J Bus Psychol (2015) 30:299–311
DOI 10.1007/s10869-014-9357-6

Detecting Insufficient Effort Responding with an Infrequency Scale: Evaluating Validity and Participant Reactions

Jason L. Huang · Nathan A. Bowling · Mengqiao Liu · Yuhui Li

Published online: 12 April 2014

Abstract

Purpose Insufficient effort responding (IER), which occurs due to a lack of motivation to comply with survey instructions and to correctly interpret item content, represents a serious problem for researchers and practitioners who employ survey methodology (Huang et al. 2012). Extending prior research, we examine the validity of the infrequency approach to detecting IER and assess participant reactions to such an approach.

Design/Methodology/Approach Two online surveys (*Studies 1* and *2*) completed by employed undergraduates were utilized to assess the validity of the infrequency approach. An on-line survey of paid participants (*Study 3*) and a paper-and-pencil survey in an organization (*Study 4*) were conducted to evaluate participant reactions, using random assignment into survey conditions that either did or did not contain infrequency items.

Findings *Studies 1* and *2* provided evidence for the reliability, unidimensionality, and criterion-related validity of the infrequency scales. *Study 3* and *Study 4* showed that surveys that contained infrequency items did not lead to more negative reactions than did surveys that did not contain such items.

Implications The current findings provide evidence of the effectiveness and feasibility of the infrequency approach for detecting IER, supporting its application in low-stakes organizational survey contexts.

Originality/Value The current studies provide a more in-depth examination of the infrequency approach to IER detection than had been done in prior research. In particular, the evaluation of participant reactions to infrequency scales represents a novel contribution to the IER literature.

Keywords Insufficient effort responding · Careless responding · Random responding · Inconsistent responding · Data screening · Online surveys

Introduction

Applied survey researchers are often faced with the challenge of poor data quality arising from inattentive or unmotivated survey respondents (e.g., Behrend et al. 2011; Hough et al. 1990). The concern over participants' attentiveness and motivation is likely heightened as survey research has moved from traditional paper-and-pencil format to the Internet (Johnson 2005). The current research focuses on *Insufficient Effort Responding* (IER), "a response set in which the respondent answers a survey measure with low or little motivation to comply with the survey instructions, correctly interpret item content, and provide accurate responses" (Huang et al. 2012, p. 100).

An earlier version of this manuscript was presented at the annual conference of the Society for Industrial and Organizational Psychology in San Diego, CA in 2012.

J. L. Huang (✉) · M. Liu
Department of Psychology, Wayne State University, 5057
Woodward Ave, Detroit, MI 48202, USA
e-mail: jasonhuang@wayne.edu

M. Liu
e-mail: mengqiao.liu@wayne.edu

N. A. Bowling
Department of Psychology, Wright State University, Dayton,
OH, USA
e-mail: nathan.bowling@wright.edu

Y. Li
School of Labor and Human Resources, Renmin University,
Beijing, China
e-mail: yuhui_li@ruc.edu.cn

The definition of IER is inclusive: It does not specify the pattern of the IER, nor the underlying reason for IER's occurrence. IER may manifest itself as either random responding (Hough et al. 1990) or as non-random response patterns (Costa and McCrae 2008; DiLalla and Dollinger 2006), and it encompasses both unintentional, occasional careless responding (Schmitt and Stults 1985), and intentional "speeding-through" of survey items (Huang et al. 2012).

As IER's potentially undesirable effects (Huang et al. 2012, 2014) are often ignored by researchers and practitioners in industrial/organizational psychology (Liu et al. 2013), there is a need to better measure IER across various types of surveys. In the current paper we focus on the infrequency approach as a particularly promising means of assessing IER. As described in more detail below, the infrequency approach utilizes counterfactual or improbable items that have clear correct responses and it equates incorrect responses with IER (Huang et al. 2012). Unlike some approaches to IER measurement that can be cumbersome to implement (e.g., psychometric antonym) or are format-dependent (e.g., long strings; see Huang et al. 2012), the infrequency approach holds promise as a simple method to detect IER applicable across different surveys.

Despite the potential ease in designing and implementing, mixed evidence exists for the infrequency approach's effectiveness in detecting IER. Huang et al. (2012) noted that the infrequency approach could be confounded with social desirability. Indeed, the same approach has been used to measure socially desirable responding (Pannone 1984; Paulhus et al. 2003). Conversely, Meade and Craig (2012) offered some evidence that an infrequency scale can capture IER. Thus, additional research on the approach is needed to further ascertain its effectiveness in a generic low-stakes survey context. Further, although respondents' negative reactions to the counterfactual or improbable items may be a potential concern, no studies thus far have assessed reactions to the infrequency items.

The goal of our paper is to examine the infrequency approach to detecting IER with two primary foci. First, extending Meade and Craig's (2012) work, we provide further construct and criterion-related validity evidence for the infrequency approach. Second, we investigate participant reactions to infrequency items. This second goal is of particular importance, because it speaks to the frequency approach's feasibility within applied survey contexts.

In the sections below, we briefly review different approaches to IER detection, discuss the rationale behind the infrequency approach, and present specific considerations for evaluating the infrequency approach.

IER Detection and the Infrequency Approach

Researchers have employed various methods to detect IER (see Huang et al. 2012 for a review). First, the *inconsistency approach* assumes that inconsistent response to items assessing the same construct is indicative of IER. For example, a participant who strongly agrees with both the statements "I love my job" and "I hate my job" would be identified as engaging in IER according to this approach (see Scandell 2000). *The response pattern approach* detects IER when a participant provides a suspicious pattern of responses, such as repeated selection of the same response option for a large number of consecutive items (Costa and McCrae 2008). *The response time approach* assumes that providing valid answers to survey questions requires a minimal amount of time to complete the measure, and thus overly fast survey completion times are indicative of IER. Although these three approaches can effectively detect IER (Huang et al. 2012; Meade and Craig 2012), they tend to necessitate complex analysis, a large number of survey items, and/or particular survey formats. The inconsistency approach, for example, generally requires a large number of items on the survey (Huang et al. 2012). Two less complex alternatives—*self-report of survey effort* (e.g., "I carefully read every survey item"; Meade and Craig 2012) and *instructed response items* (e.g., "please respond with *Somewhat Agree* for this item")—have received mixed empirical support (Huang et al. 2012; Meade and Craig 2012).

Unlike the methods above, the infrequency approach may satisfy the need for a simple yet effective means to assess IER in applied survey situations. *The infrequency approach* uses items on which all or virtually all attentive participants should provide the same response (e.g., Beach 1989; Green and Veres 1990). In other words, these items have one highly probable response, and deviations from the probable response are assumed to reflect IER. For example, Green and Stutzman (1986) embedded task statements that were clearly unrelated to a focal job in a job analysis inventory administered to incumbents, and Beach (1989) asked participants such questions as "I was born on February 30th." Each of these items has one clear correct answer, and the presence of IER is inferred when a participant selects improbable response options. A participant who incorrectly answers many infrequency items is presumed to display high levels of IER when responding to substantive survey items in the same questionnaire.

Despite some evidence that infrequency items can detect *instructed random* responses, Huang et al. (2012) cautioned that the infrequency approach could be confounded with impression management and faking. Indeed, the same approach has been applied to assess socially desirable responding. For example, having applicants rating prior work experience on job relevant tasks in a high-stake selection context, Pannone (1984) inserted an item involving a piece of nonexistent equipment, which was used to detect faking (and not IER). Similarly, in an attempt to measure self-enhancement response bias, Paulhus et al. (2003) instructed participants to rate their familiarity with a number of objects, 20 % of which were nonexistent and specifically included to assess self-enhancement. If endorsing counterfactual or nonexistent items can be an indicator of self-enhancement, then using these items to detect IER can be problematic. Thus, to use an infrequency scale to detect IER, one has to show empirical evidence that items on the scale do not inadvertently assess social desirability.

Perhaps making matters more challenging, past applications of the infrequency approach were either used in specific measurement contexts such as task analysis (e.g., Green and Stutzman 1986; Green and Veres 1990) and clinical personality assessment (e.g., Baer et al. 1997), or developed as part of a proprietary inventory (e.g., Hogan and Hogan 2007; Hough et al. 1990; Jackson 1974). As a result, there has yet to be comprehensive evaluations of the infrequency approach as an IER detection method for general survey contexts.

Recent work by Meade and Craig (2012) showed that the infrequency approach is promising for detecting IER. In a personality questionnaire administered to a sample of undergraduate psychology students, the authors embedded nine items with clear correct answers (e.g., “All my friends say I would make a great poodle”; “I am enrolled in a Psychology course currently”). The authors used the sum of those items to indicate lack of attention to the survey. Meade and Craig found that this scale loaded on the same underlying factor as the more complex inconsistency indices, while not being confounded with socially desirable responding.

Present Studies

Meade and Craig’s (2012) promising results led us to ask whether an infrequency detection scale can be developed in a generic survey context and satisfy rigorous psychometric evaluation. Answering this question is important because survey researchers and administrators may wish to develop their own detection scales to maximize similarity in item content and format to their particular surveys. The current research question can be addressed in three ways, none of which has been done in the literature thus far. First, as an individual item may capture variance due to error (e.g., lack of understanding of a particular word, mistaken selection of a response option) and unintended construct (e.g., conscientiousness, social desirability; see Gorsuch 1997), any individual item that comprise a scale for IER needs to be examined in terms of contribution to the underlying IER construct. In other words, an infrequency scale for IER should consist of items that are *internally consistent* and *unidimensional*, neither of which has been examined in the existing literature. Second, such a scale should not only exhibit *convergent validity* with other IER indices as shown in Meade and Craig (2012), but also should predict behavioral outcomes of IER (i.e., *criterion-related validity*). Lastly, considering practical survey implications, the use of such a detection scale should not result in respondents’ negative *reactions* to the survey. Respondent reactions are of particular concern with infrequency measures because in many instances their items are odd or even humorous (e.g., “I am paid biweekly by leprechauns”; Meade and Craig 2012). It is possible that otherwise attentive participants may lose confidence in the legitimacy of the questionnaire after being exposed to unusual infrequency items.[1]

We conducted four studies that examine the validity and feasibility of the infrequency approach. In *Study 1*, we evaluated an infrequency IER scale’s reliability and unidimensionality, as well as the scale’s convergent validity evidence against three other IER indices. In *Study 2*, we assessed an infrequency IER scale’s criterion-related validity, focusing on the scale’s relationship with survey completion time and length of participants’ response to an open-ended question. In *Study 3*, we compared respondent reactions to an infrequency IER scale against social desirability and impression management scales (Paulhus, 1991). In *Study 4*, we examined participant reactions to the infrequency approach within an organizational sample.

Study 1—Method

Study 1 was conducted for initial scale validation. To be considered a valid measure of the underlying IER construct, the scale should exhibit unidimensionality, high internal consistency, and high convergent validity with other IER indices. Three other IER indices were used in this study: the psychometric antonym IER index, the individual reliability index, and self-reported IER (see the “Measures” section below).

Participants

Part-time and full-time working adults enrolled in undergraduate psychology courses at a large suburban public university in the Midwestern U.S. participated in *Study 1*. Participants completed an online survey in return for extra course credits. The survey included nine scales measuring

[1] We thank three anonymous reviewers and Action Editor Scott Tonidandel for noting this potential concern that led us to conduct *Study 3* and *Study 4*.

variables such as personality, general health, and life and job satisfaction, with the IER items interspersed throughout the 105-item survey. The sample consisted of 284 participants (64 % females, 36 % males; average age = 20, SD = 3). Participants worked an average of 22 h each week (SD = 10). Racial and ethnic information was not collected for this study.

Measures

Infrequency IER Scale

The goal of item generation was to produce a set of items for which most, if not all, attentive respondents would endorse the correct responses. Thus, we developed a set of eight items using counterfactual statements, deviation from "common sense," and improbable events. We varied the eight items on the degree to which they pertain to the work context so that the item generation process could be applicable to a wide range of applied surveys. Furthermore, we considered the social desirability of the item content such that endorsement of the items would not be dictated by the desire to portray oneself in a positive light[2] (cf. Pannone 1984; Paulhus et al. 2003). An example item was, "I work twenty-eight hours in a typical work day." The items (see Table 1) were administered on a 7-point Likert scale (1 = *Strongly Disagree*; 7 = *Strongly Agree*).

We scored the infrequency items such that disagreement with a false statement in any way (*Slightly Disagree*/*Somewhat Disagree*/*Strongly Disagree*) was coded as attentive responding (i.e., non-IER) while agreement was coded as IER (IER = 1, attentive = 0). Dichotomization of responses was appropriate here because the infrequency items resemble items on an ability test with *correct* versus *incorrect* answers, and there is no degree of correctness as the Likert scale might suggest. For example, for "I eat cement occasionally," failure to disagree in any way would be considered IER, while the difference from *Neutral* to *Strongly Disagree* simply reflects random error.

Psychometric Antonym IER Index

Based on the inconsistency approach to IER detection, psychometric antonyms were computed by first selecting 30 pairs of items with the highest negative inter-item correlations in the sample (see Johnson 2005). A within-person correlation was calculated for each respondent based on the 30 pairs of items. Assuming that the highly negative correlation would be observed among normal respondents, respondents displaying insufficient effort would have relatively more positive within-person correlations. The resulting within-person correlations were then used as the psychometric antonym IER index, with higher scores indicating higher probability of IER behavior. Unlike previous studies (Huang et al. 2012; Johnson 2005; Meade and Craig 2012), we chose not to reverse score these within-person correlations so that the index is in the same direction as the IER construct.

Individual Unreliability IER Index

Based also on the inconsistency approach, individual reliability assumes attentive respondents' responses to two halves of the same measure should be positively related (see Johnson 2005). Half-scale scores were computed for odd- and even-numbered items on each substantive scale. A within-person correlation was obtained for each respondent across corresponding odd and even half-scale scores. The within-person correlation was used to indicate individual reliability, which was then reverse-scored as the individual unreliability IER index in the same direction as the IER construct.

Single Self-Report IER Item

We included a single-item measure, "I have paid no attention to this survey so far," near the end of the survey to capture participants' self-report IER (see Meade and Craig 2012). The item was rated on the same 7-point Likert scale, with higher scores suggesting likely IER behavior.

Study 1—Results and Discussion

To assess the unidimensionality of the IER items, we conducted a confirmatory factor analysis in MPlus 6.11 (Muthén and Muthén 2011). Specifically, all eight dichotomous IER items were regressed onto a single latent factor using probit regression with robust weighted least square estimation (Muthén et al. 1997). The model yielded very good fit to the

[2] In a pilot study based on 59 part-time and full time working undergraduates, the eight-item infrequency scale was not significantly correlated with Paulhus's (Paulhus 1991) impression management scale ($r = -0.20$, $p =$ ns) or self-deception scale ($r = -0.20$, $p =$ ns). If the infrequency items assessed faking (cf. Pannone 1984; Paulhus et al. 2003), the scale should be *positively* related to impression management and self-deception. Thus, these nonsignificant *negative* correlations indicate that the infrequency items measured response behavior distinct from impression management and self-deception. As suggested by an anonymous reviewer, we further examined each infrequency item's correlations with impression management and self-deception. Five items (items #1, 2, 3, 5, and 7) had negative correlations (rs ranging from −0.23 to −0.36, ps < 0.10), while two items (items #4 and #8) had near zero correlations (rs ranging from −0.05 to 0.06, ps = ns) with the two social desirability measures. Item #6 stood out for having weak positive associations with both self-deception and impression management (rs = 0.21 and 0.11, ps = ns). Overall, however, none of the infrequency IER items was saturated with social desirability, thus alleviating the concern that these items measured socially desirable responding.

Table 1 Confirmatory factor analysis for the eight-item IER Scale for *Study 1*

	M	SD	Loading
1. I can run 2 miles in 2 min	0.16	0.36	0.81***
2. I eat cement occasionally	0.11	0.31	0.97***
3. I can teleport across time and space	0.16	0.37	0.93***
4. I am interested in pursuing a degree in parabanjology	0.16	0.36	0.87***
5. I have never used a computer	0.10	0.29	0.96***
6. I work fourteen months in a year	0.15	0.35	0.82***
7. I will be punished for meeting the requirements of my job	0.17	0.37	0.80***
8. I work twenty-eight hours in a typical work day	0.02	0.13	0.44*

$N = 284$

* $p < 0.05$; *** $p < 0.001$. Standardized loadings are shown

data, $\chi^2(20) = 44.88$, $p = 0.001$; CFI = 0.99, TLI = 0.98; RMSEA = 0.066. All item loadings were significant (Table 1). The confirmatory factor analysis provided support for the unidimensionality of the eight IER items. Scale scores were computed as the mean of the eight dichotomous items ($M = 0.12$, SD = 0.23). Cronbach's α for the scale was 0.85.

Next, we examined the convergent validity evidence for the infrequency approach. As expected, the infrequency scale yielded significantly (all *p*s < 0.001) positive correlations with the psychometric antonym IER index ($r = 0.58$), the individual unreliability index ($r = 0.50$), and the single-item measure of participant attention ($r = 0.62$). Further, an exploratory factor analysis on these four IER indices resulted in a clear one-factor solution, accounting for 65 % of observed variance. While all indices loaded strongly (>0.63) on the underlying IER factor, the highest loading (0.81) came from the infrequency scale.

As a whole, *Study 1* provided evidence for the construct validity of the infrequency measure. Items based on the infrequency approach were found to be reliable and unidimensional, while the infrequency scale showed convergent validity with the other three IER indices. In *Study 2*, we focus on criterion-related validity for a shortened infrequency scale.

Study 2—Method

Study 2 was conducted to examine the criterion-related validity of the infrequency approach. Specifically, we examined the correlation between an infrequency scale and two objective manifestations of respondent effort: survey completion time and length of response to an open-ended question.

Participants

Participants in *Study 2* consisted of 133 part-time working adults enrolled in undergraduate psychology courses at a second large suburban public university in the Midwestern U.S. Participants were on average 20 years old (SD = 2), with a majority of males (55 %) and Whites (80 %). The survey consisted of 70 personality items administered online.

Measures

Shortened Infrequency Scale

Due to the brevity of the survey (70 items), we only embedded three infrequency IER items: "I eat cement occasionally," "I can teleport across time and space," and "I have never used a computer." We adopted these three items because they were similar in item context (i.e., non-work context) and length to the substantive personality items. Similar to the substantive items in the online survey, the three IER items were administered on a 6-point Likert scale, ranging from 1 (*Very Inaccurate*) to 6 (*Very Accurate*). Responses indicating that an IER item was accurate in self-description (*Slightly Accurate*/*Somewhat Accurate*/*Very Accurate*) were coded as IER (IER = 1, attentive = 0). The scale was calculated as the mean on the three items ($M = 0.13$, SD = 0.27, Cronbach's α = 0.75). It should be noted that within the *Study 1* dataset, the shortened infrequency scale was internally consistent (α = 0.79) and strongly correlated with the remaining five infrequency items ($r = 0.74$, $p < 0.001$).

Indicators of Response Effort

We included two objective behavioral outcomes of response effort: (a) total survey time and (b) number of letters typed in an open-ended question about one's job description (letters typed). *Total survey time* was the number of minutes elapsed from the initiation to the completion of the survey. Rapid responding can indicate IER (Huang et al. 2012; Meade and Craig 2012). Thus, the total minutes for survey completion was recorded as an objective albeit imperfect manifestation of participants' effort ($M = 10.62$, $SD = 7.20$). We applied a natural log transformation on this time measure before the analysis.

The second indicator of response effort, *letters typed* was obtained from an open-ended question about respondents' current jobs. Responses varied in length, ranging from simple listing of job titles (e.g., "server," "sales associate"), to somewhat specific descriptions of jobs (e.g., "information desk person at a museum," "sales associate for a small tshirt (sic.) company"), to short sentences describing the jobs (e.g., "Help desk at library. Work the cashier/check books out to patrons," "I work for a financial company and help clients understand and correctly fill out their financial portfolios"). The number of letters typed in response to this open-ended question was recorded to indicate respondents' effort ($M = 19.45$, $SD = 18.94$). As letters typed is a count variable, we performed a square root transformation on this variable prior to the analysis.

Study 2—Results and Discussion

The shortened infrequency scale was significantly correlated with total survey time ($r = -0.30$, $p < 0.001$) and letters typed ($r = -0.19$, $p < 0.05$), while the two objective outcomes of response effort were positively associated ($r = 0.26$, $p < 0.01$). Thus, respondents who tended to endorse the ostensibly improbable statements in the shortened infrequency scale also tended to spend less time completing the survey and to type relatively fewer letters when describing their jobs. *Study 2* thus provided support for the criterion-related validity of the IER infrequency measure.

Results from *Studies 1* and *2* contribute beyond Meade and Craig's (2012) research to offer a more comprehensive evaluation of the infrequency approach to IER detection. Beyond validity, however, the feasibility of the infrequency approach also depends on respondents' reactions, due to potential concern that such a scale may result in negative reactions. In *Study 3* we compared the eight-item infrequency scale against items from the self-deception and impression management scales (Paulhus, 1991). These latter two scales were selected as comparison to the infrequency scale because they have been widely applied in survey studies to capture socially desirable response sets.

Study 3—Method

Participants

The *Study 3* participants were recruited from Mechanical Turk (MTurk). MTurk, a crowdsourcing service that offers cost-effective access to a large number of respondents online (Behrend et al. 2011), has been shown to provide high-quality data that are comparable to those from other samples (Buhrmester et al. 2011; Goodman et al. 2013). In the current study, we restricted the sample to adults from the US. A total of 98 people were surveyed, with a majority being female (61 %) and White (76 %) with an average age of 30 ($SD = 9$). The sample size was determined with an a priori power estimate based on 70 % power for a hypothetical medium effect size (Cohen's d of 0.50).

Procedure

Participants were told that the study was designed to evaluate reactions toward a newly developed survey measure. Upon completion of the study, each respondent received $0.50.

Respondents were randomly assigned into two survey conditions ($n = 49$ each). The control group completed all 40 items for self-deception and impression management in their original order. In contrast, the experimental group completed a random subset of 32 of the 40 self-deception and impression management items in the same order and eight infrequency IER items. Specifically, for each experimental group respondent, a random subset of eight items was replaced by eight infrequency items. Upon completion of their respective 40 items, participants in both groups responded to a set of reaction measures.

Reaction Measures

We measured reactions to the survey with three scales from Croteau et al. (2010): *enjoyment* (3 items, $\alpha = 0.94$), *ease of responding* (3 items, $\alpha = 0.76$), and *intention to respond* to a similar future survey (2 items, $\alpha = 0.94$). Responses were made on a 7-point Likert scale ranging from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). Sample items include: "The survey was enjoyable to fill out" (enjoyment), "The survey was clear and understandable" (ease), and "Assuming I have access to a similar survey, I intend to respond to it" (intention).

Study 3—Results and Discussion

We first evaluated whether the three respondent reaction scales reflected the underlying constructs using confirmatory factor analysis in MPlus 6.11 (Muthén and Muthén 2011). A three-factor model provided reasonable fit to the data, $\chi^2(17) = 27.35$, $p = 0.05$; CFI = 0.98, TLI = 0.97; RMSEA = 0.079. All items loaded significantly on their respective intended constructs, with an average standardized loading of 0.86. The three factors were moderately correlated, with latent correlations ranging from 0.42 to

0.49. Thus, the confirmatory factor analysis suggests that the items measured the intended constructs.

We used independent samples t tests to examine whether the experimental and control groups reacted differently in terms of enjoyment, ease, and intention. After responding to their respective 40-item surveys, the experimental group reported slightly higher enjoyment (experimental group: $M = 5.20$, $SD = 1.43$; control group: $M = 4.88$, $SD = 1.48$; $t(96) = 1.11$, $p =$ ns, Cohen's $d = 0.22$) and slightly lower ease of responding (experimental group: $M = 5.63$, $SD = 1.21$; control group: $M = 5.93$, $SD = 0.97$; $t(96) = 1.38$, $p =$ ns, Cohen's $d = - 0.28$). However, neither difference was statistically significant. Finally, the two groups reported the same levels of intention to respond to a similar future survey (experimental group: $M = 6.02$, $SD = 1.12$; control group: $M = 6.02$, $SD = 0.94$; $t(96) = 0.00$, $p =$ ns, Cohen's $d = 0.00$).

Taken together, results from *Study 3* suggest that participants exposed to the infrequency IER items did not react more negatively than did participants only exposed to the widely used socially desirable responding items. The absence of negative reactions observed in *Study 3* may be attributed to how we created these infrequences items: In designing the infrequency items we considered how attentive respondents may react to the item content and deliberately limited potential confounding factors such as social desirability and offensive content.

Although *Study 3*'s results revealed no significant negative reactions to the eight infrequency items, the use of MTurk respondents may limit generalization of the results to working adults. Thus, an investigation in an organizational survey context is warranted. Informed by *Study 3*, we expect that the presence of properly designed infrequency items within an organizational survey will not produce negative respondent reactions.

Given the potential concern over negative reactions to the detection items, the survey designer may consider the option of forewarning respondents about the existence of detection methods to both deter potential IER (see Huang et al. 2012) and to circumvent potential adverse reactions. We expect an organizational survey with an introduction containing a benign warning and explanation to result in more positive reactions than an otherwise identical survey without the warning and explanation. That is, we assume that most participants will empathize with the survey administrator's desire to obtain accurate data. When informed of efforts to screen for IER, attentive participants may view their survey completion as more meaningful (i.e., data will be put to good use; see Hackman and Oldham 1975) and thus feel more positive toward the survey. In addition, offering explanation about the planned screening of IER is consistent with organizational justice research that shows the benefit of information justice (e.g., Colquitt et al. 2001).

Finally, we explore the interactive effect between infrequency items and warning. Given the presence of infrequency items is consistent with the warning information, we expect a survey with both infrequency items and benign warning will result in the most favorable reactions.

Study 4—Method

Study 4 used survey data collected from workers employed by an IT company in China. The survey was conducted as a means for the company to obtain feedback regarding their employees' attitudes, perceptions, and behaviors. As an addendum to the main survey, we designed *Study 4* to investigate the feasibility of the infrequency approach, using a 2 (Infrequency items: absent vs. present) × 2 (Warning: absent vs. present) between subjects factorial design.

Participants

A randomly selected subset of 240 of the company's employees were invited to respond to the anonymous study questionnaire. Different versions of the paper-and-pencil survey booklets were identical except for the instructions provided to participants (see below). Employees received randomly distributed questionnaires (60 booklets per condition) together with stamped, postmarked envelopes.

A total of 157 (65 %) completed questionnaires were returned, with roughly equal numbers of respondents in each of four conditions: (a) no IER items, no warning ($n = 40$); (b) IER items, no warning ($n = 37$); (c) no IER items, warning ($n = 36$); and (d) IER items, warning ($n = 44$). Seventy-six percent of respondents were female; the majority held Bachelor's degree (59 %) or higher (38 %). Although age was not measured, company tenure was assessed: 27 % of respondents had worked in the company for 1–2 years, 26 % had worked for 3–5 years, 29 % had worked 6–10 years, and 18 % had worked for 11–20 years.

Procedure

At the end of the overall introduction to the survey, participants in the *Warning condition* were provided with an additional paragraph that stated: "In our past survey work, we found that a very small number of respondents answered the questions carelessly. In this survey, we employed several methods to assess whether a respondent answered the questions carefully, so as to ensure the quality of survey data."

In the infrequency items (referred to as Items hereafter) condition, five infrequency items were scattered throughout the survey. The items were slightly modified to fit in the other items they were inserted to: "Work less than twenty-eight

Table 2 Descriptive statistics and intercorrelations for *Study 4* variables

	1	2	3	4	5	6	7
1. IER items							
2. Warning	0.07						
3. Enjoyment	0.11	0.22					
4. Ease of responding	0.10	0.14	0.55				
5. Intention	−0.01	−0.01	0.50	0.26			
6. Perceived data quality	0.16	0.04	0.51	0.19	0.33		
7. Perceived design quality	−0.05	0.30	0.46	0.47	0.33	0.29	
M	0.52	0.51	4.08	4.77	5.09	5.04	4.43
SD	0.50	0.50	1.40	1.29	1.31	1.27	1.12

$N = 157$. When $r \geq 0.16$, $p < 0.05$; when $r \geq 0.21$, $p < 0.01$; when $r > 0.26$, $p < 0.001$

hours each day"; "Work more than fourteen months each year"; "I will be punished for meeting the requirements of my job in this organization"; "Do not possess the skills to teleport across time and space"; and "Eat cement occasionally." Jason Huang, who is also a certified English-Chinese translator, translated the focal measures and instructions into Chinese.

At the end of the study, participants were asked to report their perceptions and attitudes regarding the design of the questionnaire they had just completed.

Reaction Measures

The three scales for enjoyment ($\alpha = 0.95$), ease ($\alpha = 0.75$), and intention ($\alpha = 0.98$) from *Study 3* were used in *Study 4* to assess participant reactions. In addition, we were interested in what respondents thought about the quality of the survey data (*perceived data quality*) and the designers of the survey (*perceived design quality*).[3] A literature search did not reveal any previously validated scales that assess these constructs. Thus, we wrote items for the perceived data quality (two items; e.g., "Overall, I believe the data collected from this survey project will be valid"; $\alpha = 0.86$) and perceived design of the survey (five items; e.g., "The designers of this questionnaire appear to have solid expertise in survey design"; $\alpha = 0.85$). All scales were administered on a 7-point Likert scale (1 = *Strongly Disagree*; 7 = *Strongly Agree*). An exploratory factor analysis supported a two-factor solution (70 % observed variance explained) for these seven new items, with each item clearly loading on its intended factor. Average loading was 0.87 for perceived data quality and 0.73 for perceived design quality.

Study 4—Results and Discussion

Table 2 presents the descriptive statistics and intercorrelations for *Study 4* variables. The research question in *Study 4* pertains to applicant reactions toward the survey with and without the infrequency items and benign warning. Due to slight difference in return rates across conditions in the field experiment, we checked the orthogonality of the manipulations prior to hypothesis testing: Items and Warning shared a weak correlation ($r = 0.07$, $p =$ ns), indicating the intended orthogonal manipulations were achieved. Further, neither Items nor Warning had an effect on return rate, $\chi^2(1) = 0.16$ and 0.06, respectively, $p =$ ns, suggesting the manipulations did not significantly influent respondents' decision to complete and return the surveys.

Given that the five scales for respondents' reactions were correlated moderately to highly (*r*s range from 0.19 to 0.55), MANOVA was appropriate for analyzing the effect of Items and Warning on reactions. Results from a 2 (Items: absent vs. present) × 2 (Warning: absent vs. present) MANOVA on these five reaction variables revealed no significant main effect for Items (Wilk's Lambda = 0.93, $F[5, 149] = 2.24$, $p = 0.05$), a significant main effect for Warning (Wilk's Lambda = 0.86, $F[5, 149] = 4.98$, $p < 0.001$), and a significant Items × Warning interaction (Wilk's Lambda = 0.88, $F[5, 149] = 4.19$, $p = 0.001$). In other words, the benign warning had a significant impact on respondents' reactions in general, whereas the frequency items did not. In addition, the benign warning and the infrequency items also interacted to exert an effect on how respondents reacted to the survey.

Given the significant results from MANOVA, we conducted follow-up ANOVAs to examine Warning's main effect and Items × Warning interaction effect (see Table 3). Results showed that warning had a significant positive impact on enjoyment and perceived design of the study, such that respondents presented with survey instruction containing benign warning rated the survey as more enjoyable and of higher design quality. The main effect of warning on perceived design quality was qualified by an Items × Warning interaction (see Fig. 1). Follow-up simple effects analysis revealed a positive effect of warning on perceived design quality in the presence of IER items, *F*(1,

[3] We thank Scott Tonidandel for bringing these issues to our attention.

Table 3 Descriptive statistics for each condition in *Study 4*

	No warning		Warning		ANOVA		
	M	SD	*M*	SD	Items	Warning	Items × warning
Enjoyment							
No IER items	3.58	1.26	4.29	1.34	1.60	7.70**	0.22
IER items	3.95	1.25	4.46	1.58			
Ease of responding							
No IER items	4.65	1.41	4.62	1.30	1.21	2.49	2.97
IER items	4.52	1.17	5.20	1.21			
Intention							
No IER items	4.85	1.48	5.39	1.29	0.02	0.00	6.87**
IER items	5.36	0.71	4.82	1.48			
Perceived data quality							
No IER items	4.69	1.30	4.96	1.39	5.48*	0.07	1.29
IER items	5.36	0.92	5.19	1.20			
Perceived design quality							
No IER items	4.34	0.73	4.65	1.29	0.93	15.13***	4.25*
IER items	3.83	1.21	4.84	0.96			

$N = 157$

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

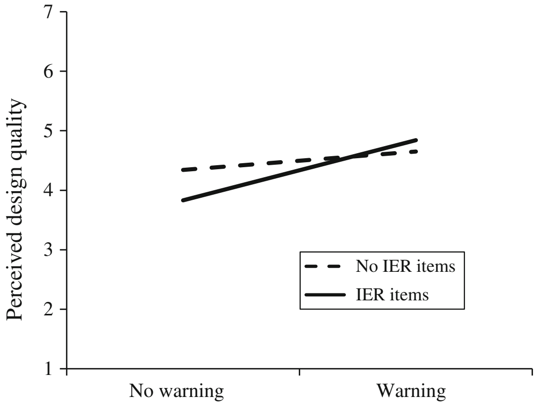

Fig. 1 The interaction effect between items and warning on perceived design quality

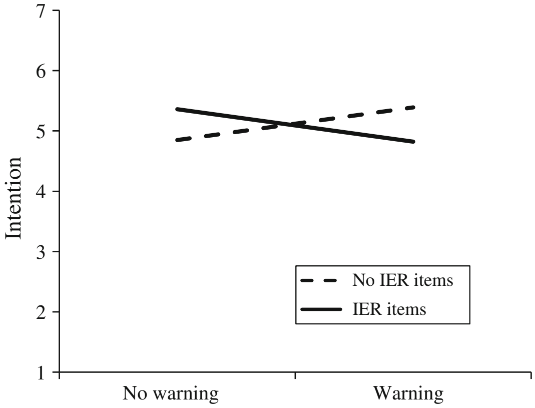

Fig. 2 The interaction effect between items and warning on intention to use

153) = 18.25, MSE = 1.12, $p < 0.001$. However, in the absence of IER items, warning did not have a significant effect, $F(1, 153) = 1.62$, MSE = 1.12, p = ns.

The interaction effect between Items and Warning was also significant for intention to use (see Fig. 2), although simple effects for Items were nonsignificant at either level of Warning. That is, after receiving benign warning, respondents indicated slightly greater intention to complete a similar survey when there were no infrequency items, $F(1,153) = 3.85$, MSE = 1.67, $p = 0.05$; in the absence of warning, however, marginally higher intention emerged when the infrequency items were included, $F(1,153) = 3.05$, MSE = 1.67, $p = 0.08$.

Finally, although the main effect of Items was not significant from MANOVA, results from ANOVA suggested that the presence of infrequency items resulted in better data quality perceptions.

Results from *Study 4* echo the conclusion from *Study 3*: the presence of infrequency items did not result in significantly adverse reactions toward the survey. The only significant influence from the infrequency items was in fact in favor of their implementation, as respondents perceived higher data quality in the Items condition.[4] Moreover, *Study 4* revealed

[4] We conjecture that the infrequency items might have grabbed the attention of some participants. That is, the presence of the infrequency items might have led some participants to find the items funny and

some benefits of providing a benign warning about detecting IER. Specifically, respondents enjoyed filling out the survey more and had better perceptions of the survey design as a result of the warning. Finally, having both infrequency items and warning resulted in the most positive perceptions of the survey design.

Although *Study 4* used a Chinese sample, we suspect that the benefits of coupling the infrequency items with benign warning/explanation are likely to generalize to other cultures because the psychological principles behind the inclusion of warning/explanation were largely derived from research conducted within Western cultures (see Colquitt et al. 2001).

General Discussion

Despite potential concern about the infrequency approach's applicability to capture IER (Huang et al. 2012), Meade and Craig (2012) provided initial evidence that an infrequency scale can detect IER. The current research builds on Meade and Craig's work on the infrequency approach to IER detection by providing a comprehensive examination of validity evidence and assessing respondent reactions. In terms of validity evidence, we demonstrated that IER can be captured with a measurement scale consisting of multiple infrequency items. Specifically, *Studies 1* and *2* showed that items developed on the premise of the infrequency approach yielded a high level of internal-consistency reliability, conformed to a unidimensional factor structure, and was related to other IER indices and to objective criteria of response effort. In terms of respondent reactions, *Studies 3* and *4* indicated that items based on the infrequency approach did not result in significantly more negative reactions.

Research Implications

The design of the current infrequency items as well as the study context provides ample grounds for interpreting our findings vis-à-vis prior research. First, our infrequency items were developed with a low level of social desirability in mind, as socially desirable responding may limit the success of infrequency items in detecting IER. For example, if an item reads "I am knowledgeable about parabanjology," agreement to this item may be driven by self-enhancement (i.e., because people often want to appear knowledgeable; see Paulhus et al. 2003). Interestingly, applying such socially desirable infrequency items in a low-stakes survey context (e.g., anonymous research with some incentive for participation) might end up capturing both self-enhancement and IER simultaneously.

Second, the survey context in which the infrequency approach is utilized should be taken into account. The current surveys were all administered in a low-stakes context, where respondents in general are motivated to respond carefully and yet IER is a valid concern. Our findings of internal consistency and unidimensionality in *Studies 1* and *2* reflected the consistent difference across respondents on the underlying IER response set—specifically, a small proportion of respondents engaged in IER behavior while the remaining respondents did not (Meade and Craig 2012). As a thought experiment, if each respondent had diligently responded to the survey, we would have found poor psychometric properties for the infrequency IER scale due to limited true score heterogeneity (see Furr and Bacharach 2014). We would also have found poor convergent validity evidence among IER indices due to range restriction.

Third, the infrequency approach may not be appropriate or necessary in a high-stakes context (e.g., selection), as respondents will be motivated to respond attentively. Indeed, the infrequency approach has been applied to capture both faking (Pannone 1984) and IER (Green and Stutzman 1986; Green and Veres 1990) on ratings of task statements, with a critical difference on survey contexts. The former was administered to applicants, who were presumed to be motivated to fake, whereas the latter were used in incumbents, whose motivation to respond attentively was in question. Considering a different type of high-stakes survey context, if participants are motivated to present themselves in a negative light (e.g., fake psychological symptoms), failing the infrequency IER items may indicate malingering rather than IER.

Fourth, the validation evidence in *Studies 1* and *2* was based on infrequency items worded uniformly in the same direction, a practice that can potentially lead to false negatives (Meade and Craig 2012). For example, if some people engage in IER by consistently endorsing "disagree," they will not be detected by those items. The high convergent validity evidence between the infrequency items and the other indices in *Study 1* somewhat mitigated this concern. Further mitigating this concern on false negatives due to uniform wording direction, the current eight-item infrequency scale resulted in *higher* scores than a modified scale containing four of the current items and four reversed items when embedded in an otherwise identical survey administered to two groups of student workers (Huang et al. 2014).

The current studies lay the groundwork for future research on IER. First, as an initial investigation of respondent reactions to researchers' effort to detect and deter IER, our studies showed that the use of a benign warning together with the infrequency IER items led to the

Footnote 4 continued
thus had a piqued interest in the questionnaire. This is consistent with presence of infrequency items resulted in a marginal increase in enjoyment. As a result, attentive participants might have become even more attentive after reading the infrequency items, and subsequently tended to think the data quality being higher.

most favorable reactions in general. However, we found an unexpected Items by Warning interaction effect on intention to completing a similar survey. It might be possible that, given both the warning and the infrequency items, the cognitive load for completing the survey was high, reducing respondents' willingness to complete another similar survey. As another possible explanation, the effect could be due to slight differential attrition across conditions: Unlike the other conditions, perhaps respondents in condition (d) were most likely to complete the survey once they started, despite their low intention to respond to surveys in general. Finer-grained understanding of respondents' motivation in the survey process, coupled with better assessment of reactions, can help answer this question.

Second, the current examination of IER as a *response set* gives rise to the study of IER as a *response process*. Current findings of high internal consistency estimates of the infrequency scale reflect consistency on IER behavior. That is, across the infrequency items interspersed throughout the survey, some respondents were consistently inattentive while others responded carefully. Future studies may turn to a process view of IER behavior, where respondents' attention and effort may wax and wane during the response process. For example, a respondent may start a survey being quite attentive and gradually loses motivation over time and starts engaging in IER behavior over time, whereas another respondent may realize he/she has lost focus for a short while (perhaps triggered by an infrequency IER item) and revert back to attentive responding. Similar to other existing IER measures such as psychometric antonym and individual reliability indices, the infrequency IER scale may not be particularly sensitive to such changes. Future studies may include objective measurement, such as fine-grained response time across sections of a survey to detect IER as a process.

Finally, future research should further examine the practicality of using infrequency items in organization-sponsored questionnaires. A potentially fruitful area of research is to examine how infrequency IER items' characteristics influence an expanded set of reaction outcomes, including constructs such as perceived fairness, perceived predictive validity, and attrition. We believe that infrequency items, when properly designed and implemented (e.g., with a benign explanation/warning), can indeed be effective in such contexts without raising respondents' negative reactions. The key would be in writing infrequency items that are inconspicuous when interspersed among the survey's substantive items. For instance, a post-training assessment of self-efficacy may include pieces of knowledge clearly unrelated to the focal training (see Chiaburu et al. 2014), and an employee survey may embed items about familiarity with a fictitious company policy. A similar approach was used by Green and Stutzman (1986) to detect careless responding to job analysis questions.

Practical Implications

We examined the validity and feasibility of the infrequency IER items, offering survey designers a simple yet effective method to detect IER in low-stakes survey contexts. The current findings point to two promising future applications. First, the particular infrequency items from the current paper may be adopted in low-stakes surveys in work and non-work contexts. The nature of these items makes it relatively easy to intersperse them with various substantive items. Second—perhaps more importantly—the validation of the infrequency approach highlights the feasibility that survey administrators can develop their own infrequency IER items that optimally fit their survey contexts and item content.

We discuss several practical considerations when implementing an infrequency IER scale. First, the infrequency approach is flexible in that the survey designer can write items to fit the particular survey context, such as job attitudes, values, job analysis, and so forth. As noted above, these items should be written with clear answers and low levels of social desirability. Second, we recommend the use of multiple infrequency items on a detection scale so that a reliable measure of the IER behavior can be obtained. The tradeoff between brevity and reliability in the design of psychological measures applies here as well: while a short (e.g., 3-item) infrequency scale can reduce participant burden and minimize potential negative reactions, a longer (e.g., 8-item) infrequency scale can more adequately reflect the underlying IER response set. The bottom line is to screen responses with a reliable and valid IER measure, so that misidentification due to transient errors is minimized. Third, the survey administrator may want to vary the infrequency items' wording directions to avoid missing some long-string responses.

Finally, while several IER detection approaches can capture the underlying response set (Huang et al. 2012; Meade and Craig 2012), the selection of a particular IER approach can be guided by the survey context. For instance, the infrequency IER items are particularly useful in shorter surveys (e.g., see *Study 2*), as the more complex approaches such as psychometric antonym and long string index require longer surveys. As another example, if an organization is concerned about workers spending too much time on a survey and screens survey items meticulously, the inclusion of several infrequency IER items may not be supported by the organization due to the lack of face validity. In addition, to the extent that different IER detection approaches emphasize different ways a respondent may engage in IER behavior, the combination of several detection approaches such as an infrequency scale and minimal response time, whenever feasible, can ensure greater coverage in ensuring the quality of survey data.

Selection and Development of Effective Infrequency Items

Existing research suggests that infrequency IER items (see Table 1; also see Meade and Craig 2012) may vary in several important ways, which may have implications on item design and selection. First, some infrequency items include *impossible* content, whereas others include *improbable* content. Teleporting across time and space, for instance, is impossible; not using a computer while filling out a long online survey, however, is improbable but not impossible, given the potential alternative of smartphones. We believe in situations where survey designers have concerns about respondent reactions, they may include some highly improbable items, as the co-occurrence of several improbable events is highly unlikely. However, such practice comes with a drawback: the improbable infrequency items can contain more measurement error than the impossible items, as careful respondents could legitimately endorse an improbable item, hence creating a false appearance of IER.

Second, infrequency items vary in the extent to which they include humorous or ambiguous content, a feature that could influence item endorsement irrespective of one's level of IER. Meade and Craig (2012) expressed concerns that attentive respondents might endorse some infrequency items because they found those items humorous or simply interpreted them figuratively. For example, some attentive respondents might indicate agreement to the statement "All my friends are aliens" because they interpret the word alien in its legal sense. Third, infrequency items may include socially desirable content. As our pilot study suggested, agreeing to the statement "I work fourteen months in a year" could be slightly confounded with impression management. Finally, infrequency items may vary in the degree to which they apply to respondents with different cultural backgrounds and reading ability. For instance, respondents who acquired English as a second language may have trouble realizing that "parabanjology" is not a real major. A survey designer may find it difficult to create items that balance and satisfy each of these four considerations. Thus, we recommend the use of multiple infrequency items so that the occasional error in one item will be compensated for by other items.

Conclusion

The present studies build on past research by documenting comprehensive validity evidence for the infrequency approach to IER detection and they provide the first examination of respondent reactions to the infrequency approach. The validity evidence and the absence of negative reactions indicate that the infrequency approach is a feasible tool for detecting IER in low-stakes surveys.

Acknowledgments We would like to thank Neal Schmitt, Fred Oswald, and Adam Meade for comments on earlier drafts of this paper, and Jessica Keeney and Paul Curran for suggestions during the early stages of this research. We also thank Travis Walker for assisting with data collection.

References

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

Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology: Interdisciplinary and Applied, 123*, 101–103.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43*, 800–813.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.

Chiaburu, D. S., Huang, J. L., Hutchins, H. M., & Gardner, R. G. (2014). Trainees' perceived knowledge gain unrelated to the training domain: The joint action of impression management and motives. *International Journal of Training and Development, 18*, 37–52.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology, 86*, 425–445.

Costa, P. T., Jr, & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The Sage handbook of personality theory and assessment: Personality measurement and testing* (pp. 179–198). London: Sage.

Croteau, A.-M., Dyer, L., & Miguel, M. (2010). Employee reactions to paper and electronic surveys: An experimental comparison. *IEEE Transactions on Professional Communication, 53*, 249–259.

DiLalla, D. L., & Dollinger, S. J. (2006). Cleaning up data and running preliminary analyses. In F. T. L. Leong & J. T. Austin (Eds.), *The psychology research handbook: A guide for graduate students and research assistants* (pp. 241–253). Thousand Oaks, CA: Sage.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*, 213–224.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68*, 532–560.

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

Green, S. B., & Veres, J. G. (1990). Evaluation of an index to detect inaccurate respondents to a task analysis inventory. *Journal of Business and Psychology, 5*, 47–61.

Hackman, J. R., & Oldham, G. R. (1975). Development of the job diagnostic survey. *Journal of Applied Psychology, 60*, 159–170.

Hogan, R., & Hogan, J. (2007). *Hogan Personality Inventory manual* (3rd ed.). Tulsa, OK: Hogan Assessment Systems.

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

Huang, J. L., Bowling, N. A., & Liu, M. (2014). The effects of insufficient effort responding on the convergent and discriminant validity of substantive measures. Unpublished manuscript.

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

Huang, J. L., Liu, M., & Bowling, N. A. (2014, May). Insufficient effort responding: Uncovering an insidious threat to data quality. In J. H. Huang & M. Liu (Co-chairs), Insufficient effort responding to surveys: From impact to solutions. *Symposium to be presented at the Annual Conference of Society for Industrial and Organizational Psychology*, Honolulu, HA.

Jackson, D. N. (1974). *Personality Research Form manual*. Goshen, NY: Research Psychologists Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*, 103–129.

Liu, M., Bowling, N. A., Huang, J. L., & Kent, T. A. (2013). Insufficient effort responding to surveys as a threat to validity: The perceptions and practices of SIOP members. *The Industrial-Organizational Psychologist, 51*(1), 32–38.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.

Muthén, L. K., & Muthén, B. O. (2011). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Pannone, R. D. (1984). Predicting test performance: A content valid approach to screening applicants. *Personnel Psychology, 37*, 507–514.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology, 84*, 890–904.

Scandell, D. J. (2000). Development and initial validation of validity scales for the NEO-Five Factor Inventory. *Personality and Individual Differences, 29*, 1153–1162.

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