics predict IER, so deleting data may cause sampling bias. Therefore, preventing IER becomes paramount, but research seems to ignore whether IER prevention itself may create systematic error. This study examines the detection and prevention of IER in Amazon’s Mechanical Turk (Mturk) by evaluating three IER detection methods pertinent to concerns of attentiveness on the platform and using two, promising, IER prevention approaches—Mturk screening features and IER preventive warning messages. I further consider how these issues relate to organizational research and answer the call for a more nuanced understanding of the Mturk population by focusing on psychological phenomena often studied/measured in organizational literature—the congruency effect and approach-avoidance motivational theories, Big Five personality, positive and negative affectivity, and core self-evaluations. I collected survey data from screened and non-screened samples and manipulated warning messages using four conditions—no warning, gain-framed, loss-framed, and combined-framed messages. I used logistic regression to compare the prevalence of IER across conditions and the effectiveness of warning messages given positively or negatively valenced motivational tendencies. I also used 4x2 factorial ANCOVAs to test for differences in
personal characteristics across conditions. The sample consisted of 1071 Mturk workers (turkers). Results revealed differences in IER prevalence among detection methods and between prevention conditions, counter-intuitive results for congruency effects and motivational theories, and differences across conditions for agreeableness, conscientiousness, and positive and negative affectivity. Implications, future research, and recommendations are discussed.
# Table of Contents

Abstract ......................................................................................................................... i  
Introduction .................................................................................................................. 1  
Background .................................................................................................................. 11  
  Mturk for Organizational Research ........................................................................... 11  
  Insufficient Effort Responding .................................................................................. 13  
  Confounding effects of IER ....................................................................................... 14  
  IER detection .............................................................................................................. 17  
IER Prevention ............................................................................................................. 20  
  Mturk screening. ........................................................................................................... 21  
  Preventive survey instructions. .................................................................................... 23  
Methods ....................................................................................................................... 32  
  Participants and Sampling .......................................................................................... 32  
  Power ............................................................................................................................ 32  
  Measures ...................................................................................................................... 33  
    State approach-avoidance ....................................................................................... 34  
    Approach-avoidance temperament ......................................................................... 34  
    IER ............................................................................................................................. 34  
    Personality ................................................................................................................ 36  
    Positive and negative affectivity .............................................................................. 36  
Research Design .......................................................................................................... 36  
  Experimental manipulation ....................................................................................... 37  
Analytic Strategies ....................................................................................................... 38  
  Manipulation checks ................................................................................................... 38  
  Hypotheses tests ......................................................................................................... 38  
Results .......................................................................................................................... 41  
Preliminary Analysis .................................................................................................... 41  
  Data .............................................................................................................................. 41  
  Manipulation checks ................................................................................................... 42  
Hypotheses Tests ......................................................................................................... 43  
  Hypotheses 1-2 .......................................................................................................... 43  
  Hypothesis 3 ............................................................................................................... 44  
  Hypothesis 4 ............................................................................................................... 45
List of Tables

Table 1: Raw descriptive statistics for focal variables and conditions.......................................................... 71

Table 2: Hypotheses 1-2: Stepwise results predicting IER on each detection protocol................................. 72

Table 3: Hypotheses 4a-f: Congruency across AP-AV constructs and IER detection protocols.......................... 73
List of Figures

| Figure 1: | Prevalence of IER across conditions and IER outcomes | 76 |
| Figure 2: | Parodoxically congruent interactions for infrequency IER | 77 |
| Figure 3: | Differences in personal characteristics across conditions | 78 |
Crowdsourcing organizational research through platforms such as Amazon’s Mechanical Turk (Mturk) presents multiple advantages that make it increasingly popular; however, scholars scrutinize its use for methodological disadvantages that affect data quality and results (Cheung, Burns, Sinclair, & Sliter, 2016; Woo, Keith, & Thornton, 2015). Mturk research is fast and inexpensive (Mason & Suri, 2012). Mturk workers (turkers) are more representative and diverse than student samples (Buhrmester, Kwang, & Gosling, 2011), allow researchers to tap broader samples compared to organizational convenience sampling (Woo et al., 2015), and replicate results in classic phenomena from multiple fields (e.g. Berinsky, Huber, & Lenz, 2012; Wolfson & Bartkus, 2013) including psychology (e.g. Crump, McDonnell, & Gureckis, 2013; Horton, Rand, & Zeckhauser, 2011; Paolacci, Chandler, & Ipeirotis, 2010). However, perspectives and supporting evidence differ on these advantages (e.g. Harms & DeSimone, 2015; Ipeirotis, 2010; Roulin, 2015; Woo et al., 2015). Likewise, some evidence supports Mturk’s research disadvantages such as (in)dependence of sampling, non-naïve participants, repeat participation, and the influence of compensation and turker motivation to participate, but the literature shows that these issues prove non-problematic and/or controllable. For instance, the Mturk platform/compensation system makes it possible and incentivizing to participate in many surveys a day and/or use multiple profiles to participate in the same study multiple times. However, a sufficient sample size (i.e. 800) reduces the number of identical turkers across studies to negligible levels, very few turkers repeat participation (approximately 3%), and one can check IP addresses and ask identifying questions (e.g. gender, age, work status) in such a way to address repeat participation and violation of
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

other study eligibility criteria (see Berinsky et al., 2012; Buhrmester et al., 2011; Crump et al., 2013; Harms & DeSimone, 2015; Shapiro, Chandler, & Mueller, 2013; Stewart et al., 2015). Given the above, scholars have called for researchers to consider the relative advantages, disadvantages, and controls, given a specific study’s needs, and called for a more nuanced understanding of Mturk samples (e.g. Cheung et al., 2016; Landers & Behrend, 2015; Roulin, 2015).

One disadvantage, insufficient effort responding (IER), presents a unique problem in crowdsourcing samples and in general. IER denotes a respondent’s random and/or intentional lack of regard for survey content (Huang, Bowling, Liu, & Li, 2015; Huang, Curran, Keeney, Poposki, & DeShon, 2012), and it differs from other forms of response bias such as missing not at random (MNAR). For instance, MNAR applies to the case where missing data requires multitudes of analyses to assess (i.e. whether gender, age, or other characteristics explain it, and effects on results given why responses are missing may vary). On the other hand, IER subsumes a number of invalid survey response tendencies such as inadvertently answering inaccurately (e.g. mistakenly selecting agree when meaning to select disagree) or intentionally answering items randomly or inaccurately/untruthfully. To be clear, I came across no literature validating IER as an MNAR analysis, but since respondents might inadvertently or intentionally skip a question, IER may explain missing responses as well as it explains different types of non-missing invalid responses. Thus, IER represents a somewhat comprehensive term, but the present research focuses on non-missing, invalid responses and their effects on data and analyses.
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

Similar to other concerns, the Mturk platform requires special consideration for IER and IER detection. First, multiple IER detection methods exist and each to its own merits; I describe IER detection methods in greater detail in the pertinent section below (I refer to these methods in three different ways—detection methods refer to the type of scale or statistic, detection protocols refer to the process, detection index/indices refers to the resulting operationalization). Based on review data (Cheung et al., 2016) and experiences with colleagues, researchers largely use only a single type of method—oft referred to as “attention checks” but I call them “instructed responses”—in their surveys (i.e. asking as a survey item “please select strongly agree”); they also tend to use otherwise non-validated detection questions or only a single item. However, different types of IER detection methods may detect different types of IER and research suggests that turkers know and scan for instructed responses (e.g. Blanchard & Banerji, 2015). That is, turkers might scan for instructed responses but speed through substantive survey items leaving only a narrow window of IER that instructed responses can detect and much IER undetected. Unfortunately, until a recent paper (Wood, Harms, Lowman, & DeSimone, in press), minimal evidence alluded to the validity of using response times as a measure of IER, and no research validated a specific measure of response times. Even still, although better evidence toward the validity of a speed index exists, research lacks for how a response time index holds up to the broader concepts of IER. For instance, a response time based IER index should relate to other IER indices and substantive scale variables in predictable ways (see the IER and IER detection sections in the present study). Similarly, other validated IER detection methods, such as the infrequency scale, are available and these use items that better fit in an organizational survey (e.g. I work
and might better withstand quick identification. Second, research yields mixed results when comparing the prevalence of IER between subject pools—students, work/field samples, Mturk, and other online forums such as discussion boards (see Goodman, Cryder, & Cheema, 2012; Hauser & Schwarz, 2016; Klein et al., 2014; Paolacci et al., 2010). And, this may stem from variations in protocols used, pass/fail definitions, or research purposes (e.g. consumer research vs social psychology and experiments vs surveys). Nevertheless, Mturk receives great scrutiny regarding IER; Chandler and others (2014) found that although 73% of turkers completed surveys when alone, 28%, 27%, 18%, 14%, and 6% were reading blogs, with other people, watching TV, listening to music, or chatting online during surveys, respectively. These distractions may increase IER. (Meade & Craig, 2012).

Most importantly, perhaps, IER presents an “insidious confound”. If left unaddressed, IER can invalidate factorial structures (Woods, 2006) and attenuate bivariate relationships (McGrath, Mitchell, Kim, & Hough, 2010), which inflates type II errors. Also, IER can increase correlations and multiple regression coefficients (Huang, Liu, & Bowling, 2015; McGonagle, Huang, & Walsh, 2016) inflating type I errors. I describe these effects in later sections. Standard recommendations include detection, deletion, and reporting IER. However, these strategies may prove costly and time consuming, especially, when data contains high IER. Ran, Liu, Marchiondo, and Huang (2015), among other scholars, have called for increased practice of, and research for, IER prevention. What is more, the latest research shows that IER extends beyond methodological issues. Individuals who commit IER likely do so consistently (Camus, 2015), and personal characteristics (e.g. conscientiousness, agreeableness, extraversion,
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

openness, and neuroticism) predict IER (Bowling et al., 2016). As such, removing IER may create a specific profile of respondents resulting in sampling bias; the literature has not examined this effect. In sum, researchers face a dilemma where they cannot leave IER in their data due to its influence on factorial structures, means, and relationships but cannot remove it without, potentially, systematically removing those having particular characteristics.

Research on IER prevention reveals two promising approaches. However, both approaches may come with flaws and leave questions unanswered. Perhaps most promising, Mturk tracks turker activities, and researchers can use the system’s screening features to set eligibility criteria such that only turkers known to do quality work may participate. Quality work in this regard refers to the percent of previous work completed by turkers that the party making it available deemed sufficient and accepted. I describe screening in more detail below. Using a relatively short survey in two studies, Peer, Vosgerau, and Acquisti (2014) showed much lower IER prevalence among screened turkers; moreover, those below screening cut-offs, that passed IER protocols, replicated previous research, and since IER prevalence was low above cut-offs, screened turkers replicated previous research regardless. Nevertheless, multiple considerations for IER not apparent in their study warrant further investigation into the screening approach. First, since IER extends beyond methodological concerns, it remains possible that sampling only those above cut-offs creates sampling bias by selecting in participants having a unique profile of individual characteristics. Second, they used a short survey; longer survey’s predict more IER (Yentes, 2015). Third, for IER detection, the authors used instructed responses such as one instructed response item (i.e. mark “2” for this
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

question). Because turkers likely know and look for instructed responses (Blanchard & Banerji, 2015; Peer et al., 2014), the findings using this type of IER detection protocol may reveal little about whether those above cut-offs pay attention to survey content or consider the meaning of items. Indeed, when using novel, IER, instructed responses, Peer and others (2014) actually found more IER within the above cut-off condition.

Another promising approach uses preventive survey instructions (warning messages). Warning messages simply entail a script before a survey that informs participants that attentive/careful responses are important, will be assessed, and that lack of attentiveness will carry consequences. Research among student samples shows that well-constructed, IER preventive, survey instructions reduce IER (Blackmore, 2014; Huang et al., 2012; Meade & Craig, 2012). Nonetheless, the current use of warning messages may have limitations that warrant further refinement. First, previous studies only used negatively worded consequences such as you will not receive credit for participation. Approach-avoidance motivation theories and research suggests that promotive and preventive messages prove most salient and effective when congruent with a person’s positively and negatively valenced motivational tendencies, respectively (e.g. Carver, Sutton, & Scheier, 2000; Elliot & Thrash, 2010; Hamamura, Meijer, Heine, Kamaya, & Hori, 2009; Higgins, 1997). Scholars know this as the congruency effect. Therefore, employing only negative framing may limit the overall effectiveness of warning messages. What is more, since negatively framed messages may reduce IER more among people higher in negatively valenced characteristics, similar to screening, it could result in sampling bias by leaving only those having specific characteristics in the data after one deletes participants that responded with insufficient effort. Relatedly,
warning messages may create measurement error by influencing moods and/or activating traits such that participants respond differently than if they saw no warnings. Finally, it remains unclear whether the effects of this prevention approach generalize to Mturk populations where sample composition, compensation, and motivation to participate may differ.

The congruency effect, approach-avoidance motivational theories, and personal characteristics (germane to IER and IER prevention) hold significant implications for conducting industrial and organizational psychology (I/O) experiments and survey research on Mturk. Although previous research replicated classic psychological phenomena on Mturk (e.g. Crump et al., 2013; Horton et al., 2011; Paolacci et al., 2010), they almost exclusively focused on cognition such as the Stroop task (Stroop, 1935), task-switch costs (Jersild, 1927), and the “Linda problem”, or, conjunction fallacy (a faulty heuristic/bias for guessing event probabilities; Tversky & Kahneman, 1983). None attempted to replicate the congruency effect. Moreover, they all center on experiments and mostly conducted replications in screened samples of turkers, which may represent only certain types of people/turkers. Thus, although previous research provides promising evidence for the validity of conducting psychological experiments on Mturk, they tell us much less about the Mturk population at whole and experiments and survey research more apropos to I/O studies.

Psychologists and I/O psychologists have advanced numerous gradations of approach-avoidance theories (described in more detail in later sections). Elliot and Thrash (2002, 2010) proposed that approach-avoidance temperament (AAT) represents an overarching, bi-dimensional, latent construct on which positively (e.g. extraversion)
and negatively (e.g. neuroticism) valenced personality and motivation constructs load, and it influences one’s more proximal-to-behavior motivational tendencies. For example, Ferris and colleagues (2011) incorporated core self-evaluations (CSE; defined below) into the AAT framework showing that CSE loaded onto AAT and—the more proximal—approach-avoidance work motivation mediated the relationship between CSE and organizational citizenship behaviors (OCB). More to the point, approach-avoidance theories, constructs, principles, and phenomena (such as the congruency effect) permeate most all of the major and/or burgeoning topics in I/O. Examples include job satisfaction (Ferris et al., 2013) workplace health promotion (e.g. see Malik, Blake, & Suggs, 2014; also see prospect theory, Tversky & Kahneman, 1981), OCB (e.g. Schusterschitz, Stummer, & Geser, 2014), leadership and leader behaviors (e.g. Jackson, Hobman, Jimmieson, & Martin, 2009; Kark & Dijk, 2007; Nifadkar, Tsui, & Ashforth, 2012), and, of course, goal setting and other motivational constructs (e.g. Elliot & Sheldon, 1997; Gist, 1987; Hirst, Knippenberg, & Zhou, 2009; Kanfer & Ackerman, 2000). More than that, I/O survey research often measures the personality and motivation constructs that predict IER, may predict IER given their place in the AAT framework, and on which a sample may become biased due to IER prevention, detection, and deletion (such as the Big Five, approach-avoidance temperament, positive/negative affectivity, and CSE). Therefore, I rely heavily on Elliot and Thrash’s (2002, 2010) AAT theories.

The present research endeavors to assess the effectiveness of screening features and warning messages to prevent IER on Mturk. It differs from previous research and contributes to the literature in several ways. First, I extend previous, preventive, screening research by comparing screened participants to the general Mturk population as
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

opposed to turkers above and below screening cut-offs as in previous research. Second, this study extends screening research by assessing the effectiveness of screening on longer surveys, which organizational research commonly uses. Third, I scrutinize the effectiveness of using the screening approach by detecting IER using three different IER protocols that address criticisms of previous screening research, different aspects of inattentiveness, and specific concerns regarding IER on Mturk. 1) I use instructed responses such as *please select always* that survey research on Mturk typically uses but that turkers might easily identify. 2) I use response times, which addresses the concern that turkers might scan for detection protocols but speed through surveys. 3) I use the infrequency scale, which contains multiple items specifically designed and validated for organizational survey research. In so doing, I evaluate the effectiveness and validity of these detection methods. I compare their relative risks and extend the validity and utility data on *spi* as a detection protocol by assessing how it performs as a page-scale level control and relative to deeper principles of IER (described in later sections). Fourth, drawing from approach-avoidance motivation theories, I extend research on warning messages by testing them in Mturk and assessing the relative effectiveness of different message framings (none, gain-, loss-, and combined-framed), which differs from previous research and general practice that only frame preventive messages negatively. In this regard, I also seek to replicate the congruency effect and understand the Mturk population in terms of psychological phenomena and survey research of direct interest and common measurement in the I/O field [such as approach-avoidance motivation and personal characteristics—Big Five personality, positive affectivity (PA), negative affectivity (NA), approach (AP), avoidance (AV), and CSE]. This differs from previous research that
focused on cognitive phenomena in solely experimental situations. Fifth, I add to research on approach-avoidance theories by testing whether the congruency effects apply to other constructs (PA, NA, and CSE), which differs from recent research suggesting how these constructs fit into the approach-avoidance framework based on how they commonly load. Finally, in an effort to address the dilemma of removing or preventing IER versus biasing samples or measurement, I examine the potentially biasing effects of screening and warning messages by evaluating psychological profiles (e.g. personality) between conditions.
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

Background

Mturk for Organizational Research

Like other fields, the use of Mturk in industrial and organizational psychology grows increasingly popular each year, and I/O scholars study an array of topics in this platform that make considerations for IER important. Harms and DeSimone (2015) conducted a review of methodological issues pertinent to Mturk and organizational literature. They found that, since Mturk’s inception in 2005, Google Scholar search results for Mturk had increased from zero in 2004 to 4000 in 2013 alone. Although few works appeared in top I/O publications at the time, their refined search showed an increase from zero to 100 for each of job satisfaction and job performance and zero to approximately 350 for leadership. In an updated review, Cheung and colleagues (2016) searched the top 20 I/O journals. They found that studies using Mturk samples increased from seven, in 2012, to 44 in 2015; for instance, Journal of Applied Psychology published two, four, and six Mturk studies in 2013, 2014, and 2015, respectively. A sample of topics studied includes coping, the effect of personality on resume evaluations, functioning of personality scales, NA and aggression, trait desirability in leaders, employee voice, justice/fairness/other motivation, self-efficacy and performance, and stress/strain (see supplementary materials in Cheung et al., 2016). What is more, 36.4% of studies reported removing data due to contamination such as failing attention check items, and only 21.2% reported using IER detection; a minority of studies, 3%, reported using Mturk screening features. The authors go on to make recommendations for conducting Mturk research given multiple, potential, methodological considerations—concluding that design, collection, and analytic strategies define data quality on Mturk.
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

not the data source itself. These data and practices carry significant implications for my work.

Specific to IER prevention and other considerations, collection of data on Mturk may influence results and conclusions one might draw. Although the majority of organizational studies using Mturk in the top 20 publications used turkers in conjunction with or to supplement other samples, 12.1% relied on Mturk only; common research purposes included scale validation and generalizability of findings (Cheung et al., 2016). IER considerations permeate all of these practices. For example, and not to isolate a given study, DuVernet, Wright, Meade, Coughlin, and Kantrowitz (2014) assessed the differential functioning of personality items given general mental ability (GMA); they used Mturk participants to assess the generalizability of their findings from health care applicants to a broader population. They concluded that personality items function differently given differences in GMA, Mturk samples provided further evidence of their results, and they inferred that differential item functioning likely stems from information processing of survey content. However, comparing their correlations and focal results between studies reveals attenuation and inflation of correlations, changes in significance, reduction of means, and different sets of items between samples that functioned differently across GMA (all of these results can indicate IER). They did not report detection, deletion, or prevention of IER (Cheung et al., 2016). Therefore, it remains unclear whether their results (given similarities and/or differences) indeed generalize, support their conclusions, or represent artifacts given IER. Similarly, Cao, Drasgow, and Cho (2015) investigated the psychometric properties of personality items using Mturk to generalize results. They tracked IER in students and Mturk and deleted inattentive
participants. However, they used instructed response items (e.g. please select always),
and even though the Mturk survey (94 items) was longer than the student survey (68
items), only 1.7% of Mturk participants were deleted compared to 5.7% of students. This
is consistent with other evidence suggesting that turkers easily identify certain types of
IER detection. Moreover, although deleting 1.7% of the data is unlikely to create a
specific profile of participants, other organizational researchers, using a variety of IER
detection, reported deleting much more—for instance, 9% (Feitosa, Joseph, & Newman,
2015), 18% (Mawritz, Greenbaum, Butts, & Graham, 2016), 26% (Cho & Allen, 2012),
and 31% (Burns, Christiansen, Morris, Periard, & Coaster, 2014). I, specific to IER
prevention and considerations, echo previous recommendations that organizational
researchers practice evidence-based, data-quality control.

**Insufficient Effort Responding**

As noted above, IER represents a respondent’s random and/or intentional lack of
regard for survey content (Huang, Bowling, et al., 2015; Huang et al., 2012). For
example, survey respondents might intentionally ignore survey items and answer at
random or in a specific pattern, or they might inadvertently key an unintended response
due to a random lapse in attention or unintentionally misinterpret a question due to a
random distraction. Thus, Huang and colleagues (2012) coined the IER label and
definition to propose a comprehensive conceptualization of invalid, problematic, survey
responses that appropriately subsumes other conceptualizations. Other conceptualizations
include random responding—responses at random from unmotivated participants, content
non-responsivity—lack of reading, understanding, and accurate/truthful responding, and
careless responding—occasional inaccuracies due to inattentiveness (Huang et al., 2012;
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

Meade & Craig, 2012). Researchers suggest that different conceptualizations of IER and their measurement capture different aspects of IER best (DeSimone, Harms, & DeSimone, 2015). Meade and Craig (2012) supported this notion empirically; they correlated 16 indicators of IER finding small to moderate effects suggesting that detection methods minimally overlap in what they detect; nevertheless, EFA, latent profile analysis, and factor mixture modelling revealed that participants fit best into a two class solution (valid responders and invalid responders). Therefore, regardless of whether an invalid response stems from intentional or unintentional reasons, random or non-random causes, or, random or non-random response patterns, a comprehensive definition proves appropriate because it all represents either valid or non-valid responses/respondents. Furthermore, IER, as a comprehensive conceptualization, yields predictable effects, thus, extending what we know about the influence of invalid problematic responses by defining conditions under which they introduce a source of systematic as opposed to random error.

**Confounding effects of IER.** In its various forms, IER has been an issue in research for decades (see Buechly & Ball, 1952; McGrath, Mitchell, Kim, & Hough, 2010; Spearman, 1904). Although some researchers might argue, and some results might suggest, that IER detection and/or removing IER from data provides little benefit (e.g. McGrath et al., 2010; Dunn & Heggestad, 2017), I hold that this depends on how one defines the problem and conceptualizes and measures IER. Undeniably, other research/researchers support this notion. Research shows that IER can and does influence results, and lower levels, or removal, of IER diminishes such a concern. For example, Schmitt and Stults (1985) noted that negatively keyed items often result in a
single factor on which researchers make substantive inferences. They posited that these factors actually arise from inattentiveness. They tested this hypothesis using a simulation study showing that when inattentiveness exceeded 10% of responses, negatively keyed items loaded on a single factor, thus, concluding that IER may invalidate factorial structures. Woods (2006) found similar results. Likewise, we know that random sources of variance attenuate associations between two measures (Crocker & Algina, 2006; Spearman, 1904). However, McGrath and others (2010) reviewed 41 studies in applied assessment pertinent to invalid responding and “real-world, criterion-related validity”.

That is, they reviewed studies that tested whether indicators of random or bias (equivalent to intentional IER) sources of variance revealed, as theorized, attenuation or moderation of relationships between substantive scales of personality and outcome criteria. They found that only a small set of reviewed studies confirmed that random or biased responding accounted for attenuating or moderating effects of invalid responses; correspondingly, they found weak support for using indicators like those representing IER to clean relationships. However, they suggested that poor validity of indicators and the adequacy of criteria might explain the lack of robust, consistent findings.

With updated concepts, and better IER indicators now available, Huang and others (2015) presented a different possibility—specific conditions under which IER results in systematic as opposed to random error. They postulated that when actual scale means (the resulting variable scores) deviate from the scale midpoints (e.g. 3 on a 1 to 5 scale; which often occurs), IER influences the mean by pulling it closer to the mid-point. This occurs because regardless of the type, why, or how respondents respond invalidly, some will answer high, some will answer low, and others will answer in the middle of the
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

They held that under the condition of scale-midpoint-deviation, on two survey scales, IER could inflate bivariate relationships and increase the probability of type I errors. They conducted two studies using student and simulation data and confirmed that IER responses gather near scale midpoints. They further found that 1) IER inflates bivariate relationships 2) this influence approached zero as scale means also approached midpoints, and 3) as IER prevalence decreased so did these insidious effects. These results can account for the lack of insidious/consistent findings in previous research if bivariate relationships were increasing (vs. decreasing) and that others found no benefit of removing IER since no biasing effects would exist if scale means showed little to no deviation from scale midpoints and/or IER is low. More specifically, under Huang and colleagues proposed conditions, IER can cause systematic as opposed to random error.

Drawing on this work, McGonagle and others (2016) examined commonly measured work and occupational health psychology variables having scores that tend to deviate from scale midpoints—such as perceived work ability, negative affectivity, general self-efficacy, and incivility. Among Mturk, work/field, and student samples, they found similar results as Huang and colleagues (2015), additionally, noting the significance of changes in relationships between full samples and attentive subsamples and in more commonly used statistics such as multiple regression. For instance, the correlation between work ability and negative affect was significant when a sample contained IER and non-significant when it did not, and the relationship between supervisors’ experienced incivility and instigated incivility significantly changed from .61 to .43. These effects of IER still occurred between some relationships when IER
IER prevalence equaled only 6% of the data. It certainly appears that IER presents an
insidious confound to survey data.

**IER detection.** This section seeks to discuss IER detection methods pertinent to
IER prevention on Mturk that this research highlights. Previous research, as noted above
(e.g. DeSimone et al., 2015; Huang, Bowling, et al., 2015; Huang et al., 2012; Meade &
Craig, 2012), point to the idea that different detection methods and/or conceptualizations
of IER capture different aspects of IER. However, no research investigates this idea on
specifics such as how different aspects of IER relate to the influence of personality on
IER or contexts in which different aspects of IER present greater or lesser concern. First,
a central concern to this study stems from criticisms of the techniques used in previous
studies that examined the preventive benefits of screening. That is, turkers may easily
identify instructed response items such as please select always embedded in scales, or,
*only select [blank] from the list below* embedded in instructions such that Mturk
participants may be familiar with these forms of detection and scan for them versus
diligently (or sufficiently) responding to survey items (Peer et al., 2014). Although
Mturk populations refresh over time and researchers can control and prevent issues
involving non-naiveté, Chandler, Mueller, and Paolacci (2014) suggested that turkers do
learn from previous exposure to tasks. Indeed, regarding IER detection, Hauser and
Schwarz (2016) compared easily identifiable instructed responses (commonly used IMC)
to novel instructed responses (structurally altered IMC) among screened turkers and
found marked differences in IER pass rates (95-96% versus 26%) though these
performances still outshined student samples. Likewise, Blanchard and Banerji (2015)
found differential functioning of easily identifiable detection between samples.
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

Second, some detection methods may capture the comprehensive nature of IER better making them more useful for research purposes (i.e. analysis can account for more IER and better support results). Of detection methods used by Meade and Craig (2012) leading to classifications of valid vs invalid responders, Huang and others (2015) revealed a clear, one-factor CFA that they dubbed IER, and on which the infrequency method loaded best. Thus the infrequency method may be the most comprehensive and explanatory method of IER detection. This approach includes post-hoc detection statistics such as individual reliability and Mahalanobus $D$ as well as a priori detection items such as psychological antonyms/synonyms or infrequency items (sometimes called bogus items; see scholarly works noted above for descriptions); more than that, infrequency detection methods may prove less learnable or identifiable by turkers. Ostensibly, turkers cannot learn to scan for post-hoc statistical measurements; similarly, turkers may encounter difficulty scanning items for parallel/opposite forms or other more naturally worded, substantive, survey items. That is, please select always may present itself to observation and memory more readily than *I work 14 months in a year* embedded in an organizational psychology survey. Bogus items such as *I am paid bi-weekly by leprechauns* received scrutiny for the reactions they might yield among participants—such as responding differently due to humorous reactions or negative reactions such as finding the survey/research pointless (Harms & DeSimone, 2015; Huang, Bowling, et al., 2015; Meade & Craig, 2012). Consequently, Huang and others (2015) compared participant reactions (enjoyment, ease of responding, and intent to participate in future surveys) in an Mturk sample; they found no significant differences in reactions between those that took a standard survey versus those that took the same survey embedded with
infrequency items. Additionally, researchers may be able to tailor infrequency scales for content and number of items and still reliably and validly detect IER (Huang, Bowling, et al., 2015; McGonagle et al., 2016), and they pose less difficulty when coding and defining cut-offs (i.e. what constituted inattentiveness).

Response times may prove a bit cumbersome when defining cut-offs (i.e. how fast/slow depicts IER), but they maintain the benefits of unidentifiability and easy calculation (for example Qualtrics measures time per page, total survey time, and time until first click directly), and they address pertinent concerns for Mturk research. For instance, some turkers may automate responses to fill out surveys and earn more money, or, more importantly, scan for attention checks but speed through substantive survey items. Response times per page that are too fast show convergent validity with similar indices such as individual reliability and long-string responses, and at 95% specificity, they produce greater sensitivity than either (Huang et al., 2012). However, the definition of response times relative to IER plays a significant role in their utility. Turkers might save a page then come back (increasing survey time), take a break with a page open (increasing time per page), which says little about their attentiveness while responding. As well, they may get distracted and fail an IER item (or substantive items), read more carefully and consider item content to a greater degree, read faster than others answering validly (or invalidly for that matter), or scan for attention checks and hurry through other items. All of these possibilities speak to survey time but may have different effects regarding response validity. Indeed, research using broadly defined response times such as time-per-page or total survey time finds little criterion or construct validity (Blackmore, 2014; Huang et al., 2012; Meade & Craig, 2012). In contrast, Huang and
colleagues (2012) showed initial evidence for the utility of response times by defining IER as response times that are too fast to reflect regard (reading/thought, etc.) for survey content using their judgment of two seconds per item and 52 seconds per page, on their survey. Wood and others, (in press) empirically derived cut-points using an index for seconds per item \((spi)\). They found a sharp drop in data quality at \(spi = 1\) such that those below this cut-off strongly affected consistency within scales and subsequent relationships between the constructs the scales measured. What is more, they found no such effect from all response times above it. Therefore, to address issues with response time, the current research adopts the cut-point approach and defines IER relative to response times as those completing items faster than one could while reading and considering content.

**IER Prevention**

Given the apparent need for IER prevention, researchers have investigated several techniques. Many proved ineffective. Yentes (2015) tested the preventive effects of survey length disclosure and the presence of a survey progress bar on Mturk and found that these methods did not significantly reduce IER. Likewise, Young (2016), proposed that the activation of subconscious, achievement goals by priming participants with a word task or a picture might work. These methods did not significantly reduce IER. Similarly, Bragg, Bowling, and Wynne (2017) attempted to reduce IER through psychological mechanisms by showing video appeals to students before a survey. In their present form, these appeals failed to reduce IER. Finally, Ward and Pond (2015), tried to reduce IER on internet-based surveys in a student sample by using a variety of virtual presence devices (e.g. an animated shape or a video file of a simulated person). The
virtual presence device alone unsuccessfully reduced IER, but did so significantly when paired with a warning message. Although most of the results above proved non-significant, the researchers propose that some approaches may better suite one context than another (e.g. Mturk surveys vs student participants or job assessments). Therefore, refining and devising appropriate prevention strategies represents a significant endeavor; moreover, some methods currently show promising results.

**Mturk screening.** Researchers can employ multiple screening methods and criteria to ensure data quality on Mturk related to given concerns (see discussion and scholars noted above). This section discusses a particularly effective tool for preventing IER and a particular problem it presents. In this regard, screening refers to turker HIT statistics. HIT acceptance (or approval) rate represents the percent of HITs (such as data coding, file sorting, or survey research) that a turker completes and for which a requester deems sufficient and remits payment. Requester refers to anyone posting a HIT on Mturk (researchers, employers, and the like). HIT experience depicts the number of tasks that a turker has completed. To stay consistent with the literature, from this point on, I refer to approval rate and experience as reputation and productivity respectively. Requesters can set their HIT to allow participation from only turkers having specific reputation/productivity. Peer and colleagues (2014) manipulated these criteria and isolated their effects on IER prevalence in surveys. In one condition, they set screening criteria to those above 95%; in the other, they set criteria to those below 90%. They found that only 2.6% of those above 95% acceptance failed at least one attention check, whereas, 33.9% of the latter group failed at least one. A similar and significant, albeit a weaker, effect stemmed from productivity, 19.7% compared to 28.1% IER, respectfully.
Notwithstanding the research question posed in later sections, I expect broadly similar effects (lower IER prevalence).

_Hypothesis 1:_ Prevalence of IER will be lower in a screened than in a non-screened Mturk sample.

**Drawbacks of screening.** However, recent research concerning stable antecedents of IER, evoke an interesting question. Camus (2015) tested the temporal stability of IER by measuring it at different time points and in different settings using students (online and in the lab). Using multiple IER detection methods, the author found evidence of temporal stability in both settings, $r = .55$ and $r = .43$, across two time points. Explicitly, the students committing IER at one time point likely did so at another. Certainly, Bowling and co-authors (2016) found rank-ordered, situational stability—those committing the most IER committed the most across situations. They also found that personal characteristics, rated by informants, predicted IER—those lower in conscientiousness, extraversion, agreeableness, emotional stability, GPA, and class attendance committed more IER. It may prove that those having greater reputation and productivity appear more attentive because they have specific characteristics. In addition, Peer and colleagues’ IER preventive screening research used relatively short surveys (demographics plus approximately 50 items). In reality, survey research often involves multiple studies embedded in the same survey, and, sometimes, researchers omit the entire length of the survey from their methods (often >100 items). Yentes (2015) varied survey length (approximately 209 items versus 35 items) on Mturk to test its effect on IER as currently conceptualized, the author found moderate effects suggesting that those taking longer survey’s pay less attention to survey content. In sum, screening in
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

Practice may reduce IER to lesser extent than previous research, due to the use of longer
surveys in organizational research and the exacerbating effect of survey length on IER.
Since the prevalence of IER carries measurement implications, understanding the utility
of screening for reducing IER in actual practice carries importance. That is, if screening
limits the available population yet IER remains above negligible levels, requiring further
deletion, then screening loses its practical utility, especially, if screening and post-hoc
data quality control creates sampling bias. However, to the best of my knowledge, no
research exists that substantiates these assumptions or defines what levels of IER
prevention/deletion might result in sampling bias. Therefore, I pose the following
research question:

Research Question 1: Controlling for IER, will the screened sample have different
psychological profiles (e.g. differences in personality) from the non-screened sample?

Preventive survey instructions. The Mturk platform lends itself well to the use
of warning messages. Requesters can use the system to give bonuses for high quality
responses or reject data from low quality responses, and this can be automated (Berinsky
et al., 2012). However, empirical evidence is lacking for the effectiveness of these tools
relative to IER. In regard to gains, bonuses may increase attentiveness to tasks and this
effect may be independent of bonus amount or structure (D. Chandler & Horton, 2011; J.
Chandler, Paolacci, & Mueller, 2013), but the literature on gains is lacking for survey
research and IER. Concerning losses, researchers commonly use phrases such as if you
do not pay attention, your HIT will be rejected and you will not be paid. However,
studies using these messages focus elsewhere rather than IER prevention and are not
informed by updates on what makes IER preventive messages effective.
A handful of research has refined IER-preventive warning messages into an effective tool in student populations. Huang and others (2012) and Meade and Craig (2012) both tested warnings, the former found significant IER reduction of small to moderate effects on multiple, IER, detection scales; the latter found non-significant results on those same IER indicators. Blackmore (2014) reasoned that this discrepancy stems from the disparate composition of their messages. Drawing on Vroom’s 1964 expectancy theory, the author theorized that for a warning to motivate attentive behavior it must make IER evaluation salient and relate inattentive behavior to a valued consequence. The message used in Huang and others (2012) had these characteristics but the one used in Meade and Craig (2012) did not. Still, the former message used the term “advanced statistical techniques” to make IER detection salient, which may not have been intuitive for students. To improve on Huang and others’ (2012) method, Blackmore (2014) told participants that IER would be assessed using a quiz and that students might not receive credit if they did not pass the quiz. Blackmore (2014) found significant differences between warning and control for multiple, IER, detection methods all at or slightly above moderate effect size. Other researchers, Ward and Pond (2015), adopted the Huang et al., (2012) “advanced statistical technique” portion of the message but did not provide valued consequences and found small effects for only one, unrelated, IER indicator. This string of research suggests that warning messages effectively reduce IER if constructed well.

Motivation and framing. Approach-avoidance motivation theories root in the hedonic principle that people are motivated to approach pleasure and avoid pain (Higgins, 1997). For some time scholars have sought to understand the nature of this
principle and how it manifests in motivational, emotional, and behavioral strategies and processes (for a good review, see Carver et al., 2000). Researchers in neuropsychology and neuroscience posit that approach and avoidance motivational tendencies arise from two, fundamental, neural systems—the behavioral activation system (BAS) and the behavioral inhibition system (BIS; Gray, 1987). This research holds that the BAS responds to incentives, such as signals of reward, and that activity in the BAS causes people to move toward goals; in contrast, the BIS responds to threats, such as signals of punishment, and activity in the BIS inhibits movement toward goals (Carver et al., 2000). More than that, such activity mediates the relationships between positive or negative emotions/cognitions and experiences, such as the facial expressions of others or rewarding or punishing stimuli (Gray, 1990; for a good review see Lang & Bradley, 2010). Neuroimaging and brain lesion studies confirm these functions and tendencies (e.g. Banks, Eddy, Angstadt, Nathan, & Phan, 2007; Costafreda, Brammer, David, & Fu, 2008; Gee et al., 2013; Schlund, Magee, & Hudgins, 2011; Spielberg et al., 2011; Yang et al., 2002). Moreover, people vary on approach-avoidance characteristics that correspond to said functions (Davidson & Irwin, 1999; Gee et al., 2013; Spielberg et al., 2011).

Seeking to move beyond basic hedonic principles, Higgins (1997) championed the theory of regulatory focus. The author organizes the basic approach and avoidance principles into promotion and prevention focus of goal attainment. These foci represent both a state and chronic sensitivity to positive or negative outcomes such that desired end-states motivate individuals to approach matches or avoid mismatches between current and desired end-states as a strategic means. Particularly pertinent to the present research, a tenant of this theory, strategic compatibility, suggests that individuals perceive
incentives as more relevant to goal attainment when they are compatible with their strategic inclination (promotion and/or prevention). Shah, Higgins, and Friedman (1998) tested this idea showing that promotion-framed incentives to gain points on an anagram task and prevention-framed incentives to avoid losing points improved performance better for those having a chronic promotion or prevention regulatory focus, respectively.

However, perhaps as long as the history of the hedonic principle, philosophers and scholars tended to conceptualize and study approach and avoidance as a unidimensional bipolar construct placing positive/good valences on one end and negative/bad valences on the other (Cacioppo, Gardner, & Berntson, 1999). Cacioppo and Berntson, (1994) and Cacioppo and colleagues, (1999) reviewed theories and empirical evidence and argued that form follows function and that human neural anatomy and physiology maintain both parallel and integrative properties such that approach and avoidance are both distinct as well as integrative (i.e. multi-dimensional). That is, people can be high or low on approach, avoidance, and any combination of the two. Not only that, both can activate simultaneously having additive, independent, as well as integrative influence on motivation, emotion, attitude, and behavior (Cacioppo & Berntson, 1994; Cacioppo, Gardner, & Berntson, 1997; Cacioppo et al., 1999; Koopmann, Lanaj, Bono, & Campana, 2016; Lanaj, Chang, & Johnson, 2012). This has become the accepted view among approach-avoidance scholars.

Although the few studies investigating warnings affords auspicious efficacy, they all employ a loss-framed design. Consistent with the tenants of approach-avoidance motivation, research in advertising, compliance, and health promotion show that message framing carries important implications for motivating behavior. Results consistently
show that people are more responsive to a cue and information is more salient when it is congruent with a person’s approach or avoidance tendencies (e.g. Hamamura, Meijer, Heine, Kamaya, & Hori, 2009; Sherman, Mann, & Updegraff, 2006). For instance, Sherman and colleagues (2006) showed that gain-framed and loss-framed messages more effectively increased flossing behavior for those with a higher AP and AV, respectively. Moreover, combined-framed messages motivate behavior to greater extent than gain-framed and loss-framed messages (Slater, 2014; Treiber, 1986; Wilson, Purdon, & Wallston, 1988; Wilson, Wallston, & King, 1990). Treiber (1986) showed that combined-framed messages increased infant car restraint use more so than loss-framed messages. Similarly, Wilson and others (1990) showed that combined-framed messages significantly reduced smoking behavior more so than gain-framed messages. The reasoning for this rests in the idea of moving toward end-states. When both AP and AV tendencies activate in an individual, avoidance increases faster as one progresses toward a goal. However, AP also increases, and, in spite the slower increase in approach activation when moving toward the desired end-state, it starts out stronger (Cacioppo & Berntson, 1994). Put simply, approach and avoidance tendencies jointly influence action in an additive fashion as individuals move toward end-states. Thus, since environmental contingencies (such as combined-framed messages) simultaneously activate approach and avoidance tendencies, combining frames may maximize the utility of warning messages by decreasing IER to greater extent.

**Hypothesis 2:** Prevalence of IER will be lower in the (a) gain-, (b) loss-, and (c) combined-framed warning conditions compared to control.

**Hypothesis 3:** Combined-framed messages will more effectively reduce IER than (a) gain and (b) loss-framed messages.
Nomological, motivation moderators. In defining work to integrate various perspectives of personality, affectivity, and motivation into a unifying framework, Elliot and Thrash (2002) argued that relevant personality traits that align with positive and negative valences serve as indicators of approach or avoidance temperament. They further argued that although this temperament provides the initial impetus for behavior, indicators more proximal to attitudes and behaviors, such as approach-avoidance motivations, channel that influence. In their study, they found that neuroticism, negative temperament, and the BIS loaded on a single, latent, avoidance factor; furthermore, they found that extraversion, positive temperament, and the BAS loaded on a single, latent, approach factor. Their full structural equation model showed that these indicators related to self-imposed approach and avoidance goals in the expected directions. Investigators in many fields began to theorize how other psychological constructs fit in an approach-avoidance temperament framework (e.g. Corr, DeYoung, & McNaughton, 2013; Elliot & Thrash, 2010; Ferris et al., 2011; Johnson, Rosen, & Levy, 2008; Lanaj et al., 2012).

Drawing on Elliot and Thrash’s (2002) work, Ferris and colleagues (2011) sought to resolve inconsistencies involving theoretical assumptions about core self-evaluations as indicative of approach or avoidance temperament and integrate CSE into an approach-avoidance framework. CSE represents a compound, broad, latent, higher-order trait that comprises the general appraisal of one’s worth, effectiveness, and capabilities and subsumes four personality traits—self-esteem, generalized self-efficacy, neuroticism, and locus of control (Judge, Erez, Bono, & Thoresen, 2003; Judge & Locke, 1997). Self-esteem depicts the value people ascribe to themselves; generalized self-efficacy entails the belief people have about their ability to perform across various situations; neuroticism
indicates the tendency to focus on negative aspects of the self and form negative cognitions; and, locus of control represents whether a person attributes the cause of life events to their own behavior (internal) or other factors (external; Judge et al., 2003). Thus, researchers considered whether CSE indicated approach and/or avoidance temperament (Chang, Ferris, Johnson, Rosen, & Tan, 2012). Ferris and others (2011) showed that CSE serves as an indicator of both, loading positively on approach temperament and negatively on avoidance temperament. They also tested a PATH model showing that CSE, work approach-avoidance motivation, and behavioral outcomes (e.g. citizenship behaviors and performance) fit best when work approach-avoidance motivation mediated the relationship between CSE and behavioral outcomes. This suggests that CSE fits the approach-avoidance framework well as an indicator of approach-avoidance temperament. Given the findings above and research relative to CSE, PA, and NA within an approach-avoidance framework, I make the following hypotheses:

**Hypothesis 4:** Motivational disposition will moderate the relationship between warning messages and IER such that gain-framed messages will most effectively reduce IER for those higher in (a) AP (b) CSE and (c) PA, and loss-framed messages will most effectively reduce IER for those higher in (d) AV (e) NA and lower in (f) CSE.

**Warning message drawbacks.** In spite the intended effect of IER prevention, warning messages may have unintended effects that negate their usefulness. Meade and Craig (2012) showed that punitive warning messages decreased student attitude toward a survey. Ward and Pond (2015) suggested that more positive prevention methods (relative to vague threats) might prove preferable. Other scholars corroborate this idea. For instance, Bragg and others (2017) suggested that warning messages might suit some
contexts better than others. For example, pre-employment or other purposed surveys among employees may preclude giving bonuses following a gain-framed warning message, and negative reactions to loss-framed messages might sour application processes or employment relationships (Liu, King, & Lustenberger, 2017). However, since Mturk survey-research occurs among a constantly changing population and small increments of payment (or loss of payment) represent normal logistics and practice, warning messages appear appropriate. Nevertheless, concerns remain that negative reactions to warning messages might change response patterns and congruency between framing and personal characteristics might select in those having specific characteristics by reducing IER among specific types of people, leaving data from other types for deletion.

More than that, warning messages themselves carry the potential to alter mood or cognitive states. Indeed, latent state-trait theory (LST; Steyer, Schmitt, & Eid, 1999) holds that survey’s don’t “take place in a situational vacuum”; the person, situation, and interactions between them at the point of measurement yield important sources of variance in psychological measurement. As such, a state represents a trait and a situational/interaction residual. Therefore, if participants respond to survey items differently when they encounter warning messages than if they encountered no warnings, the research results based on those participants’ responses, in their psychological state, may be biased by the systematic error attributable to the situational/interaction residual. Consistent with this, Koopmann and colleagues (2016) showed that daily positive and negative events influenced promotion and prevention foci levels, respectively, which influenced positive and negative mood-states. However, when approach and avoidance
systems activated jointly through occurrence of positive and negative events, the authors found an integrative effect. That is, mood-states became neither very positive nor very negative compared to when one system activated more strongly than another did. Therefore, combined-framed messages may further maximize the utility of warning messages, beyond exerting additive effects on IER reduction, by exerting integrative effects on psychological states, thus, minimizing potentially biasing effects on relevant personality and/or temperament constructs. Nevertheless, both foci increased, and empirical evidence is lacking for what degree of positive or negative events, such as a warning before a survey, may result in biasing effects. Given that warning messages potentially create sampling bias and measurement error, I ask the following research question:

*Research Question 2:* Controlling for IER between framing conditions, will warning messages result in samples having different psychological profiles (e.g. differences in personality) relative to the control group?
Participants and Sampling

I collected data by joining an existing research project. My colleagues and I invited Mturk workers to participate. The existing research protocol required full-time employment, defined as at least 35 hours per week at their primary job for at least six months. My colleagues used Mturk’s qualification system to make the study available to qualified participants. For approximately half of the present sample, the existing protocol used the additional criteria that turker reputation met or exceeded 98% and their productivity exceeded 5000 accepted HIT. My colleagues and I drew this portion of the sample from a pool of 1500 turkers who participated in a previous study on occupational health and job conditions hosted by said colleagues. I invited the other half from the general Mturk pool. The institutional review board at Florida Institute of Technology approved this research.

Data collection began with the latter half of participants and ended after 160 respondents. I used this subset to pilot test the experimental manipulation. Data collection resumed after the manipulation appeared functional based on comparisons of IER prevalence and manipulation checks. Once sampling in the general population completed, my colleague invited the screened participants. The final sample consisted of 1071 turkers—472 from the screened collection and 584 from the non-screened collection. Of the total, males comprised 49%. Participants received $1.00 in compensation.

Power. I calculated the necessary sample for a logistic regression and a factorial ANCOVA. I used logistic regression for the main hypotheses comparing screening
condition groups, warning condition groups and the interaction between screening and warnings. This yields four independent variables and three interactions. Peduzzi, Concato, Kemper, Holford, and Feinstein (1996) provide guidelines for determining sample size in logistic regression derived from a simulation study using the equation $N = 10k/p$ where $k$ is the number of predictors and $p$ is the lowest probability that the outcome will equal one. I set $p$ conservatively at .1 to reflect the midpoint of IER prevalence within screened Mturk samples. Thus $(10 \times 7)/.1 = 700$. Although updated computations allow for the account of interactions between binary predictors (see Demidenko, 2007, 2008), this method currently allows for only two binary predictors and one interaction. Therefore, I conducted a second power analysis using the more conservative estimation of the required sample from a 4x2 factorial ANCOVA with one covariate used for Research Questions 1-2. The effect size of screening in previous studies was quite large (i.e. Peer et al., 2014), so I derived the needed sample size from the small to moderate effects of warning messages (Blackmore, 2014; Huang et al., 2012) and congruency between warning messages and personal characteristics (e.g. Sherman et al., 2006). I converted the effects sizes from $d$ and $r$ to $f$ using computation software at [https://www.psychometrica.de/effect_size.html](https://www.psychometrica.de/effect_size.html). Using small effect sizes $d = .2$ and $r = .1$, the conversion yielded $f = .1$. I used G-power 3.1 to calculate the needed sample size. I set power at 0.8 as suggested by Cohen (1962). Using $f = .1$, three degrees of freedom, and eight groups, the required sample size equaled 1095, which is close to the actual sample size I have (1071).

**Measures**
State approach-avoidance. I measured this construct using a six-item self-regulatory-foci scale adapted from a work specific measure developed by Lin and Johnson (in press). An original item reads *at work I am focused on preventing negative outcomes*. An adapted item reads *right now I am focused on preventing negative outcomes*. Participants responded on a 1-5 agreement scale, 1 = “strongly disagree” and 5 = “strongly agree”.

Approach-avoidance temperament. I measured AAT using the Approach-Avoidance Temperament Questionnaire (ATQ) developed by Elliot and Thrash (2010). It uses a 1-7, Likert-type agreement scale, 1 = “strongly disagree” and 7 = “strongly agree”. This scale contains 12 items, six each for approach and avoidance. Higher scores correspond to greater trait level. Example items are *thinking about things I really want energizes me* and *when it looks like something bad can happen, I feel a strong urge to escape*. See Table 1 for all reliabilities on all focal survey scales.

IER. I measured IER using three methods. The Infrequency scale (Huang et al., 2012) uses eight items such as *I can run two miles in two minutes*. I recorded responses on a Likert-type agreement scale from 1-7, 1 = “strongly disagree” and 7 = “strongly agree”. Consistent with previous research (Huang, Bowling, et al., 2015; Huang et al., 2012; McGonagle et al., 2016), I defined failing an item as any response on a scale that indicated agreement with the item. I developed an instructed response scale using the criteria described by multiple researchers (Blackmore, 2014, Chandler et al., 2014; Huang et al., 2015b). I created three items that instructed participants to select a specific response. I placed the correct answer at three locations between 1 and 7 on Likert-type scales—the low end, the middle, and the high end. An example item is *please select*
strongly disagree. I defined failing an item on this scale as any response other than that instructed. I embedded infrequency and instructed response items throughout the survey at random. Finally, I measured response times using Qualtrics’ timer features. I calculated an spi index using equation 1, and I defined IER as spi cut-off $< 1$ where $T_{CS} =$ time to submit on a page-scale, $T_{C1} =$ time to make first click on a page-scale, and $N_R =$ the number of items on that page-scale, and spi cut-off scoring $= \log_2 spi = 0$ (Wood et al., in press).

$$spi = \frac{T_{CS} - T_{C1}}{N_R - 1}$$ (1)

IER operationalization. To be consistent with previous work and provide an intuitive understanding of IER prevalence and its differences between conditions, I operationalize IER for each index as dichotomous variables coded $0 = \text{no}$ and $1 = \text{IER}$. For each of the infrequency and instructed response scales, passing all items $= 0$ and failing any items $= 1$. For response times, having all page-scale $spi$ above zero $= 0$ and having any $spi$ on a page-scale at or below zero $= \text{IER}$. However, research provides no clear definitions of how many failures truly represent participant level IER but evidence shows that removing participants at different numbers of items failed influences results in the same direction (e.g. removing people failing one item significantly reduced correlations and removing people failing two items significantly reduced correlations further; McGonagle et al., 2016). Therefore, I also operationalize IER as a count variable. I summed all item-level $0 = \text{no}$ and $1 = \text{IER}$ for the instructed response and infrequency scale, thus, for instructed response and infrequency scale count IER, range $= [0, 3]$ and $[0, 8]$, respectively. I did the same for $spi$ for each page-scale, range $= [0, 13]$. 

Therefore, count IER represents the number of items/page-scales failed on a given detection method.

**Personality.** I measured personality using the agreeableness and conscientiousness subscales (one subscale for agreeableness and one for conscientiousness) of the Mini-IPIP (Donnellan, Oswald, Baird, & Lucas, 2006). Each subscale has four items measured on a Likert-type agreement scale from 1-7, 1 = “strongly disagree” and 7 = “strongly agree”. Four of the eight items are reverse coded, two for each subscale. Example items for agreeableness and conscientiousness, respectively, are *I sympathize with others’ feelings* and *I make a mess of things*, reverse coded. Higher scores indicate higher trait levels.

**Positive and negative affectivity.** I used the shortened version of the positive and negative affectivity scale I-PANAS-SF (Thompson, 2007) to measure affectivity. The scale asks participants to indicate the extent of their felt emotions to one-word stems such as *upset* and *inspired*. Responses occur on a scale from 1 = “not at all”, to, 6 = “extremely”.

**Research Design**

This study used quasi-experimental, between-subjects designs. The screening condition contained two natural groups—screened and non-screened—using the defined criteria, but turkers’ HIT characteristics determined these comparison groups. I manipulated the warning message condition into four randomized groups—gain-, loss-, combined-framed, and control. Both conditions employed post-test only measurement.

I used Qualtrics to collect data; participants followed a protected link (only accessible via invitation or Mturk) taking them to a consent form. I informed them that
the study focuses on job conditions and occupational health. I used incomplete disclosure to keep participants blind to the hypotheses and manipulations regarding IER. I set the Qualtrics system to navigate turkers to the first page of the study after signing by clicking agree. I used Qualtrics features to randomize which warnings turkers saw and page order (one scale per page) with the exception that the warning message always came directly after consent, manipulation check always came second, and optional questions about Mturk user habits and demographics always came last just before debriefing.

**Experimental manipulation.** All participants received standard survey instructions such as *you will be presented with instructions* and *please respond honestly.* Experimental groups saw standard instructions plus the following message:

> It is important to our study that participants give their full attention to this survey and respond with sufficient care.

> Please be aware that you will be given a multiple-choice quiz for the last part of the survey. This quiz will assess your knowledge of the survey’s contents to determine if you have payed attention and provided sufficient care.

Along with one or both of the following messages, which I counter-balanced for the combined-framed condition such that half of participants saw gain first and the other half saw loss first:

> If you pass this quiz, you will gain an additional 50 cents in compensation.

> If you fail this quiz, your HIT will be rejected and you will not be paid.

The experimental group had to answer two questions correctly, one about how attention is assessed and the other about the consequence of inattention, before continuing to the first scale.
One might note that these two reward/loss conditions, relative to each other, create a net loss situation when combined and different increments of gain or loss independently. However, nuanced conditions of conducting research on Mturk made this necessary and acceptable. First, the Mturk platform does not allow for decrements in pay, but it does allow for bonuses in pay. Therefore, a decremental message becomes meaningless to turkers; what is more, researchers can reject HITs and tend to use these types of messages, so the full rejection message balances the requirements of Mturk with the practices of research, which in part, as apparent in the title, guided the present research. Second, and as described in previous sections, non-survey research shows that different increments of bonuses carried no effect, only the presence of a bonus seems to matter (D. Chandler & Horton, 2011; J. Chandler et al., 2013), and budgeting constraints given sampling requirements made actually providing a bonus impossible. Thus, this research required a bonus that balanced the relative value of the gain/loss conditions with logistic concerns and potential negative reactions to not receiving a bonus.

**Analytic Strategies**

**Manipulation checks.** To assess whether warning messages functioned as expected, psychologically, I used independent samples $t$-tests to check for differences in state AP (between control and reward) and state AV (between control and loss).

**Hypotheses tests.** For Hypotheses 1-2, I ran hierarchical logistic regressions entering the categorical screening conditions and categorical warning message conditions as independent variables. I entered the screening condition in the first step followed by the warning message condition, and for exploratory purposes, the interaction between screening and warning in the third step. I dummy coded warning message groups for
gain, loss, and combined setting control as reference. I used dichotomized IER as the dependent variable(s). For Hypothesis 3, I ran planned comparisons between the combined-framed condition and each of gain- and loss-framed conditions for each IER index using a chi-squared test.

For Hypothesis 4, I employed hierarchical logistic regressions analyses for each IER detection method. I set the warning message condition as the IV dummy coded 0 = control and 1 = gain (for Hypotheses 4a-c) and 1 = loss (for hypotheses 4d-f). I used dichotomized IER as the DV. I set individual means of (a) AP (b) CSE (c) PA (d) AV (e) CSE and (f) NA as a continuous moderator. I entered the warning condition first followed by the moderator then their interaction.

For Research Questions 1-2, I used 4x2 factorial ANCOVAs. I set the independent variables as screening and warning message conditions and their interaction (for exploration). I set IER count variables as the continuous covariate to address the fact that the outcome variables consist of self-reports of personal characteristics that IER of the respondent likely influenced. I set the dependent variables as the individual mean scores on each of agreeableness, conscientiousness, and trait positive and negative affectivity.

Although no research empirically assessed PA and NA as predictors of IER (as did extraversion and neuroticism), the present research required a balance between the number of scales/items I could add to an existing research project with what I needed to test considering what scales the existing research already included. The existing project included the PANAS. Research (e.g. Elliot & Thrash, 2010 and related works noted above) consistently shows very strong relationships (often above $r = .60$) and/or similar
loadings on higher order constructs between positive affectivity and extraversion as well as negative affectivity and neuroticism. Therefore, I use PA and NA as covarying proxies for extraversion and neuroticism, respectively, as well as a more direct assessment of the influence of positive and negative warning messages on commonly researched, I/O and AAT relevant constructs. Since no meta-analytic evidence supports the notion that the remaining Big Five traits predict IER better than one another (the known effects stem from informant report), I simply chose two more dissimilar Big Five constructs (i.e. agreeableness and conscientiousness as opposed to agreeableness and openness).
Results

Descriptives are reported in Table 1. Raw correlations show that categorical and count IER variables relate to personal characteristics in the expected directions for the vast majority of correlations. These relationships appeared largely similar across IER indices. However, a few relationships relating to dichotomous IER indices differed (e.g. agreeableness related negatively to failing any infrequency items but positively to failing any pages for spii). For count IER, every relationship/correlation was in the correct direction and showed no inconsistencies across indices. In addition, these relationships appeared stronger for count IER (the number of items/page-scales failed) than for dichotomous IER indices. Raw correlations also reveal significant relationship for screening and/or warning conditions with personal characteristics. Finally, raw relationships between similarly valenced AAT constructs (trait PA, AP, and CSE and between trait NA, AV, and CSE) ranged between .52 and .66 and in the expected direction. Only Trait AP and CSE related outside this range at .46.

Preliminary Analysis

Data. I ran confirmatory factor analyses using AMOS 19 for agreeableness and conscientiousness, state and trait AP-AV, PA/NA, CSE, and the IER infrequency scale. Results showed good to acceptable fit for two-factor agreeableness/conscientiousness (RMSEA = .079 [.066, .092], GFI = .974, CFI = .972) two-factor, trait, AP-AV (RMSEA = .067 [.062, .077], GFI = .957, CFI = .962), two-factor PANAS (RMSEA = .058 [.048, .067], GFI = .973, CFI = .978), single-factor CSE (RMSEA = .081 [.073, .090], GFI = .953, CFI = .960), and IER (RMSEA = .057 [.044, .070], GFI = .982, CFI = .983). State
AP-AV results revealed acceptable fit using GFI (.974) and CFI (.971) but poor fit on RMSEA (.097 [.079, .115]).

I removed outliers defined by the ±1.5*IQR (or interquartile range) rule using SPSS 24. This resulted in removing 0-2% of responses (or 0-22 responses) on each scale. Although psychological research tends to use 2*IQR, many disagree on how to address univariate and/or multivariate outliers (e.g. Orr, Sackett, & Dubois, 1991), and most statistics texts (e.g. Mertler & Vannata, 2013; Moore, McCabe, & Craig, 2017) use the more conservative rule. Consistent with these texts, I ran analyses before outlier removal, the pattern of results did not differ save for that results appeared more significant with outliers included; I report the conservative results. Examination of histograms and Q-Q plots suggested non-normal distributions for each personality/motivation construct. Kolmogorov-smirnov tests showed significant skewness/kurtosis on each (p < .05). I normalized distributions using a two-step, fractional-rank, inverse-normal transformation. This transformation simultaneously addresses skewness and kurtosis regardless of the direction/shape of non-normality, and achieves normality more often than commonly used transformations (Templeton, 2011). Resulting Q-Q plots appeared normal. Finally, when calculating \( \delta \), I found seven invalid values where the timer for “page submit” occurred before “first click”. I traced the data back to the raw excel report from Qualtrics and determined that this anomaly occurred due to measurement, reporting, or some other error and not a downstream data management error. I removed these seven participants from the data.

**Manipulation checks.** I ran independent-samples \( t \)-tests to test for differences in state AP (between control and reward) and state AV (between control and loss). Results
indicated significance for state AP $t(531) = 2.03, p < .05$ such that the reward group ($M = 3.99, SD = 0.88$) showed greater state AP than the control group ($M = 3.83, SD = 0.88$), confirming successful manipulation. For state AV, the loss group ($M = 3.06, SD = 1.01$) did not differ significantly from the control group ($M = 3.07, SD = 1.01$).

**Hypotheses Tests**

**Hypotheses 1-2.** I used hierarchical logistic regressions to test Hypotheses 1 and 2a-c that IER will be lower in the screened than non-screened group and lower in gain, loss, and combined warning groups than control, respectively. I set outcomes to dichotomized IER coded $0 =$ no and $1 =$ IER. I set predictors to screening condition (non-screened = 0, screened = 1) and three warning dummies coded for gain, loss, and combined defaulting control as reference. I included interaction terms between warning dummies and screening for exploratory purposes, and I mean centered all predictors. Screening was entered first followed by warnings and then interactions. Full results for each IER outcome are shown in Table 2.

**Infrequency.** Results showed a significant overall model [$\chi^2(7) = 66.5, p < .05$, Nagelkerke $R^2 = .10$]. The first step [$\chi^2(1) = 28.3, p < .05$, Nagelkerke $R^2 = .04$] showed that the screened group was 2.4 times less likely to fail at least one infrequency item than the non-screened group. Hypothesis 1 was supported. The second step [$\chi^2(4) = 25.4, p < .05$, $\Delta$Nagelkerke $R^2 = .04$] showed that the loss and combined warning message groups were 1.87 and 2.81 times less likely to fail an infrequency item than the control group, respectively. However, results for the loss group should be further qualified as the third step [$\chi^2(7) = 66.5, p < .05$, $\Delta$Nagelkerke $R^2 = .02$] revealed a significant interaction between screening and loss-framed messages such that loss messages significantly
reduced IER in the non-screened group but not the screened group (Figure 1).

Hypothesis 2a was not supported; Hypothesis 2b was partially supported; Hypothesis 2c was supported.

**spi.** Results revealed a significant first step [$\chi^2(1) = 6.89, p < .05$, Nagelkerke $R^2 = .04$] such that screened turkers were 2.03 times more likely to exhibit IER. Hypothesis 1 was not supported. Omnibus results for warning conditions and interactions between screening and warnings were non-significant. Hypotheses 2a-c were not supported.

**Instructed responses.** Results of the omnibus test of the overall and stepwise models were non-significant. Hypotheses 1-2 were not supported.

In summary, the results supported Hypothesis 1, which stated that IER prevalence would be lower in the screened group compared to the non-screened group, for the infrequency protocol but not **spi** or instructed responses. The results failed to support Hypothesis 2a that the prevalence of IER will be lower in the gain group compared to the control group on all three protocols. The results partially supported Hypothesis 2b that IER prevalence would be lower in the loss group compared to the control group on the infrequency protocol. And, the results partially supported Hypothesis 2c that IER prevalence would be lower in the combined group compared to the control group for the infrequency protocol but not for **spi** or instructed responses.

**Hypothesis 3.** I employed planned comparisons using chi-square analyses to test Hypothesis 3 that combined-framed messages more effectively reduce IER than gain- or loss-framed messages. Since loss-framed messages functioned differently between screening conditions for infrequency, I compared the combined-framed group to the loss-framed group within screening subgroups on this outcome. I used the full data for all
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

remaining analyses of Hypothesis 3. For infrequency, results showed lower IER prevalence in the combined group than the gain group \(\chi^2 (1) = 11.0, p < .05, \text{ odds ratio } = 0.46\) and loss group in the screened \(\chi^2 (1) = 6.02, p < .05, \text{ odds ratio } = 0.33\) but not the non-screened condition. Hypotheses 3a was supported; Hypothesis 3b was partially supported. Results were non-significant for spi and instructed responses; Hypothesis 3 was not supported for these measures of IER. Prevalence of IER for each condition on each IER protocol are shown in Figure 1.

**Hypothesis 4.** I used hierarchical logistic regressions to test Hypotheses 4a-f that people are more responsive to warning messages that are congruent with their trait motivational orientations. Like above, I ran tests involving loss-framed messages within screening subgroups for the infrequency outcome. Predictor variables were gain or loss dummies (control = 0, gain or loss = 1), motivational orientations (PA, NA, trait AP-AV, CSE) and the interaction between a motivation variable and warnings. I mean centered predictors and calculated interactions using centered predictors. I entered warning dummies first followed by motivation variables then their interaction. I used dichotomized IER as the outcome (0 = no, 1 = IER). Full model results for moderation analyses across IER outcomes appear in Table 3.

For infrequency as the outcome, results showed a significant interaction between loss-framed messages and trait AV as well as NA. However, these results were opposite as hypothesized. I ran simple slopes tests using the Hayes Process Macro 2.16.3, Model 1. For trait AV, results showed, that at \(\pm 1SD\), those above the mean in the loss-framed condition were significantly less likely to commit IER than control (OR \(= 2.12, [1.03, 4.38]\)), whereas, below the mean a stronger effect occurred (OR \(= 8.06, [2.74, 23.70]\)). In
kind, for NA, those above the mean were only marginally significantly less likely to commit IER ($p = .08, \text{OR} = 1.91, [0.92, 3.94]$); below the mean results revealed full significance and a much stronger effect ($\text{OR} = 10.11, [3.37, 23.13]$).

For $spi$ as the outcome and instructed responses as the outcome, results were non-significant. Overall, Hypotheses 4a-f were not supported. Although results did not support moderation effects in the expected direction significant interactions are shown in Figure 2.

**Research questions.** To test whether screening and/or warning to prevent IER affect the composition of personal characteristics in a sample, corresponding to Research Questions 1-2, respectively, I ran 4x2 factorial ANCOVAs for each of agreeableness, conscientiousness, positive affectivity, and negative affectivity as the outcome. Predictors were screening (non-screened and screened) and warning conditions (control, gain, loss, and combined). Consistent with the analytic strategy, and since count IER showed stronger relationships to personal characteristics, I used count infrequency IER as the continuous covariate. Results are depicted visually in Figure 3. For each analysis, interactions between screening and warning conditions were non-significant.

For agreeableness, results revealed a significant overall model [$F(8, 1048) = 3.14, p < .05$] and a significant main effect of warnings [$F(3, 1049) = 4.95, p < .05$]. Planned comparisons showed significance such that those in the combined-framed group ($M = 5.41, SE = .07$) showed significantly higher agreeableness than those in the control group ($M = 5.23, SE = .07$).

For conscientiousness, analysis showed a significant overall model [$F(8, 1069) = 3.03, p < .05$] and a significant main effect of warning messages [$F(3, 1070) = 3.02, p <$
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

A planned comparison showed marginal significance \( (p = .06) \) such that those in the loss-framed condition \( (M = 5.23, SE = .08) \) were marginally less conscientious than those in the control condition \( (M = 5.02, SE = .07) \).

For trait PA, results showed a significant overall model \( [F(8, 1062) = 2.68, p < .05] \) and a significant main effect of screening \( [F(1, 1063) = 4.25, p < .05] \). Those in the non-screened group \( (M = 4.48, SE = .04) \) showed significantly higher PA than those in the screened group \( (M = 4.36, SE = .04) \). The main effect of warnings was also significant \( [F(3, 1063) = 5.58, p < .05] \). Planned comparisons showed that those in the gain-framed group \( (M = 4.55, SE = .06) \) and those in the loss-framed group \( (M = 4.51, SE = .06) \) showed significantly greater PA than those in the control group \( (M = 4.27, SE = .06) \).

For trait NA, results revealed a significant overall model \( [F(8, 1052) = 4.80, p < .05] \) and a significant main effect of screening \( [F(1, 1053) = 10.54, p < .05] \) such that those in the non-screened group had greater NA than those in the screened condition.

**Supplemental Analyses**

Several considerations warrant further analyses given present results and the theoretical and empirical background on IER. First, I ran chi-squared difference of proportion in failing between IER indices. Different IER protocols resulted in seemingly different levels of IER prevalence and this relates to concerns regarding turker attentiveness, research practices (such as the type of detection generally used), and turker motivations and survey response tendencies (i.e. turkers easily identify familiar forms of IER detection). I used the overall level of IER prevalence across prevention conditions for each dichotomous IER index as the proportion of interest. Results showed that in
terms of relative risk (RR) turkers were more likely to fail spi than instructed responses \([\chi^2 (1) = 197.4, p < .05, RR = 7.91]\) and the infrequency protocol \([\chi^2 (1) = 8.82, p < .05, RR = 1.28]\) and more likely to fail the infrequency protocol \([\chi^2 (1) = 132.9, p < .05, RR = 6.16]\) than instructed responses.

Second, since I relied on self-report of personal characteristics, and a potential major benefit of a page level IER protocol (i.e. spi) is that it could allow for controlling data quality at the scale level (i.e. deleting invalid responses on a scale to scale basis), I retested Hypotheses 4a-f controlling for IER where 0 = no and 1 = scale IER. I used spi as the control when infrequency and instructed responses were the outcome; when spi was the outcome, I used failing any other IER detection items on the scale in question as the page level control to avoid linear dependency. I show full model results in the appendix (Table A1). Results appeared largely the same for all analyses. However, controlling for spi at the page level, I found a marginally significant interaction between gain-framed messages and PA \([\chi^2 (1) = 3.57, p = .06]\) as well as between gain-framed messages and trait AP \([\chi^2 (1) = 2.63, p = .10]\) on the infrequency outcome. Simple slopes tests showed that, at ±1SD, those below the mean of PA were significantly less likely to fail IER detection \((p < .05, OR = 1.93 [1.03, 3.62])\). For trait AP, those below the mean were marginally less likely to fail IER detection \((p = .07, OR = 1.70 [0.95, 3.05])\). On both constructs, results for those above the mean were non-significant. I show simple slopes graphically in Figure 2. These effects occurred in the opposite direction as hypothesized. Hypotheses 4a-f were not supported.

Third, I retested Research Questions 1-2 using no control and controlling for spi and instructed responses (as count variables). As previously noted, theory and evidence
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

shows that IER can inflate both type I and II error, different IER protocols capture different aspects of attentiveness, and personal characteristics predict IER. In my research, I found signs of these characteristics of IER as well as differences across relationships to personal characteristics for different IER detection methods (e.g. spi had a stronger relationship to agreeableness and no relationship to AV compared to infrequency). Therefore, I retested Research Questions 1-2 to better explore the influence of IER, its detection, and prevention on analyses and sample composition.

The primary analysis (using count infrequency IER as the control) revealed no significant interactions across screening and warning conditions, higher agreeableness and marginally lower conscientiousness in the combined group than in the control group, and lower PA and lower NA in the screened group than in the non-screened group. As in the primary analyses, the supplementary analyses showed non-significant interactions between screening and warning message conditions. I describe in the following paragraph and depict graphically in Figure 2 the full results across all analyses for Research Questions 1-2.

I compare results from the supplementary analyses (using no control and count spi and instructed responses as control) to those from the primary analyses (which used count infrequency as the control). With no control, results were similar for positive and negative affectivity but I found no significant differences for agreeableness and conscientiousness. Controlling for spi, on agreeableness, results appeared the same for the main effects of warning messages [but changed from significant to marginally significant; $F(3, 1049) = 2.53, p = .06$] and planned comparisons (but still fully significant; $p < .05$). For conscientiousness, results differed such that the main effect of
warning messages was non-significant (vs significant as in the primary analysis), but the main effect of screening was marginally significant \([F(1, 1070) = 3.55, p = .06]\)—those in the screened group \((M = 5.21, SE = .05)\) showed marginally higher conscientiousness than those in the non-screened group \((M = 5.08, SE = .05)\). For PA, main effects and planned comparisons for warning messages were the same, but the main effect of screening was marginally significant. For NA, results were the same. Controlling for instructed responses, results were also the same as when controlling for the count infrequency IER save for that the main effect of warning messages on agreeableness were marginally significant \([F(3, 1049) = 2.47, p = .06]\) and all results for conscientiousness were non-significant.

In sum, regardless of whether I used a control or which IER index I used as control, those in the screened group were lower in NA compared to those in the non-screened group. Also regardless of control set-up, combined-framed warning messages increased agreeableness compared to no messages, and gain- and loss-framed messages increased PA compared to the control group.
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

Discussion

IER Prevention Results

Concerning Hypotheses 1-3, the results paint a somewhat dismal picture for IER prevention on Mturk. Although a) screened turkers appeared more attentive than non-screened turkers, b) loss- and combined-framed messages decreased IER compared to control, and c) combined-framed messages most effectively reduced IER [and these results are consistent with warning message IER prevention research in student samples (e.g. Blackmore, 2014; Huang et al., 2012)], this only held true for the infrequency measure. When IER detection methods may be easier to identify or familiar to turkers (i.e. instructed responses), IER was low and relatively equal across conditions and/or compared to control. When IER detection methods were impossible to recognize (i.e. spī), screened turkers committed more IER than non-screened turkers, and, in both screened (27%) and non-screened (20%) groups, the prevalence of IER appeared much higher than negligible and in previous screening studies. This finding is consistent with previous work where researchers found that screened turkers failed novel forms of IER at greater rates than non-screened turkers (i.e. Hauser & Schwarz, 2016; Peer et al., 2014).

In sum, screening as a means to reduce IER may not generalize to longer surveys, like the one I used, and more rigorously defined and less identifiable IER indices (such as spī). Likewise, warning messages as a means to reduce IER may not generalize to the Mturk population. Even the loss-framed only condition that mirrored previous warning message research only reduced IER on a single index (i.e. the infrequency index) and only in a sub-group of turkers (i.e. the non-screened group), where most researchers do not conduct their research. More than that, turkers failed IER protocols in precisely the linear
order as expected. That is, very few failed instructed responses; they were more likely to fail infrequency items (6.6x), and even more likely to fail spi (7.9x). The present results and interpretation agrees with previous concerns regarding turkers’ motivation to participate as a means to generate or supplement income and share information on message boards such as “look out for attention checks” (e.g. Berinsky et al., 2012; Harms & DeSimone, 2015). It also agrees with previous research showing that turkers pass familiar forms of IER detection at vastly higher rates than novel forms (e.g. Hauser and Schwarz, 2016).

More than that, and also in regard to painting a dismal picture for preventing IER on Mturk, the results from Research Questions 1-2 revealed that screening and warning messages related to the composition of a sample possibly selecting in those of specific traits and/or possibly changing the way participants responded to content items about stable traits, respectively. In terms of screening, if turkers in a screened sample commit less IER because they differ in personal characteristics vs some other practical reason—such as the fact that more experienced/reputable turkers qualify for higher paying HITs—one can expect predictable results. Based on previous findings (Bowling et al., 2016; Camus, 2015) and the covarying and interrelated nature of PA/NA with extraversion/neuroticism (e.g. Elliot & Thrash, 2002, 2010; Ferris et al., 2011), one could expect that those in the screened group would be lower in negative affectivity and higher on positive affectivity, conscientiousness, and agreeableness. Indeed, irrespective of which IER index I used as control, NA was lower in the screened sample and, although marginal, conscientiousness was higher when controlling for the impossible to identify spi index. Thus, a small majority (5/9) of the significant results provide evidence that
screened turkers commit less IER (at least in part given results from H1-H3) because they possess different personal characteristics. I discuss these results further as they relate to other IER indices and prevalence results (i.e. that IER defined by \( spi \) was higher in the screened group) in the IER protocols section of this discussion.

In terms of warning messages, although combined-framed messages most effectively reduced IER on the infrequency scale, they also yielded one of the few most consistent group difference across IER controls, in this case, on agreeableness such that turkers in the combined message group had higher levels of agreeableness than the control group. Moreover, warning messages yielded the majority of group differences in terms of profiles of personal characteristics (Figure 3). Additionally, above all other findings, loss-framed messages accounted for nearly half of all significant differences across preventions conditions and personality constructs irrespective of control. This result seems most problematic since researchers using Mturk most often use loss-framed messages. The potential explanations for these difference seem vast, so I will not attempt to explore them although I will point out that these results support the idea that survey responses do not occur in a “situational vacuum” in agreement with the proposition by LST theory (Steyer et al., 1999).

**Congruency Effect Results**

The results regarding the congruency effect paint a cautionary picture for conducting organizational research on Mturk, but they may also add further explanation to the apparent paradoxical influence of warning messages noted above. Although previous research concluded that turkers replicate findings on classic psychological studies and other phenomena (e.g. Crump et al., 2013; Horton et al., 2011), the present
study largely failed to establish the congruency effect (i.e. that messages motivate desired behavior when they hold the same positive or negative valence as a persons’ motivational orientations). However, the lack of replications does not necessarily establish that researchers cannot trust turkers to provide valid data (or as valid of data as other samples) and to think, feel, and behave as other humans do. To be honest, these notions are silly. One possible explanation for the lack of replication rests in principles of the proximity of personal characteristics to behaviors. Specific to AAT, latent positive and negative personality/motivational temperaments lie further away from behavior than the more proximal motivations that they predict, and psychological researchers (Hamamura et al., 2009; Higgins, 1997; Shah et al., 1998; Treiber, 1986; Wilson et al., 1990) consistently find congruency when measuring the more proximal motivations. Since prior research lacks knowledge of turker motivations more proximal to their survey response behaviors and/or mediating self-set earnings, reputation, and/ or productivity goals to participate in research, I measured the more distal temperaments and traits. Therefore, the lack of replication may simply depict the ostensibly much smaller and lesser-known effects of congruency with more distal characteristics. More than that, although minimal, some of the results align with previous congruency and AAT research. Specifically, on some measures, combined-framed messages most effectively reduced IER which is consistent with other AAT message framing research that found similar results (e.g. Slater, 2014; Treiber, 1986; Wilson et al., 1988).

Nevertheless, I still urge caution. Opposite of hypothesized, IER decreased or decreased to greater extent among those lower compared to higher in trait AP and PA (in response to gain-framed messages) and trait AV and NA (in response loss-framed
IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

messages; Figure 2). In alternative explanation terms, the nature of the infrequency scale might explain this. The infrequency scale’s current design and present use places failure on an item at the upper end of the scale; as such, the results may have captured those tending to answer inattentively by selecting upper responses. This would have to be a general tendency, as I placed no infrequency items in the PANAS scale, meaning that participants could not have failed an upper response IER item on positive and negative affectivity. Thus, a paradoxical relationship would only occur due to response position on the PANAS for those participants who responded inattentively by selecting upper responses on most/all scales. Also, if this general tendency to select upper responses inattentively did explain the results, I would have found more instances where any condition (gain, loss, or otherwise) showed congruency with the lower ends of scale variables regardless of the construct, which occurred not. The paradoxically congruent results only occurred when the condition and the scale shared positive/negative valence. However, it is possible that the position of correct responses exerted some influence, which does bring into question the design of the infrequency scale. In theoretical terms, the paradoxical findings may represent a non-biased result—psychologically similar or related to the congruency effect in that it occurred when the valence of the stimuli matched the turkers’ motivational orientation—that persists in certain contexts or specific situations for which AAT and its principles do not account. Indeed, Nguyen (2018) found a similarly paradoxical result for avoidance motivation (in this case prevention focus) in a sample of seasonal employees such that the negative relationship between perceived over-qualification and intrinsic motivation was more negative for those lower (and not higher as expected) in prevention focus.
Potential explanations for why I found paradoxical effects may already exist in the literature, but since I designed this study to test for the congruency effect, I cannot confirm them, so I explore one. The mobilization-minimization hypothesis (Taylor, 1991) suggests (among much more) that negatively valenced (or adverse/threatening) events mobilize cognitive and emotional processes to minimize the event and or threat. One of these processes includes a compensatory effect where after the event—such as moving beyond the manipulation message in this case—a person experiences an opposing wave of positive emotions (i.e. more than cessation of negative emotional response). Principles of the mobilization-minimization hypothesis appear asymmetrical given whether people experience a positive vs negative event and/or have higher or lower NA. That is, the compensatory effect does not occur in response to positive events and effects of this offset are stronger for those lower in NA (e.g. Ito & Cacioppo, 2005; Norris, Larsen, Crawford, & Cacioppo, 2011). Some of the present results also support this idea. For example, the loss-framed group reported higher PA than control (Figure 3). Thus, it is possible that those lower in negatively-valenced-personal-characteristics in the loss-framed group responded with positivity (appearing lower in negative characteristics than so oriented) and busied themselves with minimizing negative responses differently than those in the higher group (such as reducing inattentiveness vs ignoring the warnings; left panel of Figure 2). In sum and put simply, reporting of personal characteristics and responses to positively/negatively valenced events might have yielded results that appeared paradoxical when they were not, or, worked to a minimizing vs congruent effect. Unfortunately, this explanation seems less intuitive for the paradoxical results regarding gain-framed messages (right panel of Figure 2), and I cannot fully explore the
potential changes in reporting of personal characteristics relative to their “true” levels and as they might relate to the valence of warning messages since my design did not include a pretest.

Limitations

The present study contains a couple, somewhat minor, but notable limitations. These limitations might have influenced the results and interpretation. First, Wood and others (in press) derived the cut-off for \( spi \) from the general, less experienced, Mturk population. They collected data in two rounds to assess response consistency as a function of \( spi \) and noted that turkers completed the same scale in 80% of the time that it took to complete them during the first round but used response times from the first round to define the cut-point. If participants respond faster on a familiar scale, it stands to reason that more experienced turkers would respond faster than others, which might account for the differences across screening conditions and/or the volume of IER in both conditions as opposed being representative of IER. I was unable to eliminate this alternative explanation. Similarly, although the \( spi \) index relies on a mean of response times per item, the average number of words per item or per scale may differ between scales. Indeed, as apparent in Wood and others (in press), the demarcation at \( spi = 1 \) (\( \log_{2}spi = 0 \)) between consistent responses and inconsistent responses seems much clearer on one-word-per-item scales such as PANAS than on multiple-words-per-item scales such as the dark triad. Although not reported, I tested failure proportions on \( spi \) across PANAS and CSE (simply chosen at random) and a greater proportion of turkers failed the one-word-per-item PANAS [\( \chi^2 (1) = 21.2, p < .05, RR = 2.08 \)]. Thus, wordier scales may
have contained more IER than the current $spi$ cut-offs can capture. However, since the $spi$ showed multiple signs that it worked similar to other IER indices (such as its relationship to other variables; discussed in the IER protocols section). The noise due to this limitation certainly cannot account for more than a share of pertinent results.

Second, the relative probability of (in)correctly responding to an IER item could have influenced the results regarding pass/fail rates across IER protocols. The probability of correctly answering all infrequency or instructed response items equals .001 (0.11% chance) and .003 (0.29% chance), respectively. On the surface this suggests that people have a lower probability of guessing all infrequency items correctly (thus a higher probability of failing at least one item), and, indeed, the prevalence of dichotomized IER was significantly greater on this measure compared to instructed responses. I cannot adjust the relative risk using item level chance since the $spi$ protocol differs in that regard. However, the overall opportunities to fail do apply to all three indices, and the proportional difference at this level is similar for instructed responses compared to infrequency items. That is, turkers had 3, 8, and 13 opportunities to fail instructed responses, infrequency items, and $spi$, respectively; for the first two indices, the proportional difference at the item level (.11/.29 = .379) is similar to the proportional difference in number of items (3/8 = .375). To explore the present limitation (i.e. not reported elsewhere), I adjusted the relative risks (reported in the supplementary analysis section) down by the item/page level probabilities and turkers still had a greater risk of failing infrequency and $spi$ items/pages than instructed responses ($RR = 2.31$ and $1.81$, respectively).
Another limitation seems a bit more substantial. The present study relied on self-reports of personal characteristics, which may contain IER. That is, if a turker responded inattentively, the validity of reporting on their personal characteristics comes directly into question. In this case, however, I was able to control for the influence of IER using multiple IER indices (i.e. Figure 3), which certainly removes much of that concern. Similarly, the spi index allowed me to control for page-scale level IER when testing for congruency effects (Table A1). Unfortunately, the self-report issue also means that I cannot tease apart the relative influence of IER vs personality when results changed due to control. Also, I could not control for the scale level IER on the spi outcome, because I could not place any infrequency or instructed response items in single-word scales, as such, scale level IER on a few congruency tests using spi as the outcome resulted in a linear dependency (Table A1).

**Theoretical Implications and Future Research**

The results from combined messages on the infrequency IER outcome carries multiple theoretical implications. First, it supports the perspective that approach and avoidance tendencies represent independent systems that can be activated simultaneously. Whether warning messages motivated turkers to pay more attention or simply look harder for IER items, the results still represent greater action; consequently, combined messages showed the greater effect so this may still imply independent dual activation. These results brings back into question the theoretical structure of efforts to motivate behavior (e.g. health promotion). The vast majority of research in this regard has long focused on under what circumstances gain- or loss-framed messages are most effective (Slater, 2014). Second, that I found no (vs. significant) differences on positive or negative
affectivity in response to combined-framed messages (vs. gain- and/or loss-framed messages) raises further questions as to how dual activation might influence emotions and behavior. Although the motivating effects of combined messages did not occur for *spi*, I cannot rule out the alternative explanations regarding noise on the *spi* protocol (see limitations). Thus, to the extent that conclusions in the present paper prove accurate, combined-framed messages certainly have interesting and useful implications.

**Congruency implications.** The results from congruency effect analyses present new information to approach-avoidance motivation theories and raises new questions concerning congruency. Most mundanely, it extends the notion of congruency effects to similarly valenced constructs; specifically, PA and NA showed the same paradoxical congruency with the gain-framed and loss-framed messages as the parent constructs, AP and AV temperament, respectively. This provides further evidence for the idea that personality and motivation relate to each other in specific ways—AP-AV temperament represent a latent two-factor structure on which positively and negatively valenced personal characteristics load (i.e. Elliot and Thrash, 2002). However and perhaps most interestingly, in the present context, the results bring into question assumptions of and the generalizability of congruency. That is, under what circumstances or contexts might congruency between an individual’s characteristics and the valence of stimuli yield paradoxical results (such as Figure 2), and how might these effects be leveraged or avoided in motivational situations such as health promotion and other organizational behaviors/topics. Furthermore, although the present results showed no evidence of congruency between the valence of stimuli and CSE [as would be expected given Chang and others (2012), Ferris and colleagues (2011)], these results leave open the questions as
to how appropriately CSE fits in an approach-avoidance framework. That is, although results were non-significant for CSE, the paradoxical findings may prove specific to the present or similar context(s), leaving the non-significant results also specific to certain situations. Future research might seek to replicate tests of congruency between CSE and positive/negatively valenced stimuli in different contexts, populations, or desired behaviors.

**Implications for IER and IER protocols.** Overall, the results provide further evidence for theoretic assumptions underlying IER as a comprehensive term that can result in both increased type I and type II errors and raises new questions as to the conditions under which one or the other can occur. In almost all cases apparent in Research Questions 1-2 (Figure 1) controlling for IER revealed significant relationships between warning messages/screening to personal characteristics that I failed to find when I did not control for IER. This aligns with the traditional perspective that random noise can attenuate relationships (e.g. Buechly & Ball, 1952; Spearman, 1904). The same attenuation seemed apparent when I controlled for scale level IER in the supplementary congruency analyses where the relationship between gain-framed messages and positively valenced personal characteristics (Figure 2), became marginally to fully significant where they were non-significant without control. That I found more evidence for increased probability for type II errors makes sense since the conditions under which IER can inflate type I errors applies to scale midpoint deviation on two scale variables (e.g. Huang, Liu, et al., 2015; McGonagle et al., 2016).

However, in spite the fact that my analyses only contained a single survey scale, when controlling for the influence of *spi* on the relationship between screening and PA,
the results became less significant \(p = .042\) vs \(p = .087\). Although I cannot tease apart the relative influence of IER vs personality in the relationships between the different IER indices and personality variables, I can provide evidence that the change in significance when controlling for an IER index stems from IER. In all cases, controlling for IER moved the mean of the scale away from the midpoint to a greater degree in the condition which displayed higher IER on that index (consistent with Huang, Liu, et al., 2015).

Future research might focus on further refining the conditions under which IER would result in type I vs type II error.

As a final, overall note, as discussed above, the screened group in the present study should have had higher PA if they provided higher quality data (on instructed response type detection methods, in previous studies, and the infrequency scale, as I found in the present study) due to their personal characteristics. However, they dependably showed significantly or marginally significant lower PA. Interestingly, the screened group failed spi detection at greater proportions. This result provides further evidence that different IER detection methods might detect different aspects of IER and raises the question as to how different personal characteristics predict IER and what specifically do different detection methods capture (beyond the surface level differences of speed vs following instructions for example). Future research might seek to assess exactly how different indices capture different aspects of IER, how different indices might work together to serve as a single control variable, and why as well as what levels of IER and/or IER detection/deletion might bias a sample.

Concerning the infrequency scale, although it almost entirely yielded results as theoretically expected, and although less likely or influential in the present case (see the
congruency effect results section of this manuscript), there may be noise introduced by the position of correct and incorrect responses. Future research might seek to refine this measure by developing and validating items that spread correct responses across response options and/or investigating experimentally whether position of the correct responses influences IER related phenomenological results. More than that, as more researches begin using this IER detection method, turkers may become more familiar with it such that the items become more readily identifiable. Future research might seek to develop a much larger bank from which to randomly draw and define the minimum number of items to include (or test whether greater or fewer items result in different levels of IER detection) such that researchers can make informed decisions about how to incorporate them into their survey protocols.

For spi, the present research provides positive and negative implications. First, the results add to the validity of using spi as a measure of IER by showing that it relates to other well-defined IER indices and personal characteristics and/or variables that deviate from the mean in the expected directions (again, I cannot fully tease apart the influence of IER vs personality on the descriptive relationships Table 1). However, save for a few cases regarding dichotomized spi (e.g. spi related to agreeableness positively, \( r = .15 \), but infrequency and instructed responses related negatively, \( r = -.09 \) and -.14, respectively), and one case for count spi (where there was a non-significant relationship between spi and trait AV), spi correlated to survey variables above the mean negatively and below the mean positively. This matches the expected theoretical direction that IER gathers near the midpoint creating these relationships with survey variables displaying mean deviation from the midpoint (i.e. Huang, Liu, et al., 2015); it also matches the
performance of the infrequency scale (columns 5, 6, 8, and 9 of Table 1). If the influence stems from actual relationships to personal characteristics (e.g., Bowling et al., 2016; Camus, 2015), the results also match. In fact, count $spi$ showed the strongest of these correlations (column 9, Table 1). Second, controlling for $spi$ revealed previously unfound relationship (adjusting type II error) and reduced one existing relationship (adjusting type I error). In these cases, controlling for $spi$ moved the mean away from the midpoint as expected. These results extend Wood and colleagues (in press) findings by showing that $spi$ functions similarly as other IER measures based on the updated theoretical definitions of IER. More than that, the results when controlling for page-scale level IER (i.e. that multiple relationship emerged in the congruency analysis as a result) allude to the validity of using a page-scale level IER measure and raises new questions as to the theoretic definitions of what constitutes IER responses vs IER respondents.

However, some aspects of $spi$ may create noise in detecting IER and require refinement to the $spi$ as an IER protocol. As noted in the limitations, turker experience with familiar scales may result in faster responses and the current cut-offs may not account for IER in wordier scales. If turkers grow familiar with a scale such as PANAS, know their answers/themselves, and how to respond such that they respond much quicker than cutoff, this does not necessarily represent IER. Future research might seek to assess the relative differences in $spi$ cutoffs when dealing with measures turkers commonly encounter compared to those they may encounter less frequently. Similarly, future research might consider conditioning $spi$ on the mean of words per items, and/or the number of items, to better capture IER on wordier scales.
Finally, in terms of instructed responses, although the results provide some evidence that turkers easily identify them, among those that did fail compared to those that did not, this type of IER detection still showed signs of validly detecting IER. That is, among those that did fail instructed responses compared to those that did not, much like $\text{spi}$, the expected theoretical relationships occurred. Interestingly, this was the only detection protocol for which every relationship significantly aligned with scale variables as expected. In addition, similar to the other protocols, controlling for instructed responses revealed relationship not apparent when not controlling for IER.

**Practical Implications and Concluding Remarks**

First and foremost, let us treat Mturk as a special population similar to a particular industry, region, nation, or other classification of participants—such as students (who actually perform worse on IER measures; Hauser and Schwarz, 2016). Conducting research on Mturk can yield valid and generalizable results (e.g. Berinski et al., 2012; Wolfson & Bartkus, 2013), but the evidence put forth in the present research shows that some psychological phenomena, particularly those that relate to the congruency effect derived from AAT, may differ among turkers. Researchers questioning whether results on Mturk generalize to other populations should replicate studies in the population of interest, and research conducted and/or replicated on Mturk, which does its diligence when so doing, should stand alone as valid.

Second, avoid using screening and warning messages. These techniques can bias your samples. Although some research (e.g. Peer et al., 2014) showed that conducting studies from the general Mturk population results in more IER, this may not be the case after all. Future research might benefit from investigating other methods of assuring
higher data quality, but should also consider/study how so doing might influence data, results, and interpretations. Currently, if you wish to reduce IER without other biasing influence, use shorter surveys (see Yentes, 2015). Although exactly how long of a survey is too long is unknown, given the prevalence of IER in other studies, discussed above, somewhere around 50 items but much less than 100 should keep IER in more manageable ranges. Furthermore, I do not mean to suggest using IER indices as an adjustment factor (i.e. finding “true” scores on intelligence, “real” levels of racism, or “adjusted” conscientiousness on a selection tool) since no research provides concrete evidence as to which index or combination of indexes best validates results. Instead, as I demonstrated in my research, control for IER to assess whether results and interpretations change; use the general Mturk population; and avoid rejecting HITs and deleting those committing IER. Although no research shows how much IER deletion will bias a sample, deleting IER, particularly at high levels, absolutely can. Moreover, rejecting HITs causes conflict between researchers and our valued participants and might result in the same biasing effect as deleting participants post hoc, and you will remove potentially viable information since IER on one page, given the detection type such as spi, may not imply IER on another and no research suggests otherwise. Your research objectivity and ethical standards will do the rest. In addition, looking at the descriptives in my research, you can control for IER as a categorical or a count variable as your analysis requires (in fact, the count IER operationalization had the strongest relationships to other content variables with which they should have).

Third, use multiple IER detection protocols; avoid using the same ones in multiple studies (turkers follow researchers; Chandler et al., 2013), and using, as a sole
means of detection, those that turkers looking to “cheat” the system can easily identify.

Moreover, as I found in the current study, controlling for different IER indices may yield different results, so using only a single IER protocol reduces their utility and leaves the validity of results in question. I recommend using both a priori and post hoc detection methods—at least one that is not bound by position in the scale, such as instructed responses, one more rigorously defined, such as the infrequency scale, and one that turkers cannot identify, such as spi. Although instructed responses may be the easiest to identify, they are also the most versatile in that the keyed response can occur at any point in the scale (I recommend at least one for each end and one for the middle), and researchers can word them in any number of ways. I recommend wording them differently each time and masking them around other text. That is, as opposed to asking, “please select strongly agree”, you might ask something such as “I am the type of person that likes to [blank…some word that fits the scale at hand]. Ignore this first sentence and mark always”. Since the possibilities are vast, you never have to ask it the same way twice.
### Table 1

**Raw descriptive statistics for focal variables and conditions**

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Screened&lt;sup&gt;A&lt;/sup&gt;</td>
<td>.45</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
</tr>
<tr>
<td>2.</td>
<td>Gain&lt;sup&gt;B&lt;/sup&gt;</td>
<td>.24</td>
<td>.42</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
</tr>
<tr>
<td>3.</td>
<td>Loss&lt;sup&gt;B&lt;/sup&gt;</td>
<td>.24</td>
<td>.42</td>
<td>-.01</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
<td>-.31</td>
</tr>
<tr>
<td>4.</td>
<td>Combo&lt;sup&gt;B&lt;/sup&gt;</td>
<td>.28</td>
<td>.45</td>
<td>-.08</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
<td>-.34</td>
</tr>
<tr>
<td>5.</td>
<td>Inf-Dic&lt;sup&gt;C&lt;/sup&gt;</td>
<td>.18</td>
<td>.39</td>
<td>-.16</td>
<td>-.05</td>
<td>-.04</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
<td>-.11</td>
</tr>
<tr>
<td>6.</td>
<td>Spi-Dic&lt;sup&gt;C&lt;/sup&gt;</td>
<td>.24</td>
<td>.42</td>
<td>-.16</td>
<td>-.01</td>
<td>-.05</td>
<td>.01</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
</tr>
<tr>
<td>7.</td>
<td>IR-Dic&lt;sup&gt;C&lt;/sup&gt;</td>
<td>.03</td>
<td>.17</td>
<td>.03</td>
<td>.05</td>
<td>.02</td>
<td>.04</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
</tr>
<tr>
<td>8.</td>
<td>Inf-Count&lt;sup&gt;D&lt;/sup&gt;</td>
<td>.36</td>
<td>.85</td>
<td>-.17</td>
<td>-.05</td>
<td>-.10</td>
<td>.89</td>
<td>.20</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
</tr>
<tr>
<td>9.</td>
<td>Spi-Count&lt;sup&gt;D&lt;/sup&gt;</td>
<td>.63</td>
<td>1.73</td>
<td>-.12</td>
<td>-.04</td>
<td>-.02</td>
<td>.04</td>
<td>.27</td>
<td>.65</td>
<td>.33</td>
<td>.27</td>
<td>.33</td>
<td>.27</td>
<td>.33</td>
<td>.27</td>
<td>.33</td>
<td>.27</td>
<td>.33</td>
<td>.27</td>
</tr>
<tr>
<td>10.</td>
<td>IR-Count&lt;sup&gt;D&lt;/sup&gt;</td>
<td>.03</td>
<td>.20</td>
<td>.02</td>
<td>.05</td>
<td>.03</td>
<td>.03</td>
<td>.27</td>
<td>.25</td>
<td>.96</td>
<td>.32</td>
<td>.45</td>
<td>.32</td>
<td>.45</td>
<td>.32</td>
<td>.45</td>
<td>.32</td>
<td>.45</td>
<td>.32</td>
</tr>
<tr>
<td>11.</td>
<td>Big 5 Agree</td>
<td>5.23</td>
<td>1.25</td>
<td>-.05</td>
<td>-.02</td>
<td>.00</td>
<td>.09</td>
<td>-.09</td>
<td>-.15</td>
<td>-.14</td>
<td>-.09</td>
<td>-.24</td>
<td>-.14</td>
<td>-.14</td>
<td>-.24</td>
<td>-.14</td>
<td>-.24</td>
<td>-.14</td>
<td>-.24</td>
</tr>
<tr>
<td>12.</td>
<td>Big 5 Consc</td>
<td>5.14</td>
<td>1.21</td>
<td>-.06</td>
<td>-.06</td>
<td>-.03</td>
<td>-.10</td>
<td>-.09</td>
<td>-.13</td>
<td>-.11</td>
<td>-.19</td>
<td>-.12</td>
<td>-.23</td>
<td>(.86)</td>
<td>(.86)</td>
<td>(.86)</td>
<td>(.86)</td>
<td>(.86)</td>
<td>(.86)</td>
</tr>
<tr>
<td>13.</td>
<td>Trait PA</td>
<td>4.40</td>
<td>0.95</td>
<td>-.07</td>
<td>.07</td>
<td>.07</td>
<td>-.03</td>
<td>-.12</td>
<td>-.12</td>
<td>-.01</td>
<td>-.21</td>
<td>-.10</td>
<td>.34</td>
<td>.36</td>
<td>(.85)</td>
<td>(.85)</td>
<td>(.85)</td>
<td>(.85)</td>
<td>(.85)</td>
</tr>
<tr>
<td>14.</td>
<td>Trait NA</td>
<td>1.89</td>
<td>0.96</td>
<td>-.12</td>
<td>-.02</td>
<td>-.02</td>
<td>.00</td>
<td>.13</td>
<td>.05</td>
<td>.15</td>
<td>.13</td>
<td>.18</td>
<td>.14</td>
<td>-.16</td>
<td>-.35</td>
<td>-.23</td>
<td>(.83)</td>
<td>(.83)</td>
<td>(.83)</td>
</tr>
<tr>
<td>15.</td>
<td>Trait AV</td>
<td>3.69</td>
<td>1.49</td>
<td>-.02</td>
<td>-.02</td>
<td>-.03</td>
<td>.01</td>
<td>.09</td>
<td>-.01</td>
<td>.07</td>
<td>.10</td>
<td>.03</td>
<td>.07</td>
<td>-.09</td>
<td>-.31</td>
<td>-.32</td>
<td>.52</td>
<td>(.89)</td>
<td>(.89)</td>
</tr>
<tr>
<td>16.</td>
<td>Trait AP</td>
<td>5.15</td>
<td>1.04</td>
<td>-.15</td>
<td>-.01</td>
<td>.04</td>
<td>.01</td>
<td>-.07</td>
<td>-.18</td>
<td>-.15</td>
<td>-.05</td>
<td>-.29</td>
<td>-.14</td>
<td>.36</td>
<td>.25</td>
<td>.53</td>
<td>-.23</td>
<td>-.13</td>
<td>(.87)</td>
</tr>
<tr>
<td>17.</td>
<td>CSE</td>
<td>4.93</td>
<td>1.08</td>
<td>.01</td>
<td>.02</td>
<td>.04</td>
<td>-.01</td>
<td>-.08</td>
<td>-.07</td>
<td>-.12</td>
<td>-.07</td>
<td>-.17</td>
<td>-.12</td>
<td>.24</td>
<td>.46</td>
<td>.53</td>
<td>-.60</td>
<td>-.66</td>
<td>.46</td>
</tr>
</tbody>
</table>

**Note.**  
Inf = Infrequency Scale, Spi = The speed index (speed per item), IR = Instructed responses, Agree = Agreeableness, Consc = Conscientiousness, PA = Positive Affectivity, NA = Negative Affectivity, AV = Avoidance, AP = Approach, CSE = Core Self-Evaluations, Values in ( ) = Cronbach’s alpha, *p < .05, **p < .01.

<sup>A</sup> Variable is dichotomous coded, 0 = No and 1 = Yes.

<sup>B</sup> Variable is dummy coded relative to other warning message conditions, 0 = Other conditions and 1 = In condition.

<sup>C</sup> Variable is the dichotomous IER operationalization coded 0 = None failed and 1 = Failed any item/page.

<sup>D</sup> Variable is the count IER operationalization coded as the number of items/pages failed.
Table 2
Hypotheses 1-2: Stepwise results predicting IER on each detection protocol

| Variable                  | Model One |          |         |          |         |         |         | Model Two |          |         |         | Model Three |         |         |          |         |         |         |         |         |         |         |         |         |         |         |
|---------------------------|-----------|----------|---------|----------|---------|---------|---------|-----------|----------|---------|---------|-------------|---------|---------|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| **Infrequency Scale**    |           |         |         |          |         |         |         |           |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |
| Screening                | 28.3**    | -0.88** | 0.42    | [0.30, 0.58] | 25.4** | -0.96** | 0.38    | [0.27, 0.54] | 12.8*    | -0.89** | 0.41    | [0.29, 0.59] |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Gain                     | -0.14     | 0.87    | [0.58, 1.31] |        | -0.13   | 0.88    | [0.56, 1.37] |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Loss                     | -0.63**   | 0.54    | [0.34, 0.83] |        | -0.49*  | 0.61    | [0.39, 0.96] |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Combined                 | -1.03**   | 0.36    | [0.23, 0.56] |        | -0.95** | 0.39    | [0.24, 0.64] |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Gain*Screening           | -0.05     | 0.95    | [0.38, 2.38] |        |         |         |         |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Loss*Screening           | 1.47*     | 4.34    | [1.73, 10.9] |        |         |         |         |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Combined*Screening       | 0.55      | 1.74    | [0.62, 4.92] |        |         |         |         |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |

**Instructed Response**

| Screening                | 6.89**    | 0.38**  | 1.46    | [1.10, 1.94] | 1.07    | 0.39**  | 1.47    | [1.11, 1.96] | 2.57    | 0.39    | 1.47    | [1.11, 1.96] |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Gain                     | -0.01     | 0.99    | [0.66, 1.49] |        | 0.05    | 1.05    | [0.70, 1.59] |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Loss                     | -0.10     | 0.91    | [0.60, 1.37] |        | -0.06   | 0.94    | [0.62, 1.43] |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Combined                 | 0.11      | 1.12    | [0.76, 1.64] |        | 0.14    | 1.15    | [0.78, 1.71] |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Gain*Screening           | -0.66     | 0.52    | [0.23, 1.18] |        |         |         |         |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Loss*Screening           | -0.43     | 0.65    | [0.29, 1.50] |        |         |         |         |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| Combined*Screening       | -0.38     | 0.68    | [0.31, 1.49] |        |         |         |         |        |         |         |         |         |             |         |         |          |         |         |         |         |         |         |         |         |         |         |         |         |         |         |

Predictors are categorical dummies, 0 = no, 1 = yes. Est = odds ratio. OR = Odds Ratio. \( \hat{r} < .10 \). \( *p < .05 \). \( **p < .01 \).
<table>
<thead>
<tr>
<th></th>
<th>Step $\chi^2$</th>
<th>Est. as the Outcome</th>
<th>Odds Ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 3</td>
<td></td>
</tr>
<tr>
<td>Gain</td>
<td>0.70</td>
<td>-0.17</td>
<td>0.84</td>
<td>[0.56, 1.26]</td>
</tr>
<tr>
<td>Gain*PA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain</td>
<td>0.94</td>
<td>-0.20</td>
<td>0.82</td>
<td>[0.54, 1.23]</td>
</tr>
<tr>
<td>Gain*AP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain</td>
<td>0.95</td>
<td>-0.20</td>
<td>0.82</td>
<td>[0.54, 1.23]</td>
</tr>
<tr>
<td>Gain*CSE</td>
<td></td>
<td>-0.26**</td>
<td>0.77</td>
<td>[0.63, 0.95]</td>
</tr>
<tr>
<td>Screened</td>
<td>Loss</td>
<td>0.52</td>
<td>1.31</td>
<td>[0.63, 2.71]</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>0.87**</td>
<td>2.40</td>
<td>[1.52, 3.76]</td>
</tr>
<tr>
<td></td>
<td>Loss*NA</td>
<td>-0.66</td>
<td>0.52</td>
<td>[0.21, 1.31]</td>
</tr>
<tr>
<td></td>
<td>Loss</td>
<td>0.85</td>
<td>1.40</td>
<td>[0.68, 2.88]</td>
</tr>
<tr>
<td></td>
<td>Loss*CSE</td>
<td>-0.27</td>
<td>0.76</td>
<td>[0.53, 1.11]</td>
</tr>
<tr>
<td></td>
<td>Non-screened</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Loss</td>
<td>17.0**</td>
<td>-1.16**</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>0.37*</td>
<td>1.45</td>
<td>[1.04, 2.02]</td>
</tr>
<tr>
<td></td>
<td>Loss*NA</td>
<td>1.00**</td>
<td>2.71</td>
<td>[1.27, 5.78]</td>
</tr>
<tr>
<td></td>
<td>Loss</td>
<td>17.0**</td>
<td>-1.16**</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>Loss*CSE</td>
<td>-0.42**</td>
<td>0.66</td>
<td>[0.49, 0.87]</td>
</tr>
<tr>
<td></td>
<td>Note: PA/NA = Positive/Negative Affect, AP = Trait Approach, CSE = Core Self-Evaluations. $^*p &lt; .10$, $^*p &lt; .05$, $^{**}p &lt; .01$.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss as the Outcome</td>
<td>Est.</td>
<td>Odds Ratio</td>
<td>95% CI</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>-----</td>
<td>------------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td><strong>Gain</strong> 0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>[0.67, 1.51]</td>
<td></td>
</tr>
<tr>
<td><strong>PA</strong></td>
<td>-0.12</td>
<td>0.89</td>
<td>[0.71, 1.11]</td>
<td></td>
</tr>
<tr>
<td><strong>Gain*PA</strong></td>
<td>0.00</td>
<td>1.00</td>
<td>[0.64, 1.58]</td>
<td></td>
</tr>
<tr>
<td><strong>Gain</strong> 0.06</td>
<td>-0.05</td>
<td>0.91</td>
<td>[0.63, 1.44]</td>
<td></td>
</tr>
<tr>
<td><strong>AP</strong></td>
<td>-0.33**</td>
<td>0.72</td>
<td>[0.57, 0.91]</td>
<td></td>
</tr>
<tr>
<td><strong>Gain*AP</strong></td>
<td>0.00</td>
<td>-0.07</td>
<td>0.99</td>
<td>[0.66, 1.50]</td>
</tr>
<tr>
<td><strong>CSE</strong></td>
<td>-0.07</td>
<td>0.93</td>
<td>[0.76, 1.14]</td>
<td></td>
</tr>
<tr>
<td><strong>Gain*CSE</strong></td>
<td>0.00</td>
<td>1.29</td>
<td>0.46</td>
<td>[0.65, 1.49]</td>
</tr>
<tr>
<td><strong>Loss</strong> 0.68</td>
<td>-0.17</td>
<td>0.84</td>
<td>[0.46, 0.99]</td>
<td></td>
</tr>
<tr>
<td><strong>NA</strong></td>
<td>-0.15</td>
<td>1.16</td>
<td>[0.90, 1.48]</td>
<td></td>
</tr>
<tr>
<td><strong>Loss*NA</strong></td>
<td>0.26</td>
<td>-0.11</td>
<td>0.98</td>
<td>[0.60, 1.34]</td>
</tr>
<tr>
<td><strong>AV</strong></td>
<td>0.04</td>
<td>0.93</td>
<td>[0.81, 1.05]</td>
<td></td>
</tr>
<tr>
<td><strong>Loss*AV</strong></td>
<td>0.37</td>
<td>-0.13</td>
<td>0.88</td>
<td>[0.58, 1.34]</td>
</tr>
<tr>
<td><strong>CSE</strong></td>
<td>0.18**</td>
<td>1.04</td>
<td>[0.90, 1.19]</td>
<td></td>
</tr>
<tr>
<td><strong>Loss*CSE</strong></td>
<td>0.38</td>
<td>1.21</td>
<td>0.25</td>
<td>[0.47, 0.99]</td>
</tr>
</tbody>
</table>

**Instructed Response as the Outcome**

<p>| <strong>Gain</strong> 0.65 | -0.50 | 0.61 | [0.18, 2.10] |
| <strong>PA</strong> | -0.38 | 0.68 | [0.36, 1.30] |
| <strong>Gain*PA</strong> | 0.36 | 0.71 | [0.17, 2.83] |
| <strong>Gain</strong> 0.66 | -0.51 | 0.60 | [0.17, 2.08] |
| <strong>AP</strong> | -1.21** | 0.30 | [0.14, 0.63] |
| <strong>Gain*AP</strong> | 0.08 | 1.18 | 0.01 | [0.43, 0.94] |
| <strong>Gain</strong> 0.68 | -0.51 | 0.60 | [0.17, 2.07] |
| <strong>CSE</strong> | -0.81** | 0.44 | [0.24, 0.82] |
| <strong>Gain*CSE</strong> | 0.04 | 0.92 | [0.47, 2.14] |
| <strong>Loss</strong> 0.38 | 0.32 | 1.37 | [0.50, 3.74] |
| <strong>NA</strong> | 1.41** | 4.12 | [2.16, 7.82] |
| <strong>Loss*NA</strong> | -0.64 | 0.53 | [0.14, 1.94] |</p>
<table>
<thead>
<tr>
<th>Loss</th>
<th>0.42</th>
<th>0.33</th>
<th>1.39 [0.51, 3.79]</th>
<th>2.73(\dagger)</th>
<th>0.36</th>
<th>1.43 [0.52, 3.91]</th>
<th>0.08</th>
<th>0.30</th>
<th>1.35 [0.46, 3.94]</th>
<th>0.27</th>
<th>1.31 [0.93, 1.84]</th>
<th>0.10</th>
<th>1.10 [0.56, 2.16]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV</td>
<td>0.28(\dagger)</td>
<td>1.32 [0.95, 1.84]</td>
<td>0.27</td>
<td>1.31 [0.93, 1.84]</td>
<td>0.10</td>
<td>1.10 [0.56, 2.16]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss*AV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss</td>
<td>0.17</td>
<td>0.22</td>
<td>1.24 [0.44, 3.47]</td>
<td>8.60(\star)</td>
<td>0.44</td>
<td>1.55 [0.54, 4.46]</td>
<td>0.22</td>
<td>0.26</td>
<td>1.30 [0.36, 4.64]</td>
<td>0.10</td>
<td>1.10 [0.56, 2.16]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSE</td>
<td>-0.79(\star)</td>
<td>0.45 [0.26, 0.78]</td>
<td>-0.79 (\star)</td>
<td>0.45 [0.26, 0.78]</td>
<td>-0.26</td>
<td>0.77 [0.26, 2.29]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss*CSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. PA/NA = Positive/Negative Affect, AP = Trait Approach, CSE = Core Self-Evaluations. \(\overset{\cdot}{p} < .10, \overset{\star}{p} < .05, \overset{\star\star}{p} < .01\).
1a. Infrequency as the IER Outcome

![Bar chart showing infrequency as the IER outcome across different warning message conditions.]

1b. *spi* as the IER Outcome

![Bar chart showing *spi* as the IER outcome across different warning message conditions.]

1c. Instructed Responses as the IER Outcome

![Bar chart showing instructed responses as the IER outcome across different warning message conditions.]

Figure 1. Prevalence of IER across conditions and outcomes.
Figure 2. Paradoxically congruent interactions for infrequency IER without (left panes) and with (right panes) IER control.
Figure 3. Differences in personal characteristics across conditions. (∗p < .05).
References


IER IN MTURK: EVIDENCE-BASED QUALITY CONTROL

*Industrial and Organizational Psychology, 8*(2), 142–164. https://doi.org/10.1017/iop.2015.13


Roulin, N. (2015). Don’t throw the baby out with the bathwater: Comparing data quality of crowdsourcing, online panels, and student samples. *Industrial and Organizational Psychology, 8*(02), 190–196. https://doi.org/10.1017/iop.2015.24


Table A1
Hypotheses 4a-f: Congruency across AP-AV constructs and IER detection methods (controlling for scale level IER)

<table>
<thead>
<tr>
<th></th>
<th>Step 1 ( \chi^2 )</th>
<th>Est</th>
<th>OR</th>
<th>95% CI</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infrequency as the Outcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER</td>
<td>2.99†</td>
<td>0.55†    1.73 [0.94, 3.17]</td>
<td>0.94</td>
<td>0.51†    1.72 [0.94, 3.15]</td>
<td>0.36</td>
</tr>
<tr>
<td>Gain</td>
<td>-0.20</td>
<td>0.82 [0.54, 1.23]</td>
<td>-0.22</td>
<td>0.80 [0.53, 1.22]</td>
<td>-0.24</td>
</tr>
<tr>
<td>PA</td>
<td>0.07</td>
<td>1.07 [0.86, 1.34]</td>
<td>0.12</td>
<td>1.12 [0.89, 1.41]</td>
<td></td>
</tr>
<tr>
<td>Gain*PA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER</td>
<td>4.62*</td>
<td>0.97*    2.63 [1.12, 6.15]</td>
<td>0.94</td>
<td>0.97*    2.63 [1.12, 6.16]</td>
<td>0.38</td>
</tr>
<tr>
<td>Gain</td>
<td>-0.21</td>
<td>0.82 [0.54, 1.23]</td>
<td>-0.20</td>
<td>0.82 [0.54, 1.24]</td>
<td>-0.18</td>
</tr>
<tr>
<td>AP</td>
<td>-0.07</td>
<td>0.93 [0.74, 1.17]</td>
<td>-0.07</td>
<td>0.93 [0.73, 1.17]</td>
<td></td>
</tr>
<tr>
<td>Gain*AP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER</td>
<td>4.32*</td>
<td>0.99*    2.69 [1.09, 6.66]</td>
<td>0.81</td>
<td>0.98*    2.65 [1.07, 6.57]</td>
<td>5.53*</td>
</tr>
<tr>
<td>Gain</td>
<td>-0.19</td>
<td>0.83 [0.55, 1.25]</td>
<td>-0.17</td>
<td>0.85 [0.56, 1.28]</td>
<td>-0.13</td>
</tr>
<tr>
<td>CSE</td>
<td>-0.25*</td>
<td>0.78 [0.64, 0.96]</td>
<td>-0.24*</td>
<td>0.79 [0.64, 0.97]</td>
<td></td>
</tr>
<tr>
<td>Gain*CSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screened group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER</td>
<td>8.30**</td>
<td>1.43** 4.16 [1.66, 10.4]</td>
<td>1.00</td>
<td>1.48** 4.41 [1.74, 11.2]</td>
<td>13.2**</td>
</tr>
<tr>
<td>Loss</td>
<td>0.38</td>
<td>1.47 [0.69, 3.12]</td>
<td>0.46</td>
<td>1.59 [0.72, 3.50]</td>
<td>0.57</td>
</tr>
<tr>
<td>NA</td>
<td>0.34** 2.31 [1.46, 3.65]</td>
<td>0.85** 2.33 [1.47, 3.71]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss*NA</td>
<td>-0.57</td>
<td>0.77 [0.22, 1.43]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER</td>
<td>4.06*</td>
<td>1.45* 4.28 [1.14, 16.0]</td>
<td>1.01</td>
<td>1.49* 4.44 [1.18, 16.7]</td>
<td>1.91</td>
</tr>
<tr>
<td>Loss</td>
<td>0.37</td>
<td>1.45 [0.70, 3.01]</td>
<td>0.42</td>
<td>1.53 [0.73, 3.19]</td>
<td>0.51</td>
</tr>
<tr>
<td>AV</td>
<td>0.17</td>
<td>1.19 [0.93, 1.51]</td>
<td>0.19</td>
<td>1.21 [0.94, 1.55]</td>
<td></td>
</tr>
<tr>
<td>Loss*AV</td>
<td>-0.25</td>
<td>0.78 [0.48, 1.27]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER</td>
<td>13.2**</td>
<td>2.10** 8.18 [2.74, 24.4]</td>
<td>0.51</td>
<td>2.08** 8.07 [2.70, 24.2]</td>
<td>1.42</td>
</tr>
<tr>
<td>Loss</td>
<td>0.28</td>
<td>1.32 [0.62, 2.83]</td>
<td>0.35</td>
<td>1.42 [0.65, 3.09]</td>
<td>0.38</td>
</tr>
<tr>
<td>CSE</td>
<td>-0.23</td>
<td>0.79 [0.54, 1.16]</td>
<td>-0.25</td>
<td>0.78 [0.53, 1.15]</td>
<td></td>
</tr>
<tr>
<td>Loss*CSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. PA = Trait Positive Affect, NA = Trait Negative Affect, AP = Trait Approach, NA = Trait Negative Affect, CSE = Core Self-Evaluations, Scale IER = failed any IER items or failed spi on the scale in question 0 = no I = failed any IER on that scale \( p < .10 \), \* \( p < .05 \), \* \* \( p < .01 \).
Table A1 (continued)

<table>
<thead>
<tr>
<th>Non-screened group</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale IER</td>
<td>2.73* 0.74 2.08 [0.90, 4.18]</td>
<td>16.3** 0.66 1.93 [0.80, 4.66]</td>
<td>5.24* 0.71 2.04 [0.84, 4.95]</td>
<td>6.40* 0.60 1.83 [0.74, 4.52]</td>
</tr>
<tr>
<td>Loss</td>
<td>-1.15** 0.32 0.18 [0.57]</td>
<td>-1.22** 0.30 0.16 [0.53]</td>
<td>-1.53** 0.22 0.11 [0.44]</td>
<td>-1.39** 0.25 0.13 [0.48]</td>
</tr>
<tr>
<td>NA</td>
<td>0.39 1.47 [1.05, 2.05]</td>
<td>0.53 1.70 [1.17, 2.47]</td>
<td>0.94 2.57 [1.20, 5.48]</td>
<td>0.45* 1.57 [1.02, 2.42]</td>
</tr>
<tr>
<td>Loss*NA</td>
<td>2.08 0.92 2.51 [0.74, 8.50]</td>
<td>16.9** 0.94 2.57 [0.72, 9.12]</td>
<td>8.00** 0.93 2.54 [0.71, 9.11]</td>
<td>4.46* 0.94 2.56 [0.71, 9.27]</td>
</tr>
<tr>
<td>Loss</td>
<td>-1.25** 0.29 0.16 [0.52]</td>
<td>-1.21** 0.30 0.16 [0.55]</td>
<td>-1.32** 0.27 0.14 [0.51]</td>
<td>-0.16** 0.63 0.45 [0.88]</td>
</tr>
<tr>
<td>AV</td>
<td>0.28 1.32 [1.08, 1.60]</td>
<td>0.34 1.41 [1.14, 1.75]</td>
<td>0.45* 1.57 [1.02, 2.42]</td>
<td>-0.42 0.66 [0.34, 1.28]</td>
</tr>
<tr>
<td>Loss*AV</td>
<td>Scale IER 181** 23.0 ∞ [0, Und]</td>
<td>0.24 23.0 ∞ [0, Und.]</td>
<td>1.15 23.1 ∞ [0, Und.]</td>
<td>0.85 22.2 ∞ [0, Und.]</td>
</tr>
<tr>
<td>Gain</td>
<td>0.13 1.14 [0.67, 1.92]</td>
<td>0.09 1.10 [0.65, 1.85]</td>
<td>0.48 1.05 [0.61, 1.81]</td>
<td>0.18 1.20 [0.89, 1.62]</td>
</tr>
<tr>
<td>PA</td>
<td>0.16 1.18 [0.87, 1.59]</td>
<td>0.18 1.20 [0.89, 1.62]</td>
<td>0.29 1.33 [0.73, 2.44]</td>
<td>-0.42* 0.66 [0.45, 0.96]</td>
</tr>
<tr>
<td>Gain*PA</td>
<td>Scale IER 6.25* 1.63* 5.12 [1.42, 184]</td>
<td>0.03 1.62 5.10 [1.41, 184]</td>
<td>7.68** 1.61 5.05 [1.38, 183]</td>
<td>0.12 0.47 1.61 [0.33, 7.75]</td>
</tr>
<tr>
<td>Gain</td>
<td>-0.04 0.96 [0.64, 1.46]</td>
<td>-0.01 0.99 [0.65, 1.5]</td>
<td>-0.19 0.83 [0.56, 1.23]</td>
<td>-0.18* 0.83 [0.69, 1.01]</td>
</tr>
<tr>
<td>AP</td>
<td>-0.33** 0.72 [0.57, 0.91]</td>
<td>-0.33** 0.72 [0.57, 0.91]</td>
<td>-0.42* 0.66 [0.45, 0.96]</td>
<td>-0.42* 0.66 [0.45, 0.96]</td>
</tr>
<tr>
<td>Gain*AP</td>
<td>Scale IER 32.5** 2.77 15.9 [5.23, 482]</td>
<td>0.13 2.78* 16.1 [5.29, 49.0]</td>
<td>0.02 2.77* 16.0 [5.24, 48.9]</td>
<td>0.18 2.77* 15.9 [5.21, 48.6]</td>
</tr>
<tr>
<td>Gain</td>
<td>0.08 1.08 [0.71, 4.66]</td>
<td>0.08 1.08 [0.71, 4.66]</td>
<td>0.08 1.08 [0.71, 1.67]</td>
<td>0.08 1.08 [0.71, 1.67]</td>
</tr>
<tr>
<td>CSE</td>
<td>-0.01 0.99 [0.80, 1.22]</td>
<td>-0.01 0.99 [0.80, 1.22]</td>
<td>-0.01 0.99 [0.80, 1.22]</td>
<td>-0.01 0.99 [0.80, 1.22]</td>
</tr>
<tr>
<td>Gain*CSE</td>
<td>Scale IER 168** 23.0 ∞ [0, Und]</td>
<td>0.03 23.0 ∞ [0, Und.]</td>
<td>1.56 23.0 ∞ [0, Und.]</td>
<td>0.04 23.0 ∞ [0, Und.]</td>
</tr>
<tr>
<td>Loss</td>
<td>0.05 1.05 [0.62, 1.76]</td>
<td>0.04 1.04 [0.62, 1.74]</td>
<td>0.04 1.04 [0.62, 1.75]</td>
<td>0.04 1.04 [0.62, 1.75]</td>
</tr>
<tr>
<td>NA</td>
<td>0.20 1.23 [0.89, 1.68]</td>
<td>0.20 1.23 [0.89, 1.68]</td>
<td>0.20 1.23 [0.89, 1.68]</td>
<td>0.20 1.23 [0.89, 1.68]</td>
</tr>
<tr>
<td>Loss*NA</td>
<td>Scale IER 6.08* 1.43* 4.16 [1.37, 12.6]</td>
<td>0.32 1.43* 4.19 [1.38, 12.7]</td>
<td>0.05 1.41* 4.12 [1.35, 12.6]</td>
<td>1.15 1.41* 4.09 [1.34, 12.5]</td>
</tr>
<tr>
<td>Loss</td>
<td>0.89 [0.59, 1.34]</td>
<td>-0.11 0.89 [0.59, 1.34]</td>
<td>0.58 0.89 [0.59, 1.34]</td>
<td>0.58 0.89 [0.59, 1.34]</td>
</tr>
<tr>
<td>AV</td>
<td>0.02 1.02 [0.89, 1.17]</td>
<td>0.02 1.02 [0.89, 1.17]</td>
<td>0.82 1.02 [0.88, 1.17]</td>
<td>0.82 1.02 [0.88, 1.17]</td>
</tr>
<tr>
<td>Loss*AV</td>
<td>-0.42 0.66 [0.34, 1.28]</td>
<td>-0.42 0.66 [0.34, 1.28]</td>
<td>-0.42 0.66 [0.34, 1.28]</td>
<td>-0.42 0.66 [0.34, 1.28]</td>
</tr>
<tr>
<td>Table A1 (continued)</td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 3</td>
<td>Model 4</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td><strong>Instructed response as the outcome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER 13.6** 2.44**</td>
<td>11.5 [3.38, 39.0]</td>
<td>0.48</td>
<td>1.08</td>
<td>11.3 [3.32, 38.5]</td>
</tr>
<tr>
<td>Gain -0.44</td>
<td>0.64 [0.18, 2.28]</td>
<td>-0.42</td>
<td>0.33 [0.19, 2.36]</td>
<td>0.63</td>
</tr>
<tr>
<td>PA -0.17</td>
<td>0.84 [0.45, 1.57]</td>
<td>0.83</td>
<td>[0.43, 1.59]</td>
<td></td>
</tr>
<tr>
<td>Gain*PA 0.87</td>
<td>0.24</td>
<td>[3.21]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER 11.5** 2.66**</td>
<td>14.3 [3.84, 52.9]</td>
<td>0.68</td>
<td>2.66</td>
<td>14.4 [3.85, 53.6]</td>
</tr>
<tr>
<td>Gain -0.53</td>
<td>0.59 [0.17, 2.11]</td>
<td>-0.63</td>
<td>0.53 [0.14, 1.99]</td>
<td>0.35</td>
</tr>
<tr>
<td>AP -1.04** 0.35 [0.16, 0.78]</td>
<td>-1.03* 0.36 [0.16, 0.78]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain*AP 0.31</td>
<td>1.38 [0.29, 6.49]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER 18.9** 3.29**</td>
<td>26.8 [7.36, 97.8]</td>
<td>0.37</td>
<td>3.26** 26.1 [7.15, 95.7]</td>
<td>5.98*</td>
</tr>
<tr>
<td>Gain -0.40</td>
<td>0.67 [0.18, 2.47]</td>
<td>-0.24</td>
<td>0.79 [0.21, 2.98]</td>
<td>-0.74</td>
</tr>
<tr>
<td>CSE -0.88</td>
<td>0.42 [0.20, 0.87]</td>
<td>-0.95</td>
<td>0.39 [0.18, 0.84]</td>
<td></td>
</tr>
<tr>
<td>Gain*CSE -0.60</td>
<td>0.55 [0.12, 2.62]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IER 17.8** 2.39**</td>
<td>10.9 [3.89, 30.5]</td>
<td>1.10</td>
<td>2.47** 11.8 [0.63, 6.45]</td>
<td>19.6**</td>
</tr>
<tr>
<td>Loss 0.56</td>
<td>1.75 [0.52, 3.91]</td>
<td>0.54</td>
<td>1.71 [0.56, 5.22]</td>
<td>1.34</td>
</tr>
<tr>
<td>NA 1.34</td>
<td>3.83 [2.00, 7.39]</td>
<td>1.42</td>
<td>4.14 [1.10, 8.14]</td>
<td></td>
</tr>
<tr>
<td>Loss*NA -0.91</td>
<td>0.40 [0.11, 1.54]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER 19.9** 2.98**</td>
<td>19.7 [6.34, 61.4]</td>
<td>0.71</td>
<td>3.02** 20.5 [6.52, 64.5]</td>
<td>3.65*</td>
</tr>
<tr>
<td>Loss 0.45</td>
<td>1.57 [0.55, 4.54]</td>
<td>0.48</td>
<td>1.61 [0.56, 4.68]</td>
<td>0.42</td>
</tr>
<tr>
<td>AV 0.37*</td>
<td>1.44 [0.99, 2.13]</td>
<td>0.36*</td>
<td>1.44 [0.86, 1.93]</td>
<td></td>
</tr>
<tr>
<td>Loss*AV 0.09</td>
<td>1.10 [0.47, 2.36]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale IER 24.0** 3.13**</td>
<td>22.9 [7.50, 70.1]</td>
<td>0.09</td>
<td>3.13** 22.8 [8.40, 170]</td>
<td>5.53*</td>
</tr>
<tr>
<td>Loss 0.16</td>
<td>1.18 [0.58, 4.85]</td>
<td>0.33</td>
<td>1.39 [0.45, 4.24]</td>
<td>0.25</td>
</tr>
<tr>
<td>CSE -0.72**</td>
<td>0.49 [0.26, 0.91]</td>
<td>-0.71*</td>
<td>0.26 [0.26, 0.92]</td>
<td></td>
</tr>
<tr>
<td>Loss*CSE -0.12</td>
<td>0.25 [0.25, 3.09]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. PA = Trait Positive Affect, NA = Trait Negative Affect, AP = Trait Approach, NA = Trait Negative Affect, CSE = Core Self-Evaluations. Scale IER = failed any IER items or failed any of the scale in question 0 = no 1 = failed any IER on that scale. *p < .10, **p < .05, ***p < .01.