Control and Detection of Careless Responding in Questionnaire Surveys

Authors: Zhong Xiaoyu, Li Mingyao, Li Lingyan, Li Lingyan

Date: 2020-10-28T00:00:00+00:00

Abstract

Questionnaire surveys are a widely used data collection method in psychological and educational research, and careless responding by participants may compromise the validity of questionnaire data. A review of existing research reveals that: (a) careless responding can be defined from two perspectives: external response patterns and internal generating causes; (b) common proactive control methods for careless responding primarily include two categories: reducing task difficulty and enhancing participant motivation; (c) post-hoc identification methods mainly consist of three categories: embedding identification scales, response pattern identification, and response time identification. Future research should optimize and develop control methods based on studies of response mechanisms, examine the cross-situational applicability of identification methods and develop novel approaches, and conduct more in-depth investigations into the identification and handling of partial carelessness.

Full Text

Preventing and Detecting Insufficient Effort Responding in Surveys

ZHONG Xiaoyu; LI Mingyao; LI Lingyan
(Collaboration Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing 100875, China)

Abstract: Survey research represents a widely employed data collection method in psychological and educational studies. However, insufficient effort responding (IER) by participants can substantially compromise data quality. A comprehensive review of existing literature reveals three key findings: (a) IER can be conceptualized from two complementary perspectives—external response patterns and internal causal mechanisms; (b) Preemptive control

strategies primarily fall into two categories: reducing task difficulty and enhancing respondent motivation; (c) Post-hoc detection methods encompass three major approaches: embedded validity scales, response pattern analysis, and response time analysis. Future research should focus on optimizing prevention methods through deeper investigation of IER mechanisms, evaluating the cross-situational applicability of detection techniques while developing novel approaches, and conducting more thorough examinations of partial IER identification and treatment.

Keywords: insufficient effort responding (IER); data screening; invalid response; survey and questionnaire design & construction
Classification Code: B841

In scientific research, when investigators operationalize their constructs into a series of interconnected, measurable indicators and compile them into question inventories (Liu & Chen, 1991) designed to assess human behaviors or attitudes, a survey instrument is created (Che, 2001). Survey methodology constitutes a prevalent data collection approach in social sciences, yet data obtained through this method frequently contain substantial measurement error. Consequently, prior to modeling, inference, and decision-making, researchers must screen these data to identify and correct invalid responses (Huang et al., 2012).

Among these errors, insufficient effort responding represents a pervasive yet often overlooked factor due to its challenging nature. Empirical evidence indicates that IER prevalence ranges from 1% (Gough & Bradley, 1996) to 30% (Burns et al., 2014) across most survey contexts. IER contaminates data quality and substantially reduces data authenticity; if left unaddressed, it may obscure meaningful findings and generate spurious results (Curran, 2016; Maniaci & Rogge, 2014). Its detrimental effects primarily include:

First, IER compromises the reliability and validity of measurement instruments (DeSimone et al., 2018; Kam & Meyer, 2015; Zijlstra et al., 2011). For instance, reverse-coded items in unidimensional scales are more likely to emerge as separate factors from positively-worded items (Woods, 2006). Second, IER produces random data or outliers that subsequently bias inference and decision-making (Barge & Gehlbach, 2012; Huang et al., 2015; Zijlstra et al., 2011), such as affecting percentile rank scoring (Zijlstra et al., 2011) and inflating or deflating correlations between variables (Credé, 2010; Holtzman & Donnellan, 2017; Huang et al., 2015; Schneider et al., 2018).

The proliferation of electronic surveys (Evans & Mathur, 2005; Lloyd & Devine, 2010) has exacerbated IER risks due to factors including enhanced submission convenience (Johnson, 2005), response anonymity (Meade & Craig, 2012), uncontrolled testing environments (Barge & Gehlbach, 2012; Carrier et al., 2009; Meade & Craig, 2012), and reduced experimenter-participant interaction (Francavilla et al., 2019; Johnson, 2005; Ward & Meade, 2018; Zhang & Conrad, 2018) (Ward & Pond, 2015). In light of these challenges, this paper systemati-

cally synthesizes relevant research to heighten awareness among researchers and practitioners regarding IER and to provide guidance for selecting appropriate prevention and detection methods. We first clarify the conceptual boundaries of IER by reviewing related terminology in international literature, then summarize prevention and detection techniques, and conclude with future research directions.

Concepts of Insufficient Effort Responding

The concept of "insufficient effort responding" lacks a unified terminology in English-language literature, with subtle variations across studies. These terms primarily emphasize two distinct orientations: external response patterns and internal causal mechanisms.

1.1 External Response Patterns

One conceptualization of IER focuses on observable outcomes, namely response patterns, particularly the distribution of options in Likert-type scales. The widely adopted term "random responding" describes participants who select answers arbitrarily throughout a questionnaire (Beach, 1989; Berry et al., 1992; Marjanovic et al., 2015). However, researchers have noted that IER may also manifest in non-random patterns (Meade & Craig, 2012), such as straight-lining or nondifferentiation (Curran, 2016; Fang et al., 2016; Huang et al., 2012; Meade & Craig, 2012), or selecting answers according to meaningless regularities (Dunn et al., 2018). Furthermore, Grau et al. (2019) found that IER overlaps to some extent with specific response styles. Illustrative examples of various response patterns are presented in Figure 1 [Figure 1: see original paper].

Figure 1. Examples of Various Response Patterns

These studies provide intuitive descriptions of the explicit response patterns associated with IER while acknowledging that such patterns stem from participant inattentiveness or lack of effort. This "lack of effort" precisely distinguishes IER from social desirability responding—although the latter may also exhibit specific response styles (He & Van De Vijver, 2013, 2015a, 2015b, 2016), it does not reduce cognitive load during responding but rather "requires additional cognitive effort" (Grau et al., 2019; Maniaci & Rogge, 2014; McGrath et al., 2010; Meade & Craig, 2012). Nevertheless, because IER patterns are complex and diverse, enumeration is impractical, and conceptualizing IER solely through pattern description risks narrowing its scope.

1.2 Internal Causal Mechanisms

To avoid such conceptual narrowing, some researchers emphasize the underlying causes of IER. Krosnick (1991) proposed that respondent effort exists on a continuum from ideal maximum (optimization) to complete lack of effort, with task difficulty, respondent ability, and respondent motivation jointly determin-

ing one' s position on this continuum. Zhang (2013) further refined this theory by distinguishing three anchor points: ideal maximum (a), attainable maximum (b), and actual value (c). Task difficulty and respondent ability determine the position of the attainable maximum (b), while respondent motivation determines the actual value (c) (see Figure 2 [Figure 2: see original paper]). IER thus represents behaviors where low motivation leads participants to disregard questionnaire instructions, inadequately comprehend item content, or fail to provide accurate responses (Bowling et al., 2016; Huang et al., 2012; Meade & Craig, 2012). Related concepts include insufficient effort responding (Huang et al., 2012), careless responding (Grau et al., 2019; Johnson, 2005; Meade & Craig, 2012), disengaged responding (Soland et al., 2019), shirking behavior (Fang et al., 2016), inattention (Johnson, 2005; Maniaci & Rogge, 2014; Meade & Craig, 2012), and satisficing behaviors (Anduiza & Galais, 2017; Barge & Gehlbach, 2012; Zhang & Conrad, 2018).

Figure 2. Theoretical Framework of Krosnick (1991) and Zhang (2013) (Source: Zhang, 2013)

These two categories of terminology describe and define IER from complementary perspectives that are not mutually exclusive. Together, they enrich the conceptualization of IER. Building on these terms, researchers have proposed that IER can be defined as response patterns exhibited by individuals who, due to insufficient motivation, fail to comply with item requirements or respond without carefully reading item content, with explicit manifestations including random responding, straight-lining, and similar behaviors (Huang et al., 2012).

Preemptive Control of Insufficient Effort Responding

Preemptive control refers to methods implemented during questionnaire design or administration that prevent or reduce IER. These strategies primarily fall into two categories: first, reducing task difficulty to increase the attainable maximum of respondent effort, typically through adjusting questionnaire wording and length; second, enhancing respondent motivation to elevate the actual level of effort, commonly through external incentives, commitment requests, and feedback mechanisms that increase social interaction.

1.3 Reducing Task Difficulty

According to Zhang' s (2013) theoretical framework, task difficulty influences the attainable maximum of respondent effort. In survey contexts, reducing task difficulty manifests in two ways: providing clear, appropriate, and comprehensible instructions and item wording to minimize cognitive processing demands (García, 2011; Rousseau & Ennis, 2013), and shortening questionnaires to reduce respondent fatigue. Overly lengthy surveys may lead to depleted cognitive resources, inability to sustain attention (Wei & Zhang, 2019), or feelings of boredom and tedium that trigger IER (Baer et al., 1997; Berry et al., 1992). Empirical evidence demonstrates that longer questionnaires negatively impact data

quality (Nguyen, 2017). Consequently, researchers recommend using single-item measures for well-defined, unidimensional constructs that are not central to the research, particularly in large-sample or time-constrained studies, to enhance data collection efficiency (Wei & Zhang, 2019).

1.4 Enhancing Respondent Motivation

Preemptive IER control more frequently targets respondent motivation to increase actual effort levels. When participants feel unwilling or unaccountable for their responses, their actual diligence falls far below their maximum potential (Ward & Meade, 2018). Strategies to enhance motivation include:

1) **External incentives and warnings**. Since most surveys are low-stakes or uninteresting to participants (Wei & Zhang, 2019), they cannot inherently maintain high motivation, necessitating external rewards or cautions. Monetary compensation represents a common recruitment incentive, yet excessive emphasis on rewards may encourage perfunctory responding (Barge & Gehlbach, 2012; Maniaci & Rogge, 2014). Therefore, warnings complement rewards by appearing in instructions, informing participants that researchers will statistically evaluate response quality, exclude problematic data, provide feedback on data quality, or even penalize IER (e.g., withholding payment). Research indicates that warnings significantly reduce IER (Huang et al., 2012; Ward & Pond, 2015).

2) **Commitment to respond conscientiously**. Once individuals explicitly commit to an action or position, they tend to behave consistently with that commitment (Cialdini, 2001). However, direct commitments may prove insufficient. Cibelli (2017) experimentally required participants to pledge to "think carefully, recall diligently, and spend adequate time responding," thereby increasing their sense of responsibility. Results showed limited effectiveness, with commitments only increasing effort on difficult items (e.g., open-ended questions). Additionally, participants often ignore instructions, prompting researchers to propose instructional manipulation checks (IMCs) that require correct responses before survey continuation. Oppenheimer et al. (2009) found this approach improved overall response quality.

3) **Feedback and increased social interaction**. These methods primarily target electronic surveys. First, pop-up warnings triggered by excessively fast responding or consecutive identical responses can improve data quality (Cibelli, 2017; Zhang, 2013; Zhang & Conrad, 2018). Second, the absence of social interaction with experimenters in online surveys is believed to hinder sustained motivation and cognitive effort (Fang et al., 2014; Meade & Craig, 2012). Ward and Pond (2015) addressed this by incorporating "virtual humans"in electronic surveys to simulate experimenter-participant interaction, thereby enhancing attention and accountability. Experimental evidence demonstrates that combining warning instructions with su-

pervisory virtual humans significantly reduces IER prevalence. However, Francavilla et al. (2019) found limited effects, with experimental participants showing improvement only on select indicators. Further research examined the role of "social presence" in feedback, substituting human faces for yellow exclamation icons in pop-up messages, yet found no significant differences between methods (Zhang, 2013; Zhang & Conrad, 2018). Moreover, pop-up messages and virtual humans risk distracting participants (Ward & Pond, 2015).

Post-Hoc Detection of Insufficient Effort Responding

While preemptive controls reduce IER incidence, they cannot eliminate it entirely. Therefore, post-hoc identification and removal of remaining IER data after collection is essential. Researchers have developed numerous detection methods.

1.5 Embedded Validity Scales

Embedded validity scales, also termed proactive approaches (Dunn et al., 2018) or direct screening methods (Desimone et al., 2015), operate by inserting detection items within the original questionnaire to measure IER severity. These scales comprise three primary item types: bogus items, instructional items, and self-report items.

1) **Bogus items** present questions with obviously correct answers, such as "I was born on February 30th" (Huang et al., 2012) or "I have traveled around the world 92 times" (Dunn et al., 2018). Although these items use standard Likert 5- or 7-point scales like surrounding items, only "strongly disagree" represents a reasonable response. Multiple selections of alternative options flag insufficient effort.

2) **Instructional items** require participants to follow explicit directions, such as "Please select the second option for this item" (Anduiza & Galais, 2017), "Please skip this item" (Maniaci & Rogge, 2014), or "Please click the small circle at the bottom of the screen" (Oppenheimer et al., 2009). Repeated failure to comply with instructions indicates IER.

3) **Self-report items** directly assess participants' subjective judgments of their own diligence, such as "I did not pay much attention to the actual meaning of these questions" or "I responded very carelessly" (Huang et al., 2012). This straightforward approach flags participants who self-report insufficient effort.

While simple and intuitive, embedded scales face two limitations. First, IER participants may not completely ignore items; if detection items are unrelated to the main content, they require minimal cognitive resources to notice, thereby only minimally detecting IER. Second, excessive embedded items may irritate

conscientious respondents (Costa Jr & McCrae, 2008; Curran, 2016; Meade & Craig, 2012).

1.6 Response Pattern Analysis

Response pattern analysis, also called reactive approaches, examines participants' response patterns after data collection and computes detection indices representing IER severity (Meade & Craig, 2012). The detection logic primarily involves individual consistency analysis and outlier analysis.

1.6.1 Individual Consistency Analysis In Likert scales, IER commonly manifests as random or straight-line responding (Curran, 2016; Maniaci & Rogge, 2014; Meade & Craig, 2012; Revilla & Ochoa, 2015). These indices assume that excessively random or excessively consistent option distributions indicate insufficient effort (Barge & Gehlbach, 2012; Marjanovic et al., 2015). Common indices include the long string index, inter-item standard deviation (ISD), individual reliability, and positive/negative item correlations.

1) **Long string index** measures the longest consecutive run of identical responses, making it highly sensitive to straight-lining (Meade & Craig, 2012). For example, in a 10-item, 4-point scale with response pattern [1,1,1,2,1,2,2,3,4,4], runs of identical responses are [3,1,1,2,1,2], yielding a maximum run of 3 (the long string index) and a mean of 1.67 as an alternative IER indicator. Some researchers also compute long string indices per response option (Costa Jr & McCrae, 2008; Huang et al., 2012), which would be [3,2,1,2] for options 1-4 in this example.

2) **Inter-item standard deviation (ISD)**, also termed intra-individual response variability index (Curran, 2016; Dunn et al., 2018; Marjanovic et al., 2015), is calculated as:

$$ISD_i = \sqrt{\frac{\sum(X_{ig} - \bar{X}_i)^2}{(k-1)}}$$

where ISD_i represents participant i' s ISD, X_{ig} is participant i' s score on item g, $\bar{X}_i$ is participant i' s mean across all items, and k is the total number of items. Excessively random responding produces abnormally large ISD within a single dimension, while excessively consistent responding yields abnormally small ISD across the entire questionnaire (Dunn et al., 2018; Marjanovic et al., 2015). Researchers recommend ISD calculation for questionnaires containing 25-150 items with more than 5 items per dimension (Barge & Gehlbach, 2012; Dunn et al., 2018).

3) **Individual reliability** measures IER under two assumptions: each subscale assesses a single psychological construct, and IER participants respond randomly (Curran, 2016). The most common index is even-odd consistency (Huang et al., 2012; Jackson, 1976, 1977; Johnson, 2005; Meade

& Craig, 2012), computed by dividing the questionnaire into subscales, calculating mean scores for odd and even items within each subscale, correlating these vectors, and applying Spearman-Brown correction. Jackson (1977) suggested that even-odd consistency below 0.30 indicates probable IER. Curran (2016) proposed Resampled Individual Reliability (RIR), which uses repeated resampling and bootstrapping to generate numerous split-half samples for more robust estimation.

4) **Positive/negative item correlations** assess correlations between item pairs with semantically similar or opposite meanings. Item pairs can be constructed either semantically during questionnaire design or psychometrically through data-driven approaches (Curran, 2016). Following Johnson' s (2005) recommendation, pairwise item correlations can be computed from collected data, with correlations above 0.60 forming psychometric positive/negative item pairs. IER severity is reflected in the correlation values of these item pairs.

Despite their intuitive nature, individual consistency indices face limitations. Response consistency is influenced by questionnaire content, length, and format, making cross-questionnaire cutoff values difficult to establish and reducing effectiveness in certain contexts. For instance, the long string index is limited in short questionnaires (Curran, 2016). Moreover, in attitude and adaptability surveys where score distributions are often skewed rather than normal (Mou, 2017; Wang & Zhu, 2009; Yao et al., 2012; Zheng et al., 2018), selecting many "strongly agree" options is normative. Additionally, when reverse-coded items are present, score-sensitive indices like individual reliability and ISD require cautious application (Curran, 2016).

1.6.2 Outlier Analysis Outlier analysis operates on the assumption that "most participants in any given sample are thinking carefully and responding conscientiously" (Curran, 2016). Therefore, substantial deviation from group response patterns suggests IER. Common indices include Mahalanobis distance, individual respondent' s goodness-of-fit score (RGF), person-total correlation, and person-fit statistics such as Guttman error count, U3 index, lZ index, and autoencoder algorithms from neural networks.

1) **Mahalanobis distance** (Mahalanobis, 1936) is a widely used multivariate outlier detection index available in most statistical software. Defined as:

$$MD_i = \sqrt{(x_i - \mu)^T S^{-1} (x_i - \mu)}$$

where $x_i = (x_{i1}, ..., x_{ik})^T$ represents participant i' s scores on k dimensions, $\mu = (\mu_1, ..., \mu_k)^T$ is the mean vector, and S is the covariance matrix. Meade and Craig (2012) demonstrated through simulation that Mahalanobis distance is a powerful IER detection indicator. Velleman and Welsch (1981) suggested using

leverage values $h_{ii} = \frac{MD^2}{2k}$ as outlier criteria, where k is the number of variables and n is the sample size.

2) **Individual respondent' s goodness-of-fit score (RGF)** (Kountur, 2016) is calculated as:

$$RGF = \sum \frac{(X_g - \bar{X}_g)^2}{\bar{X}_g}$$

where RGF represents response conscientiousness, X_g is the participant' s score on item g, and $\bar{X}_g$ is the mean score across all participants on item g. Larger RGF values indicate greater deviation from the overall response pattern.

3) **Person-total correlation** (Curran, 2016) computes the correlation ρ_{XM} between a participant' s response pattern X and the aggregate pattern M of all other participants, where $M = E(X)$. Low person-total correlation suggests substantial divergence from the overall pattern, potentially indicating IER.

4) **Person-fit statistics**, widely used in achievement testing to identify aberrant individuals, compare observed score distributions with ideal distributions (Meijer & Sijtsma, 2001). This logic has been adapted for IER detection in surveys, requiring the assumption that most participants respond conscientiously to construct ideal distributions from group data (Meijer & Sijtsma, 2001; Wang & Xu, 2015). Common person-fit indices for IER detection include Guttman error count (G_P) and its standardized form (G_{NP}) for polytomous scoring (Emons, 2008; Guttman, 1944, 1950), the polytomous version of U3 index ($U3_P$) (Emons, 2008; Van der Flier, 1980), and the polytomous version of lZ index (lZ_P) (Melipillán, 2019).

   - **Guttman error count**: The Guttman model assumes that participants should score higher on easier items. Originally developed for dichotomously-scored achievement tests, items are ordered by decreasing difficulty π_g. If a participant misses an easier item while correctly answering a more difficult one, this violates the Guttman model and constitutes a Guttman error. More errors indicate greater aberrance. The Guttman model extends to polytomous scoring for Likert-type questionnaires (G_P) (Emons, 2008; Niessen et al., 2016). Based on dominance model theory, higher trait levels increase the probability of selecting higher response options. Option endorsement probabilities π_g can be calculated analogously to item correct rates, with Emons (2008) proposing a standardized version G_{NP} for cross-context comparisons.

   - **U3 index**: This nonparametric person-fit index demonstrates good statistical power (Karabatsos, 2003). Originating from achievement testing, the general expression for nonparametric person-fit indices is:

$$G_i = \sum_{g=k-r+1}^{k} w_g$$

where g is the item index, k is the total number of items, X_g is the participant' s score on item g, i indexes participants, n is the sample size, and r is the number of correctly answered items (Meijer & Sijtsma, 2001). The adaptive function w_g varies across indices, with $w_g = \ln(\cdot)$ for U3. Smaller absolute G_i values indicate less aberrance, with $G_i = 0$ representing perfect fit to the Guttman model. Like Guttman errors, substituting option endorsement probabilities π_g for correct rates enables U3 application to polytomous scales (Emons, 2008).

- **lZ index**: Levine and Rubin (1979) proposed the log-likelihood fit statistic, the most widely studied person-fit index. The l statistic, a parametric index, represents the discrepancy between observed score patterns and the ideal pattern predicted by IRT models, with lZ being its standardized form (Drasgow et al., 1985):

$$l = \sum \{X_g \ln P_g(\theta) + (1 - X_g) \ln[1 - P_g(\theta)]\}$$

For dichotomous scoring (e.g., achievement tests), $P_g(\theta)$ represents the probability of a correct response for ability θ; for polytomous scoring, it is denoted $P_{xg}(\theta)$, representing the probability of endorsing option x_g on item g (Melipillán, 2019). Smaller l and lZ values indicate greater aberrance.

5) **Autoencoder**: This unsupervised neural network method for identifying high-dimensional outliers is widely used in engineering and was applied to IER detection by Melipillán (2019). Autoencoders compress data through dimensionality reduction encoding then reconstruct it through decoding, comparing generated and original data. Outliers typically show larger reconstruction errors. With appropriate threshold settings, aberrant cases can be flagged. Melipillán's research demonstrated that autoencoder-based identification through four iterations outperformed lZ index detection.

However, all outlier indices heavily depend on sample characteristics, as they only indicate deviation from the group without revealing its cause, making their application to IER detection questionable. First, in low-stakes surveys where IER prevalence may be substantial (unlike achievement testing contexts with few aberrant responses), outlier-flagged participants may actually be conscientious responders rather than the numerous IER cases. Second, individual variation across items is normal, and using such differences to judge IER may exclude legitimately extreme but conscientious respondents. Third, other factors like faking can also cause aberrant data, so flagged cases may not result from insufficient effort. Additionally, these indices have specific limitations: Mahalanobis

distance requires multivariate normality rarely satisfied in survey data (Niessen et al., 2016); person-fit indices based on dominance model assumptions may not align with attitude survey cognitive processes; and neural network algorithms produce difficult-to-interpret results with questionable cross-situational stability.

1.7 Response Time Analysis

Response time analysis operates on the principle that responses provided too rapidly for basic item comprehension cannot represent genuine attitudes (Huang et al., 2012). Four methods exist for establishing response time thresholds: empirical specification, visual inspection of distributions, integration with other data quality indicators, and experimental pretesting.

Empirically specified thresholds include absolute and relative standards. The most widely used absolute standard is Huang et al.'s (2012) "informed guessing" criterion of 2 seconds per item (Curran, 2016; Soland et al., 2019). Relative standards have also been proposed, with Höhne and Schlosser (2018) summarizing five such criteria (see Table 1).

Table 1. Outlier Response Time Thresholds (Höhne & Schlosser, 2018)

Study	Lower Threshold	Upper Threshold
Mayerl (2013)	Mean-(2*SD)	Mean+(2*SD)
Schnell (1994)	Q.50-(1.5*IQR)	Q.50+(1.5*IQR)
Hoaglin et al. (2000)	Q.50-(1.5*(Q.50-Q.25))	Q.50+(1.5*(Q.75-Q.50))
Hoaglin et al. (2000)	Q.50-(3*(Q.50-Q.25))	Q.50+(3*(Q.75-Q.50))
Lenzner et al. (2010)	-	-

A second method involves visually inspecting response time distributions. Assuming conscientious responding requires at least 5 seconds to read, comprehend, and answer an item, normal response times should exceed 5 seconds, while IER participants may respond faster. This produces a bimodal distribution (see Figure 3 [Figure 3: see original paper]), with an initial "spike" of rapid IER responses followed by the distribution of normal responding (Wise, 2017; Wise & Demars, 2006; Wise & Kong, 2005).

Figure 3. Theoretical Response Time Distributions for Rapid-Guessing (IER) and Normal Responding

A third method links response times with other detection indices (e.g., long string index) to establish or validate thresholds. Soland et al. (2019) applied this strategy to OECD school testing data by dividing mean response times into intervals based on empirical criteria and computing multiple indicators (long string index, reverse-item correlations, second eigenvalue magnitude from EFA, correlations between self-efficacy and achievement scores) within each interval. Results showed poor indicator performance when mean response times fell below 2 seconds per item.

A fourth method involves experimental pretesting. Huang et al. (2012) first used laboratory instructions to create conscientious and IER groups, collecting data including response times. They then fixed specificity at 95% and 99% to derive thresholds and corresponding sensitivities for each indicator, subsequently applying these experimentally-derived thresholds to survey screening.

Response time analysis offers advantages as it is unaffected by response patterns and can be evaluated at the item level. Numerous studies have identified response time as an effective IER indicator (Huang et al., 2012; Wise & Kong, 2005). However, limitations exist. First, response time data are only available for electronic surveys. Second, like other indices, its ability to distinguish conscientious from IER participants depends on the degree of distribution overlap; identification effectiveness decreases when IER distributions do not substantially deviate from normal distributions (Curran, 2016). This is particularly problematic in surveys where even careful reading and contemplation require minimal time, potentially causing numerous "false positives."Some researchers argue that the bimodal distribution theory from cognitive testing does not generalize well to surveys (Soland et al., 2019). Third, increased response time does not necessarily indicate higher data quality (Yan & Tourangeau, 2008). Meade and Craig (2012) suggest a non-linear relationship between response time and data quality: extremely fast responding indicates IER, but extremely slow responding beyond a certain threshold may also suggest insufficient effort, as excessive times may reflect participants chatting, watching television, or listening to music during online surveys (Barge & Gehlbach, 2012; Börger, 2016).

Discussion and Future Directions

Insufficient effort responding represents a common source of noise in survey data. This paper has systematically reviewed IER concepts and summarized prevention and detection methods. Below, we discuss unresolved issues requiring further exploration.

1.8 Optimizing and Developing Prevention Methods Based on IER Mechanisms

Existing research demonstrates that adjusting questionnaire wording or length, providing rewards and warnings, using pop-up reminders, virtual humans, commitment requests, and instructional manipulation checks can reduce IER to

varying degrees. However, these methods may also produce side effects or even counterproductive outcomes—for instance, external incentives may increase respondent carelessness, pop-up reminders may distract participants, and virtual humans may disrupt the responding experience.

To avoid or mitigate these adverse effects and develop more effective prevention methods, researchers must address why these methods work. Future studies could employ techniques such as eye-tracking and EEG to monitor and explore the questionnaire responding process in detail, enriching theoretical understanding of IER mechanisms and influencing factors. These theories could then explain why side effects occur, providing a foundation for method optimization and development.

Additionally, future research should systematically evaluate existing methods by analyzing their specific effects. Current studies often assess method effectiveness by comparing IER detection indices between experimental and control groups, yet many methods only affect certain detection indices. Future experimental designs should test and compare the actual effectiveness of various methods, explaining which types of IER each method reduces based on IER mechanisms, thereby providing guidance for researchers and practitioners.

1.9 Examining Cross-Situational Applicability of Detection Indices and Developing Novel Approaches

Most detection indices were developed for personality inventories or cognitive tests, which feature numerous items and normally distributed scores, making these indices more applicable in such contexts. For example, longer questionnaires provide more items for calculating even-odd consistency and positive/negative item correlations, yielding more stable coefficients, while Mahalanobis distance and lZ indices are more effective with normal distributions.

However, attitude and behavior surveys—also common in social sciences—may not share these characteristics, reducing index effectiveness. In many attitude surveys, conscientious participants tend to select 4 or 5 on 5-point Likert scales, producing negatively skewed distributions, while some IER participants may select 5 for all items. In such cases, reduced within-individual variability diminishes the effectiveness of individual consistency indices, and small deviations from normal participants reduce outlier analysis effectiveness.

Future research must therefore focus on cross-situational applicability. For attitude and behavior surveys, researchers should: (a) combine multiple indices to address single-index limitations, while clearly understanding which IER patterns each index detects to selectively apply appropriate combinations; (b) develop new indices to address existing limitations, particularly exploring person-fit indices and machine learning applications that offer more precise outlier identification than traditional methods like person-total correlation.

Furthermore, current research predominantly uses simulation studies to evaluate

detection indices, yet simulated data parameters may not reflect attitude and behavior surveys. Future studies should utilize real data from these survey types to enhance ecological validity and generalizability.

1.10 Identification and Treatment of Partial IER

Although research often dichotomizes participants as "conscientious" or "insufficient effort," real responding scenarios include participants who are only careless on some items. For example, in lengthy questionnaires, participants may exhibit IER in middle or later sections due to fatigue or waning interest (Baer et al., 1997; Berry et al., 1992; Meade & Craig, 2012). With partial IER, embedded scale error counts, individual consistency indices, and outlier analysis values fall between fully conscientious and fully insufficient effort responders, with similarity to conscientious responding depending on the proportion of partial IER. Existing indices may fail to detect such cases.

Currently, only Dunn et al. (2018) have suggested flexibly selecting continuous item blocks to compute ISD for detecting IER within those sections. For instance, in long questionnaires, later items could be selected to identify fatigue-induced partial IER. However, participants may exhibit IER at any point, particularly in electronic surveys with uncontrolled environments where external distractions can occur anytime. Developing more flexible methods to identify specific IER sections represents a promising research direction.

Moreover, treatment of identified partial IER requires further investigation. Deleting all data from such participants wastes valid responses, while only removing IER sections risks non-random missing data. Even if non-random missingness can be ruled out, IER data are not missing but rather inaccurate data that still reflect some participant tendencies, warranting further exploration of whether and how to apply imputation methods.

Participants do not always think carefully and provide reliable answers. Researchers and practitioners must neither be blindly optimistic nor selectively ignore this phenomenon. Instead, they should implement effective measures to prevent IER during data collection and employ technical methods during data cleaning to identify and remove such noise, ensuring data are as authentic and accurate as possible for reliable subsequent analyses.

车文博. (2001). 心理咨询大百科全书. 杭州: 浙江科学技术出版社. 刘蔚华, 陈远. (1991). 方法大辞典. 济南: 山东人民出版社.

牟智佳. (2017). MOOCs 学习参与度影响因素的结构关系与效应研究——自我决定理论的视角. 电化教育研究, 38(10), 37-43.

王俪嘉, 朱德全. (2009). 中小学教师对待公开课态度的调查研究. 上海教育科研, (08), 28-31.

卫旭华, 张亮花. (2019). 单题项测量: 质疑、回应及建议. 心理科学进展, 27(07), 1194-1204.

姚成, 龚毅, 濮光宁, 葛文龙. (2012). 学生评教异常数据的筛选与处理. 牡丹江师范学院学报 (自然科学版)(03), 7-8.

郑云翔, 杨浩, 冯诗晓. (2018). 高校教师信息化教学适应性绩效评价研究. 中国电化教育 (02), 21-28.

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, 29(3), 497-519.

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment*, 68(1), 139-151.

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education*, 53(2), 182-200.

Beach, D. A. (1989). Identifying the random responder. *The Journal of Psychology*, 123(1), 101-103.

Berry, D. T., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, 4(3), 340.

Börger, T. (2016). Are fast responses more random? Testing the effect of response time on scale in an online choice experiment. *Environmental and Resource Economics*, 65(2), 389-413.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, 111(2), 218.

Burns, G. N., Christiansen, N. D., Morris, M. B., Periard, D. A., & Coaster, J. A. (2014). Effects of applicant personality on resume evaluations. *Journal of Business and Psychology*, 29(4), 573-591.

Carrier, L. M., Cheever, N. A., Rosen, L. D., Benitez, S., & Chang, J. (2009). Multitasking across generations: Multitasking choices and difficulty ratings in three generations of Americans. *Computers in Human Behavior*, 25(2), 483-489.

Cialdini, R. B. (2001). Harnessing the science of persuasion. *Harvard Business Review*, 79(9), 72-81.

Cibelli, K. L. (2017). The effects of respondent commitment and feedback on response quality in online surveys. (Unpublished doctoral dissertation), University of Michigan, Ann Arbor.

Costa Jr, P. T., & 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 Publications Ltd.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, 70(4), 596–612.

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

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology*, 67(2), 309–338.

Desimone, J. A., Harms, P. D., & Desimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, 36(2), 171–181.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67–86.

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, 33(1), 105–121.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, 32(3), 224–247.

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research*, 15(2), 195–219.

Fang, J., Prybutok, V., & Wen, C. (2016). Shirking behavior and socially desirable responding in online surveys: A cross-cultural study comparing Chinese and American samples. *Computers in Human Behavior*, 54, 310–317.

Fang, J., Wen, C., & Prybutok, V. (2014). An assessment of equivalence between paper and social media surveys: The role of social desirability and satisficing. *Computers in Human Behavior*, 30, 335–343.

Francavilla, N. M., Meade, A. W., & Young, A. L. (2019). Social interaction and internet-based surveys: Examining the effects of virtual and in-person proctors on careless response. *Applied Psychology*, 68(2), 223–249.

García, A. A. (2011). Cognitive interviews to test and refine questionnaires. *Public Health Nursing*, 28(5), 444–450.

Gough, H. G., & Bradley, P. (1996). *The California Psychological Inventory™ manual: Third edition.* Palo Alto, CA: Consulting Psychologists Press.

Grau, I., Ebbeler, C., & Banse, R. (2019). Cultural differences in careless responding. *Journal of Cross-Cultural Psychology*, 50(3), 336–357.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9(2), 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60-90). Princeton, NJ: Princeton University Press.

He, J., & Van De Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences*, 55(7), 794–800.

He, J., & Van De Vijver, F. J. R. (2015a). Effects of a general response style on cross-cultural comparisons: Evidence from the teaching and learning international survey. *Public Opinion Quarterly*, 79(S1), 267–290.

He, J., & Van De Vijver, F. J. R. (2015b). Self-presentation styles in self-reports: Linking the general factors of response styles, personality traits, and values in a longitudinal study. *Personality and Individual Differences*, 81, 129–134.

He, J., & Van de Vijver, F. J. R. (2016). Response styles in factual items: Personal, contextual and cultural correlates. *International Journal of Psychology*, 51(6), 445–452.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (2000). *Understanding robust and exploratory data analysis*. New York, NY: John Wiley.

Höhne, J. K., & Schlosser, S. (2018). Investigating the adequacy of response time outlier definitions in computer-based web surveys using paradata SurveyFocus. *Social Science Computer Review*, 36(3), 369–378.

Holtzman, N. S., & Donnellan, M. B. (2017). A simulator of the degree to which random responding leads to biases in the correlations between two individual differences. *Personality and Individual Differences*, 114, 187–192.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, 100(3), 828–845.

Jackson, D. N. (1976). The appraisal of personal reliability. Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.

Jackson, D. N. (1977). *Jackson vocational interest survey: Manual*. Port Huron, MI: Research Psychologists Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39(1), 103–129.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality. *Organizational Research Methods*, 18(3), 512–541.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4), 277–298.

Kountur, R. (2016). Detecting careless responses to self-reported questionnaires. *Eurasian Journal of Educational Research*, (64), 307–318.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213–236.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, 24(7), 1003–1020.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269–290.

Lloyd, K., & Devine, P. (2010). Using the internet to give children a voice: An online survey of 10- and 11-year-old children in Northern Ireland. *Field Methods*, 22(3), 270–289.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2, 49–55.

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

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences*, 84, 79–83.

Mayerl, J. (2013). Response latency measurement in surveys: Detecting strong attitudes and response effects. *Survey Methods: Insights from the Field*. Retrieved from https://surveyinsights.org/?p=1063

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, 136(3), 450–470.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107–135.

Melipillán, E. R. (2019). Careless survey respondents: Approaches to identify and reduce their negative impact on survey estimates. (Unpublished doctoral dissertation), University of Michigan, Ann Arbor.

Nguyen, H. L. T. (2017). Tired of survey fatigue? Insufficient effort responding due to survey fatigue. (Unpublished master' s thesis), Middle Tennessee State University, Murfreesboro.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, 63, 1-11.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867-872.

Revilla, M., & Ochoa, C. (2015). What are the links in a web survey among response time, quality, and auto-evaluation of the efforts done? *Social Science Computer Review*, 33(1), 97-114.

Rousseau, B., & Ennis, J. M. (2013). Importance of correct instructions in the tetrad test. *Journal of Sensory Studies*, 28(4), 264-269.

Schneider, S., May, M., & Stone, A. A. (2018). Careless responding in internet-based quality of life assessments. *Quality of Life Research*, 27(4), 1077-1088.

Schnell, R. (1994). *Graphisch gestützte datenanalyse* [Graphically supported data analysis]. München, Germany: Oldenbourg.

Soland, J., Wise, S. L., & Gao, L. (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education*, 32(2), 151-165.

Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse, Netherlands: Swets & Zeitlinger.

Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, 35(4), 234-242.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 68(3), 456-477.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, 67(2), 231-263.

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

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, 36(4), 52-61.

Wise, S. L., & Demars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 43(1), 19–38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2), 163–183.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28(3), 186–191.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, 22(1), 51–68.

Zhang, C. (2013). Satisficing in web surveys: Implications for data quality and strategies for reduction. (Unpublished doctoral dissertation). University of Michigan, Ann Arbor.

Zhang, C., & Conrad, F. G. (2018). Intervening to reduce satisficing behaviors in web surveys. *Social Science Computer Review*, 36(1), 57–81.

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*, 36(2), 186–212.

Note: Figure translations are in progress. See original paper for figures.