Careless Survey Respondents: Approaches to Identify and Reduce their Negative Impacts on Survey Estimates

by

Edmundo Roberto Melipillán

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in the University of Michigan
2019

Doctoral Committee:

Senior Research Scientist Steven Heeringa, Co-Chair
Professor Emeritus James M. Lepkowski, Co-Chair
Assistant Research Scientist Zeina Mneimneh
Research Assistant Professor Ting Yan

Edmundo Roberto Melipillán

robmeli@umich.edu

ORCID iD: 0000-0003-3360-433X

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my dissertation co-chair, Steve Heeringa, who was a constant source of encouragement, support and advice over the past years. I am deeply indebted to him for his guidance throughout my dissertation. Special thanks must go to Jim Lepkowski, co-chair of my dissertation, for his advice and expertise throughout my PhD. I am also very grateful to my committee members, Ting Yan, and Zeina Mneimneh for their insightful and constructive comments during the various stages of my dissertation. Many thanks to the other faculty and staff at MPSM for their help. Further I like to thank Mengyao Hu for her comments and suggestions.

Finally, my special thanks go to my family. To my wife Valeria, and my daughter and sons Martina, Clemente, and Luciano for their love, patience, and continuous support throughout my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Multi-item response scales are widely used in surveys to assess a variety of constructs including respondents' attitudes, behavior, and personality. Multi-item scales often appear in grid question formats with the same response options for a set of survey question items. In these types of questions, survey satisficing is likely to occur, where respondents might skim instructions, respond in a haphazard fashion, or rush through questionnaires. Those with these behaviors are often referred to as satisficers or careless respondents (CR).

Despite that previous literature has extensively discussed ways to identify satisficing behaviors in these type of scales (e.g., the detection of response order effects, straightlining and speeding, and the use of trap questions), there are two methods overlooked in survey literature. One method is the person-fit-statistics which identify the inconsistency of responses by comparing the expected responses to the actual reported responses. One of the most popular person-fit statistics is the standardized log-likelihood person-fit statistic, also known as the standardized log-likelihood l_z^p, which has been proven to be a useful tool in multi-item scales with a large number of items. Another is the autoencoder method, which was initially developed and used in engineering to identify anomalies or outlier cases.

This dissertation intends to fill three important gaps in the existing literature related to CR identification and reduction of their negative effects. Specifically, this dissertation examines i) the use of standardized standardized log-likelihood l_z^p and the autoencoder in identifying

careless respondents (CR) in multi-item scales; ii) the use of multiple imputation to deal with data of the identified CR; and iii) evaluate the use of standardized log-likelihood l_z^p and the autoencoder as an alternative to trap questions and explore how to best deal with trapped respondents.

The first study compares the performances of standardized log-likelihood l_z^p and the autoencoder in identifying CR in a multi-item scale with a small number of items. This research is based on a full factorial simulation study experimental design focusing on two types of CR behaviors – random response and non-differentiation of question item directions. Results indicate that the autoencoder with two iterations has increased sensitivity and acceptable false positive rates, identifying more CR (higher sensitivity) in all conditions, compared to the standardized log-likelihood l_z^p.

The second study compares three approaches in treating data from CR, including using the full sample or "complete data analysis" approach, excluding all CR data, and deleting and imputing CR data. Results of this chapter suggest that the autoencoder identification with imputation of CR data outperforms the standardized log-likelihood l_z^p identification and imputation.

The third study examines whether the standardized log-likelihood l_z^p and the autoencoder can be used as an alternative to the trap question and what is the optimal approach to deal with CR data. Data from a non-probability web survey suggest that the autoencoder provides equivalent results to the use of trap questions. In addition, it is possible to remove only a subset of trapped respondent data in analysis by using the autoencoder to identify the most-careless subset of trapped respondents.

CHAPTER I

Overview

Survey researchers have long faced a critical issue that not all respondents are as diligent as they would like them to be. Survey cognitive theory suggests that respondents in general go through four cognitive processing steps in answering survey questions (Tourangeau, Rips, & Rasinski, 2009): *comprehending* the question, *retrieving* relevant information from memory, *integrating* information to arrive at a judgment, and *formulating* and *editing* a response. Those who perform all four steps diligently for all survey questions are referred to as survey optimizers. However, not all respondents optimize in taking surveys. To reduce cognitive burden, some respondents skim instructions, respond in a haphazard fashion, or rush through questions. Respondents with these behaviors are known as survey satisficers (Krosnick, 1991) or careless respondents (CR).

One type of question for which quality is known to suffer from satisficing behaviors, are multi-item response scales, often appearing in grid question formats where the same response categories are used for a set of question items. Multi-item scales are often used in surveys to asses latent constructs through manifest variables to measure respondent attitudes, behavior, health (e.g., mental and physical health scales), wellbeing, and personality. Satisficing behavior in multi-items scales has been extensively discussed in the survey literature. Specifically, satisficing behavior in these scales have been assessed through a variety of quality indicators, including response order effects in which the order of the response categories affects

respondents' answers (Yan & Keusch, 2015); response styles such as respondents favoring extreme categories, regardless of question (Baumgartner & Steenkamp, 2001); straightlining where respondents choose the same response option without distinguishing the question items (Zhang & Conrad, 2014); and speeding.

In web surveys, a method which has achieved increasing popularity to identify CR is the use of trap questions, also known as instructional manipulation checks (Hauser & Schwarz, 2015; Oppenheimer et al., 2009). These questions often have "a lure question with lure responses" (Liu & Wronski, 2018) and an instruction asking respondents to ignore the lure question and provide a specific response following the instruction. Those respondents who fail trap questions are believed to be satisficing, not paying attention to survey instructions.

Despite the large survey literature examining satisficing behavior in grid questions, there are two methods that have largely been overlooked in the survey literature. Person-fit-statistics, have been extensively discussed in the psychometric literature to identify inconsistent responses by comparing expected responses based on a psychometric model to reported responses (van der Flier, 1982). One of the most popular person-fit statistics is called the standardized log-likelihood l_z^p. It has been proven to be a useful tool in multi-item scales with a large number of question items to identify patterns of inconsistent responses (Conijn, Franz, Emons, de Beurs, & Carlier, 2019). Another overlooked method is the autoencoder, initially developed and used in engineering to identify anomalies or outlier cases (Chen, Sathe, Aggarwal, & Turaga, 2017).

This dissertation is the first to examine both of these methods in surveys and apply them in the identification of satisficing behavior. This research expands the existing literature on survey satisficing in multi-item scale grid questions in three directions: 1) the identification of CR in multi-item scale questions using the aforementioned methods (the l_z^p and the autoencoder);

2) exploration of the approaches to deal with identified CR in survey data; and 3) the use of the l_z^p and the autoencoder as alternatives to trap questions in web surveys. The dissertation thus has three research objectives:

1. Compare the properties of the standardized log-likelihood l_z^p and the autoencoder in identifying CR in a multi-item scale with a small number of items.
2. Examine the use of deletion and multiple imputation as a means to deal with CR data in comparison to excluding or keeping all CR data.
3. Evaluate the use of the standardized log-likelihood l_z^p and autoencoder as alternatives to trap questions and explore how to best deal with trapped respondents.

Each research objective corresponds to one of the three substantive chapters in this dissertation. Chapter 2 of the dissertation is based on a simulation study, which aims to evaluate the two overlooked methods – the standardized log-likelihood l_z^p and the autoencoder in identifying CR in multi-item scales with a small number of items. Two types of careless responding behaviors are examined in Chapter 2: random responses, where respondents provide random responses, and non-differentiation of item directions in scales with both positively and negatively worded items. Specifically, Chapter 2 examines whether the autoencoder can outperform the standardized log-likelihood l_z^p in detecting these two types of CR behaviors. The comparisons of the two methods are conducted using a full factorial experimental design for six factors. Three factors are related to the characteristics of the scale: the number of question items in the multi-item scale (6 vs. 12), the number of response categories (4 vs. 7), and the quality of the scale (items with medium factor loadings vs. low factor loadings). Three other factors are associated with the characteristics of CR: proportion of CR in the simulated dataset (10%, 20%, and 30%), proportion of the items for which CR employed satisficing behavior (half vs. all

items), and types of careless response (*random* and *non-differentiation of item direction changes* for mixed items with both positive and negative wordings). The experiment has 96 experimental conditions. For each condition, a Confirmatory Factor Analysis (CFA) was used to generate the initial data (i.e., without CR). One thousand datasets were generated for each experimental condition. A subsample of10%, 20%, and 30% of each sample was randomly selected and the question items were replaced with random response patterns (or reversed coding of the responses in mixed scales with both positive and negative item wordings). Both the standardized log-likelihood l_z^p and the autoencoder were applied to each dataset and the sensitivity and false positive rates to detect CR were compared.

Chapter 3 answers the natural follow-up question to identifying CR: what should one do with their responses? One approach is to exclude all their data from analysis (e.g., listwise deletion). This is standard practice, especially when CR are identified using trap questions. This approach reduces sample size, reducing statistical power. A second approach is "complete data analysis", analyze all data available including data for CR. This approach can only be used when researchers can validate that including CR data will have no effect on results. This validation is, however, rarely done in practice. Thus, researchers using complete data analysis would risk biasing findings. Chapter 3 introduces a deletion and multiple imputation method for CR data. The analysis is based on a simulation study, similar to that of Chapter 2. The imputation uses sequential regression predicted mean match. Chapter 3 focuses solely on random response behaviors.

Chapter 4 examines satisficing behavior in web surveys where a trap question is used to identify CR. There is no consensus on how to best use data from trapped respondents. Previous literature expressed concerns about the trap question method. For example, trap questions may

introduce respondent confusion, change respondent behavior, decrease the rapport between the respondents and the survey researchers, and reduce respondent motivation in survey participation. Chapter 4 uses the standardized log-likelihood l_z^p and the autoencoder to identify CR in a web survey multi item scale with a trap question. The CR data are then deleted and multiply imputed as a remedy for careless response. The chapter addresses two research questions: 1) whether the standardized log-likelihood l_z^p and the autoencoder can be used as alternatives to the trap question method; and 2) how to best deal with data from trapped respondents. To answer the first research question, model results based on the standardized log-likelihood l_z^p and the autoencoder are compared with those of the trap question method. For the second question, different approaches are applied to identify careless response, comparing results of imputing only the identified subset with the deleted and multiply imputed method applied to all trapped respondents' data. This chapter sheds light on how to better deal with trapped respondent data.

References

Baumgartner, H., & Steenkamp, J.-B.E.M. (2001). Response styles in marketing research: A Cross-national investigation. *Journal of Marketing Research*, *38*, 143-156.

Chen, J., Sathe, S., Aggarwal, C., & Turaga, D. (2017). Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining* (pp. 90–98).

Conijn, J.M., Franz, G., Emons, W.H.M., de Beurs, E., & Carlier, I.V.E. (2019). The assessment and impact of careless responding in routine outcome monitoring within mental health care. *Multivariate Behavioral Research, 54*, 593-611.

Hauser, D.J., & Schwarz, N. (2015). It's a Trap! Instructional manipulation checks prompt systematic thinking on "Tricky" tasks. *SAGE Open*, *5*(2).

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

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

Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The Psychology of Survey Response*. New York, NY: Cambridge University Press.

Van Der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*, 267-298.

Yan, T., & Keusch, F. (2015). The effects of the direction of rating scales on survey responses in a telephone survey. *Public Opinion Quarterly*, *79*, 145-165.

Zhang, C., & Conrad, F.G. (2014). Speeding in web surveys : The tendency to answer very fast and its association with straightlining. *Survey Research Methods, 8*, 127-135.

CHAPTER II

Identifying Careless Survey Respondents: Comparing the Standardized Log-Likelihood l_z^p and the Autoencoder

2.1 Introduction

Survey researchers have to deal with respondents who are not as diligent as they would like them to be. Respondents might skim instructions, respond in a haphazard fashion, or rush through questions, a set of behaviors known as survey satisficing (Krosnick, 1991). In such cases, the quality of the survey data can be affected negatively. Unfortunately, such respondents have traditionally been challenging to identify (Oppenheimer, Meyvis, & Davidenko, 2009).

One area where survey satisficing is especially likely to occur is multi-item response scales, widely used to assess constructs such as respondent attitudes, behavior, health (e.g., mental and physical health scales), social wellbeing, and personality. Multi-item scales often appear in grid question formats with the same response options for a set of survey question items. Previous literature has extensively discussed ways to identify satisficing behavior in these types of scales, including the detection of response order effects, response styles, straightlining, and speeding (Schonlau & Toepoel, 2015; Zhang & Conrad, 2013).

A method widely discussed in the psychometrics literature to identify *aberrant responses* is person-fit indices. Person-fit methods identify inconsistent responses by comparing the responses expected from a model to the actual reported responses (Van Der Flier, 1982). The person-fit index is applied here by using one of the most popular person-fit statistics, the

standardized log-likelihood l_z^p method, to identify satisficing behavior responses to multi-item scales.

Person-fit statistics, especially the l_z^p method, have been proven to be a useful tool to identify satisficing behavior in multi-item scales with a large number of question items. However, this method has several potential disadvantages. It does not work for multi-dimensional scales or scales with a small number of items.

In this chapter, another method to identify satisficing behavior in surveys is introduced to address this issue. The *autoencoder neural network*, is a newly developed unsupervised neural network method initially used to identify anomalies or outlier cases in engineering problems. This chapter uses a simulation study to compare the performance of the standardized log-likelihood l_z^p and the autoencoder in identifying satisficing behavior in multi-item scales with a small number of items.

2.1.1 Survey Satisficing

Survey researchers have long known that individual and contextual factors can influence the quality of answers to survey questions. For the past 25 years, this understanding has been increasingly refined. One of the most influential contributions in this regard has been Krosnick's theory of survey satisficing, which has become the dominant framework for understanding data quality in surveys from a measurement-error perspective (Turner, Sturgis, & Martin, 2014).

The theory of satisficing offers a useful framework for exploring sub-optimal survey responses. In short, survey respondents satisfice when they fail to fully engage in one or more of the four stages of cognitive processing: comprehending the question, retrieving relevant information from memory, integrating information to arrive at a judgment, and formulating and

editing a response (Krosnick, 1991; Tourangeau, Rips, & Rasinski, 2000). Respondents may satisfice through a variety of strategies, such as nondifferentiation or speeding. The amount of satisficing can range from little or no effort to more substantial but still not maximal effort (Krosnick, 1991). The level of satisficing in turn results in different effects on data quality.

Krosnick (1991, 1999) distinguished two forms of satisficing, weak and strong. Weak satisficing describes the situation in which the four cognitive stages of survey response (comprehension, recall, retrieval, and judgment) are undertaken, but less thoroughly than when optimizing occurs (Krosnick, 1991, 1999). Respondents who engage in weak satisficing, may put less effort into understanding the meaning of the questions, search their memories less thoroughly for relevant information, integrate the retrieved information carelessly, or select a response imprecisely. An example of weak satisficing is selecting the first reasonable option from a list rather than considering all options and selecting the most appropriate.

Strong satisficing occurs when respondents simplify the answer process by skipping the retrieval and judgment steps altogether, but still attempt to provide answers that are acceptable or seem reasonable (Krosnick, 1991, 1999). When this happens, respondents may look to the wording of the question for a cue pointing to an easy answer that does not require much, or any, thought. If no such cue is present, the respondent may arbitrarily select an answer. Respondents randomly selecting response options in a multi-item choice question is an example of strong satisficing.

The propensity to use a satisficing strategy is thought to be determined by three factors: respondent ability, respondent motivation, and task difficulty (Krosnick, 1991). Ability is related to the extent to which respondents can perform complex mental operations. Motivation is influenced by the need for cognition, the degree to which the topic of a question is personally

important, the extent of respondent fatigue, and aspects of questionnaire administration (such as interviewer behavior) that either encourage optimizing or suggest that thoughtful reporting is not necessary (Bethlehem & Biffignandi, 2012). Task difficulty is a function of the attributes of the questions (e.g., the difficulty of interpreting a question) and how the questionnaire was administrated (e.g., the pace at which an interviewer reads the questions). Other aspects of the task can contribute to increasing the tendency to satisfice such as the length of the questionnaire, long lists of response alternatives, and mode of data collection (Krosnick & Alwin, 1987). All of these factors can have different effects on the accuracy and quality of the data obtained (Bowling, 2005).

Previous research has produced an important body of evidence regarding the negative impact of satisficing response behaviors on the quality of cross-sectional survey data. Oppenheimer et al. (2009), concluded that satisficing participants, by providing answers that do not accurately answer the questions, decrease the signal-to-noise ratio of a data set and can substantially lower the power of hypothesis test procedures. Barge and Gehlbach (2012) noted that satisficing was associated with differences in the distributions of respondents on single-item key indicators. For example, a result in their study showed that although the percentage of non-satisficers reporting a grade-point average of 10 (equivalent to A on the common US scale) was only 2.4%, it was 76% in the case of very strong satisficers. These kinds of effects were also observed in attitudinal, behavioral, and factual items.

Barge and Gehlbach also reported that as satisficing became more pronounced, the reliabilities of psychometric scales and the correlations between them tended to increase. These results are likely due in large part to the impact of straightlining. The authors stated that, "In the most extreme cases, satisficing respondents can negatively influence the data enough to

introduce correlations where, in fact, none exist." Hamby and Taylor (2016) examined the psychometric consequences of satisficing. They found evidence that satisficing in the first part of the questionnaire was associated with improved internal consistency, reliability, and convergent validity, and with poorer discriminant validity for scales located at the end of the questionnaire.

2.1.2 Satisficing in Multi-Item Scales and its Consequences for Data Quality

As mentioned earlier, multi-item scales are ubiquitous in surveys, and are important in assessing a variety of constructs. Previous research has identified different satisficing strategies associated with multi-item scales, covering both weak and strong forms of satisficing. Weak satisficing associated with multi-item scales includes response order effects (Yan & Keusch, 2015) and response styles – i.e., the tendency of respondents to select a certain response category regardless of content. Strong satisficing related with multi-item scales mainly include random responses, non-differentiation of the scale directions, and straightlining. Random responses arise when respondents blindly answer questions by choosing responses without putting thought into the question response. Non-differentiation of scale direction occurs when inattentive respondents do not notice a change of the question item directions. Straightlining describes situations where respondents select the same response option in the grid for all or most of the questions without distinguishing individual question content. These forms of strong satisficing originate from respondent inattentive response and are referred to as careless responding in the survey literature. This chapter mainly focuses on the former two types of strong satisficing in multi-item scales.

As suggested in previous literature, practical estimates of careless response in multi-item scales typical studies range from as low as about 1% (CPP, 2002) to as high as 30% (Burns, Christiansen, Morris, Periard, & Coaster, 2014). Careless response is an important issue because

inclusion of even a low proportion of these responses in datasets impacts the usefulness of the data (Huang, Liu, & Bowling, 2015; Maniaci & Rogge, 2014; Woods, 2006). The correlation between two scores that measure a similar construct is attenuated when participants provide careless responses. This is because inconsistent responses increase the amount of measurement error and obscure true relationships among the variables (Widhiarso & Sumintono, 2016).

Careless responses can have serious psychometric implications as well. Random responses and non-differentiation of the item wording direction constitute error variance, which attenuates correlations, reduces internal consistency, and potentially results in erroneous factor analysis results (Meade & Craig, 2012). Johnson (2005) illustrated how factor structures differed for subsamples of survey participants identified as CR. Additionally, careless response on reverse-coded items can contribute to the presence of so-called "method" factors, in which positively worded items for a given scale load onto one factor, while negatively worded items for the same scale load onto another (Woods, 2006). Woods found that a single-factor confirmatory model did not fit in such instances when as little as 10% or 20% of respondents were from CR on reverse coded items (see also Huang, Curran, Keeney, Poposki, & DeShon, 2012).

2.1.3 Careless Responding Detection Methods

Methods of screening for careless responding can be broken into roughly four types. The first type involves the use of motivation filtering. This technique requires the use of self-report motivation measures that collect respondent opinions about how important the scale or questionnaire was to them and the amount of effort exerted to complete it (Wise & DeMars, 2005). The scores provided by these measures can be used along with a "cutoff score" to classify respondents as CR.

The second screening method requires special items or scales to be inserted into the survey prior to administration. In the context of web surveys, an easy-to-use technique has emerged and gained rapid popularity, trap questions or instructional manipulation checks. Most trap questions, although they may differ in their formats, follow a similar pattern. They have a "lure" question with "lure" responses and an instruction asking respondents to ignore the lure question and provide a specific response (Hauser & Schwarz, 2015).

The third approach is a measure of response-time effort (RTE). The RTE measures the amount of time respondents spend completing each item on the questionnaire (Wise & DeMars, 2005). RTE assumes that respondents will answer items using a rapid-guessing behavior, which involves responding without taking sufficient time to consider the item content and response options. Similar to motivation filtering, RTE can be used to classify respondents as CR by specifying item thresholds to determine the presence of rapid-guessing behavior (Wise & Kong, 2005).

A fourth approach is the use of a person-fit analysis. Person-fit analysis is a broad set of statistical methods used to identify inconsistencies in patterns of an item scores on a test or scale (Meijer & Sijtsma, 2001). For example, a respondent that selects items on a depression scale that reflect severe symptoms (e.g., suicidal ideation), but none of the milder symptoms (e.g., feeling hopeless or pessimistic) has an inconsistent response pattern. When a response pattern is deemed inconsistent, it is then assumed that the responses to the survey items are guided by a response mechanism other than the construct specified (Meijer, 2002). For example, a person randomly selecting a response to items in order to get to the end of the questionnaire faster would produce person-misfit since the mechanism of selecting items is guided purely by guessing (Felt, Castaneda, Tiemensma, & Depaoli, 2017).

2.1.4 Person-Fit Statistics and the l_z^p Method

Person-fit statistics (PFS) were originally developed to detect invalid test scores in cognitive and educational measurement. PFS were designed to detect a lack of motivation in test completion or scoring errors (Levine & Drasgow, 1982; Meijer & Sijtsma, 2001). However, PFS also have been applied in the context of personality and attitudinal measurements to detect CR, response styles, and respondent data falsification (e.g. Conijn, Emons & Sijtsma, 2014; Emons, 2008; Ferrando, 2012; LaHuis & Copeland, 2009; Reise & Flannery, 1996; Woods, Oltmanns, & Turkheimer, 2008; Zickar & Drasgow, 1996).

The existence of many PFS has motivated several researchers to compare the existing PFS in an attempt to find the most statistically powerful (e.g., Glas & Meijer, 2003; Karabatsos, 2003; Tendeiro & Meijer, 2014). For scales including polytomous items, one type of PFS, the standardized log-likelihood l_z^p, developed by Drasgow, Levine, and Williams (1985), has been found to have higher detection rates than others (Emons, 2008). The standardized log-likelihood l_z^p method is especially useful in the identification of individuals who respond inconsistently to one-dimensional scales and its power increases as the scale length (number of items) increases.

2.1.5 The Present Study

Although that the standardized log-likelihood l_z^p is useful to identify careless responses in one-dimensional scales with a large number of items, it also has several important limitations. It does not work well with multi-dimensional scales with a small number of items. This is because when the number of items is small, the amount of information available to identify careless responding is limited. This, in turn, reduces the power and the usefulness of the standardized log-likelihood l_z^p as an index to detect this type of response (Emons, 2008). When the standardized

log-likelihood l_z^p is applied to analyze a questionnaire containing several short scales, this statistical measure must be applied separately for each scale, making it less efficient.

In order to overcome this limitation, Conijn (2013) suggested combining the outcome of the standardized log-likelihood l_z^p applied to different scales after applying the standardized log-likelihood l_z^p to each scale separately. Conijn simulated data of five scales, each of which consisted of 12 items with five answer categories each. The answers to each of the five scales were generated following the Samejima's Graded Response Model (Samejima, 2016). A subsample comprised of 30% of the cases of one of the scales was randomly selected. For this subsample, six of the 12 items for the selected scale were replaced with "contaminated" data (i.e., answers consistent with careless responding). Conijn's method detected only 32% of the contaminated cases, or a 68% false negative rate. These limitations of the standardized log-likelihood l_z^p have not been well addressed. New methods to better detect careless responses in short scales (i.e., less than 20 items) are needed.

To address the need for a method that can identify a larger proportion of CR in scales with a small number of items, a method initially used in engineering to identify anomalies or outlier cases, the autoencoder neural network, is used here. A goal of this research is to explore the use of the autoencoder to identify CR and compare the performance of the standardized log-likelihood l_z^p and the autoencoder in terms of sensitivity and false positive rate in a simulation study. Two strong forms of careless responding behaviors are examined: random response and non-differentiation of question item directions. These behaviors can negatively impact survey data quality. One common feature of these two behaviors, similar to other types of satisficing, is that they both lead to response patterns that are inconsistent with the majority of the sample. Given that the standardized log-likelihood l_z^p and the autoencoder methods are designed to detect

inconsistent response patterns or anomalies, it is expected that both methods will detect CR. The standardized log-likelihood l_z^p uses a model-based approach and detects item response patterns that are inconsistent with the population model underlying the multi-item scale. The autoencoder uses an unsupervised neural network framework designed to detect outlier response patterns (see Section 2.2).

Compared to other methods to identify CR, such as RTE and straightlining detection, the standardized log-likelihood l_z^p and the autoencoder methods have several potential advantages. The response time method depends on the use of a threshold to identify speeders. Given that respondents with different characteristics such as age and cognitive functioning, vary in their ability to read questions and provide responses, a speeding threshold for respondents with lower cognitive ability may be normal response time for those with higher cognitive ability. This response heterogeneity makes it difficult to define a single threshold. In addition, there is no standard approach to determine the threshold. Some studies define the threshold based on RTE by finite mixture models, while others use the calculations based on reading time for each word (Zhang & Conrad, 2013). In some cases, the decision on the threshold can be somewhat arbitrary and difficult to standardize.

Straightlining detection methods only work well in scales that include both positively and negatively-worded or reverse-coded items. In scales with items worded in the same direction, it is challenging to disentangle the real response patterns having a "straight–line" outcome (e.g., someone truly feel strongly agrees to all items in a scale using agree-disagree response options) and satisficing behaviors with straightlining at the "strongly agree" categories. In contrast, the standardized log-likelihood l_z^p and the autoencoder methods can work with both types of scales –

scales with same-direction worded or non-reverse coded items and scales with reverse-coded items.

2.2 The Standardized Log-Likelihood l_z^p

The standardized log-likelihood l_z^p is based on the Graded Response Model (GRM). Consider an $N \times J$ matrix where N is the number of persons and J is the number of items (items are indexed j; $j = 1,\ldots, J$). Responses to item j are categorized into $m_j + 1$ ordered categories, where higher levels of the response scale indicate more of the latent trait that is being measure by the scale. Associated with each of item j's response categories is a category score, x_j , with integer values 0, 1, …, m_j . The GRM specifies the odds for a person selecting a category corresponding to the score x_j or higher (de Ayala, 2008). According to the GRM, the probability of selecting the category corresponding to the score x_j or higher is given by

$$P_{x_j}^*(\theta) = P_{x_j}^* = \frac{\exp[\alpha_j(\theta - \delta_{x_j})]}{1 + \exp[\alpha_j(\theta - \delta_{x_j})]} \quad (1)$$

where θ is the latent variable trait, α_j is the discrimination (slope) parameter for item j, and δ_{x_j} is the category boundary location for category score x_j ($x_j = 0, 1, \ldots, m_j$). The category boundary location is the boundary between categories corresponding to scores x_j and x_{j-1}. In the GRM the parameters δ_j's are always in increasing order and there are m_j category boundary locations for item j. By definition, the probability of responding in the lowest category or higher is 1.0 ($P_0^* = 1.0$) and the probability of responding in category $m_j + 1$ or higher is 0 ($P_{m_j+1}^* = 0$).

To obtain the probability of an individual obtaining a particular category-score x_j, it is necessary to take the difference between the probabilities $P_{x_j}^*$ for adjacent categories. That is,

$$P_{x_j} = \begin{cases} P_0^* - P_1^* = 1 - P_1^* & \text{if } x_j = 0 \\ P_{x_j}^* - P_{x_j+1}^* & \text{if } 1 \le x_j \le m_j - 1 \\ P_{x_j}^* - P_{x_j+1}^* = P_{x_j}^* - 0 & \text{if } x_j = m_j \end{cases}$$

The sum of the P_{x_j}'s across the response options for a fixed value of θ is

$$\sum_{x_j=0}^{m_j} P_{x_j} = 1$$

Let d_{x_j} (m) = 1 if $x_j = m$ (m = 0, 1, …, m_j), and 0 otherwise. The unstandardized log-likelihood person-fit statistic for polytomous items, l^p, is given by

$$l^p(\mathbf{x}) = \sum_{j=1}^{J} \sum_{m=0}^{m_j} d_{x_j}(m) \ln P_{x_j} \tag{3}$$

The standardized version of l^p, that is, l_z^p is given by

$$l_z^p = \frac{l^p - E[l^p]}{\sqrt{Var[l^p]}} \tag{4}$$

where E[l^p] is the expected value of l^p, given by

$$E[l^p] = \sum_{j=1}^{J} \sum_{x_j=0}^{m_j} P_{x_j} \ln P_{x_j} \tag{5}$$

and $Var[l^p]$ is the variance of l^p, given by

$$Var[l^p] = \sum_{j=1}^{J} \left[\sum_{x_j^*}^{m_j} \sum_{x_j}^{m_j} P_{x_j^*} P_{x_j} \ln \left(\frac{P_{x_j^*}}{P_{x_j}} \right) \right] \tag{6}$$

Large negative values of the standardized log-likelihood l_z^p are indicative of a misfit. When the true θ value is used to compute the standardized log-likelihood l_z^p, this statistic follows a standard normal distribution. However, Nering (1995) showed that when the unknown true θ value is replaced by an estimated $\hat{\theta}$ value, the standardized log-likelihood l_z^p is no longer standard normal. In this case, some type of Monte Carlo method is needed to compute a true p-value to decide if the standardized log-likelihood l_z^p value indicates of misfit.

2.3 Specifications of the Autoencoder Neural Network

2.3.1 Neural Network

Artificial neural networks are computational models inspired by the nervous systems of living beings. A neural network is constructed from a number of interconnected nodes known as neurons. A typical feed-forward neural network may be drawn as in Figure 2.1. In the figure, each circle is a neuron, with incoming arrows being the neuron's inputs and outgoing arrows the neuron's outputs. Each neuron is connected to all of the neurons in the next layer in what is called a fully connected layer.

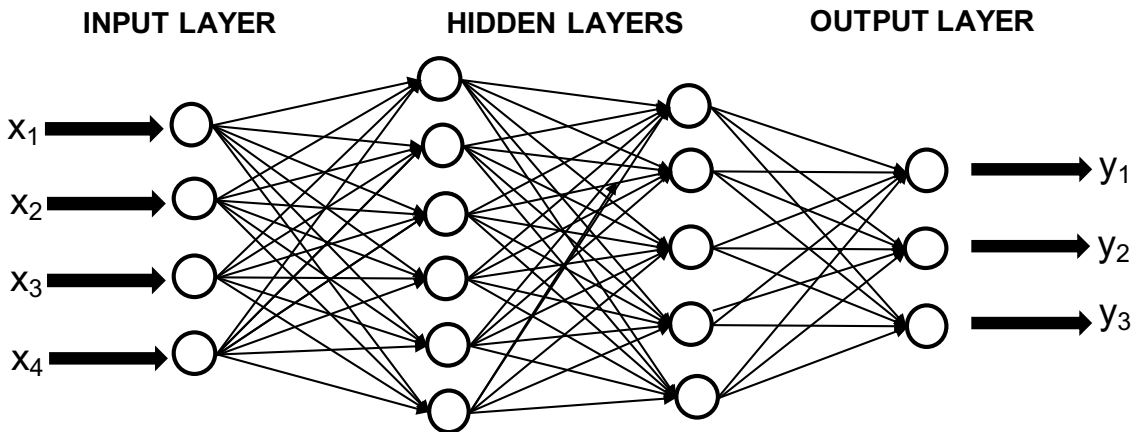

Figure 2.1. Basic feed-forward neural network.

Neurons are arranged in layers reflecting the flow of information. The first layer has no incoming arrows and is the input to the network. The input layer neurons correspond to the number of predictor variables that will be analyzed in the network. The right-most layer, the network output, has no outgoing arrows. The number of output nodes corresponds to the number of dependent variables that need to be predicted or classified. The other layers are considered "hidden." Each neuron in the hidden layers is represented by a nonlinear activation function (e.g., the logistic function) that is applied to the neuron's value before passing it to the output (Nunes, Hernane, Andrade, Bartocci, & dos Reis, 2017).

2.3.2 Artificial Neuron

In Figures 2.2 and 2.3 the basic unit of an artificial neural network, a single-input neuron is represented. Here, i corresponds to the input variable ($i = 1, \dots, n$), and j corresponds to the hidden layer where the neuron is located.

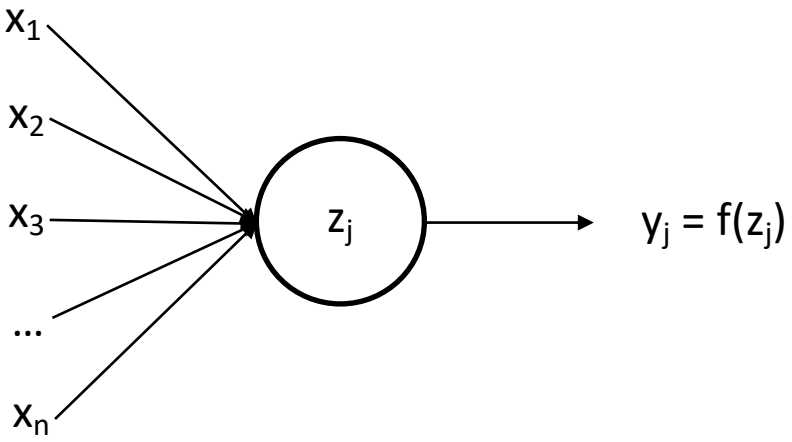

Figure 2.2. An artificial neuron.

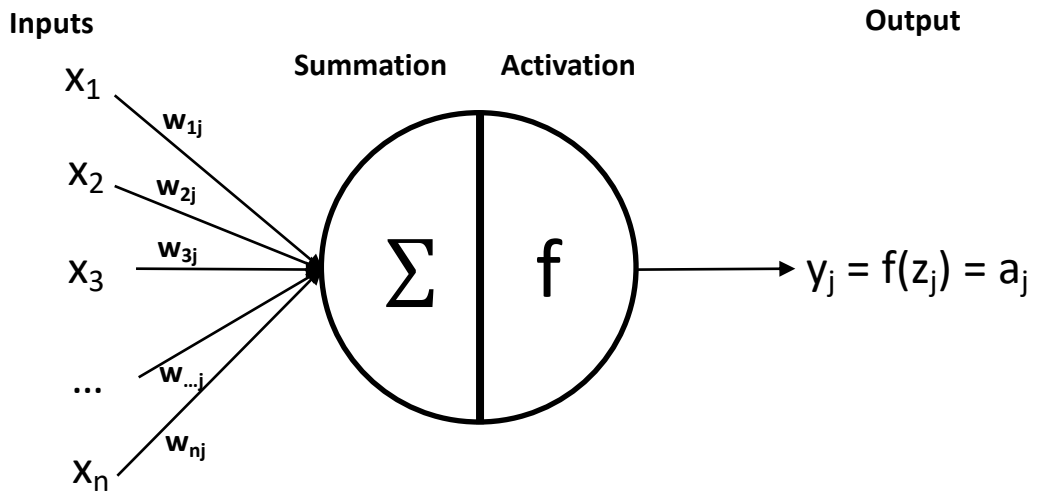

Figure 2.3. Alternative representation of an artificial neuron.

Given a vector sample of *n* input variables, $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, the output of the neuron *j* will be the result of two computation steps. In the first step, a vector of *n* weights, $\mathbf{w_j} = [w_{1j}, w_{2j}, \ldots, w_{nj}]$, and a scalar b_j called bias are used to compute a weighted sum z_j of the input vectors as:

$$z_j = \sum_{i=1}^{n} w_{ij} x_i + b_j = \mathbf{w_j^T x} + b_j \tag{7}$$

In the second step, the output a_j, of the neuron *j* is calculated as:

$$a_j = f(z_j) = f\left(\sum_{i=1}^{n} w_{ij} x_i + b_j\right) = f\left(\mathbf{w_j^T x} + b_j\right) \tag{8}$$

The output a_j depends on the specific activation function *f* selected for the neuron. The weights $\mathbf{w_j}$ and bias b_j are both adjustable parameters of neuron *j*.

Thus a typical neuron is composed of five basic elements:

- Input variables $(x_1, x_2, \ldots, x_n)$ are the samples coming from the external environment and represent the values assumed by the variables of a particular application. These input variables are usually standardized in order to enhance the computational efficiency of the learning algorithm.
- Weights $(w_1, w_2, \ldots, w_n)$ are the values used to weight each one of the input variables, which enables the quantification of their relevance with respect to the functionality of the neuron.
- The linear aggregator (Σ) gathers all input variables weighted by the weights and the bias to produce an output z_j.
- Activation function (f) whose goal is limiting the neuron output a_j within a reasonable range of values, for example, [0, 1] or [0, ∞).
- Neuron output (a_j) consists of the final value produced by the neuron given a particular set of input variables and can be used as input for other sequentially interconnected neurons.

Typically, the activation function f is chosen by the researcher and then the parameters $\mathbf{w_j}$ and b_j will be modified by a training function such that the neuron input/output relationship meets a specified objective.

The activation function f in Figures 2.2 and 2.3 can be a linear or a nonlinear function of the input vector **x**. The particular activation function used in the network is selected to satisfy the problem specification. Some of the most commonly used functions include the s*igmoid/logistic activation function* (Figure 2.4), which takes the input z_j and transforms it into a value in the range between [0, 1], expressed as:

$$a_j = f(z_j) = \frac{1}{1 + \exp^{-z_j}} \tag{9}$$

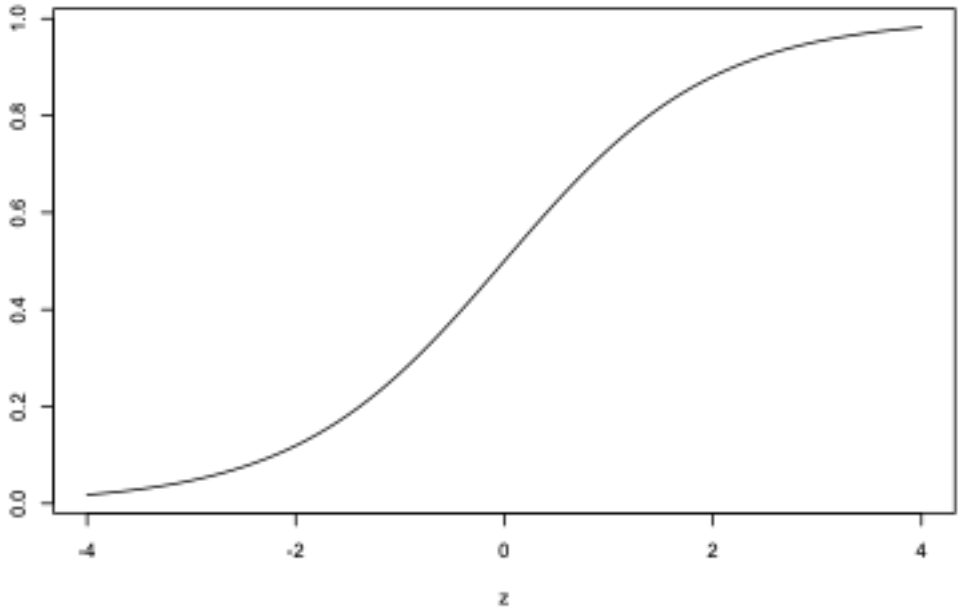

Figure 2.4. Sigmoid activation function.

Because the output produced by the sigmoid activation function is always in the range [0, 1], this function is an ideal choice as an activation function for binary classification problems.

The *hyperbolic tangent activation function (tanh)* is similar to the sigmoid activation function but with the output zero-centered value in the range [-1, 1] (see Figure 2.5). The tanh function is expressed as

$$a_j = f(z_j) = \frac{\mathrm{e}^{z_j} - \mathrm{e}^{-z_j}}{\mathrm{e}^{z_j} + \mathrm{e}^{-z_j}} \tag{10}$$

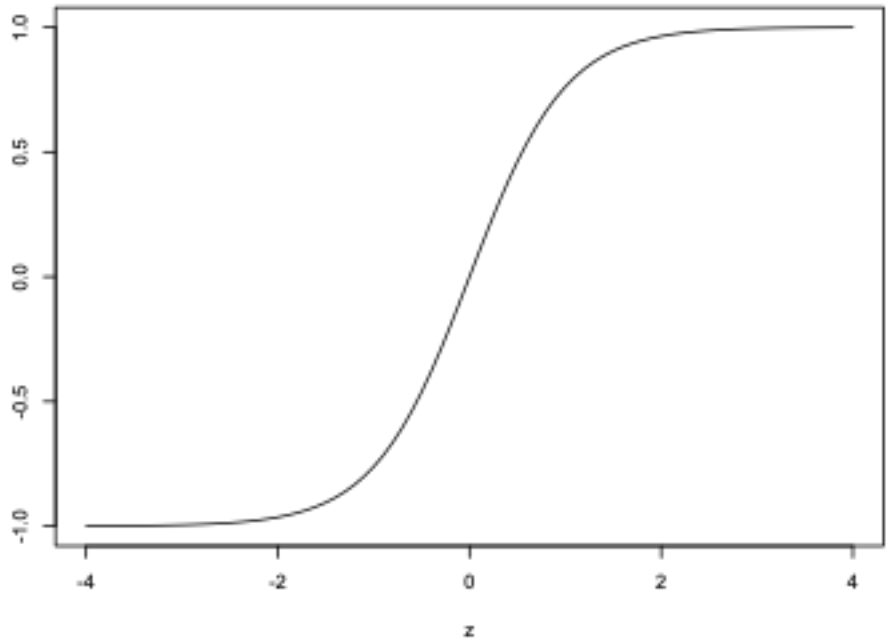

Figure 2.5. Hyperbolic tangent activation function.

The *rectified linear unit activation function (ReLu)* is a computationally simple and efficient linear function (see Figure 2.6). ReLu is expressed as $a_j = f(z_j) = \max(0, z_j)$. In the case of the sigmoid and tanh functions, all the neurons within the hidden units fire during model convergence. However, in the case of ReLu, some of the neurons will be inactive (for the negative input values) and hence the activations are sparse and efficient.

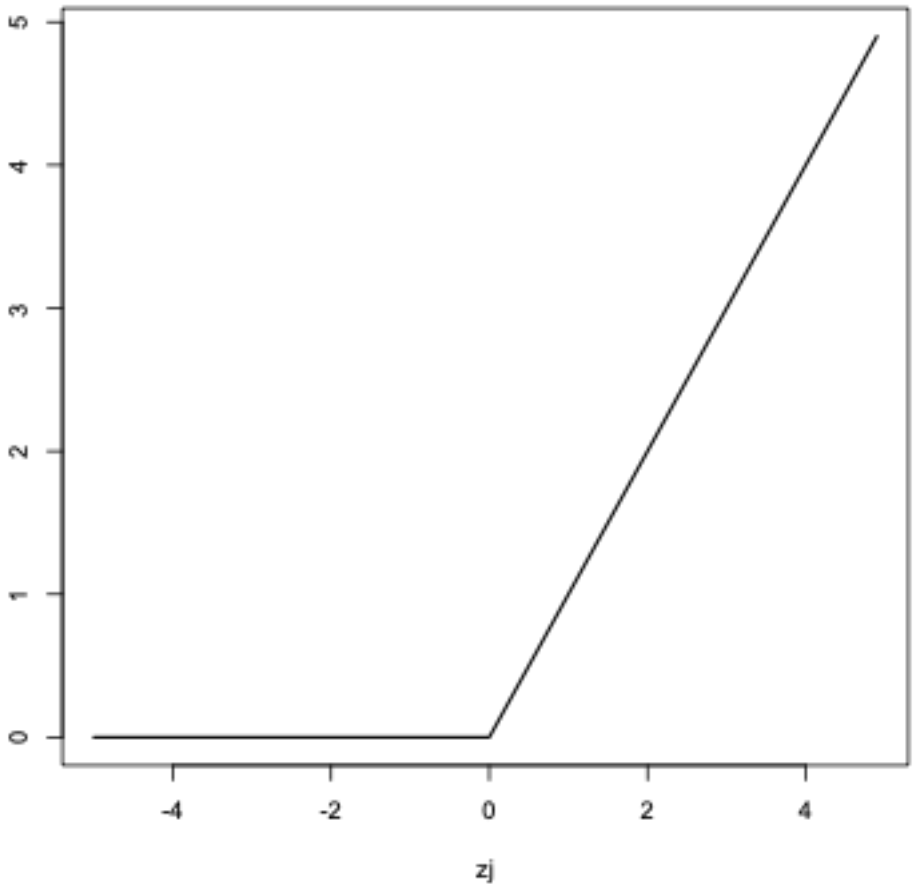

Figure 2.6. ReLu activation function.

The nonlinearity property of these activation functions is essential in order to model complex data patterns to solve regression and classification problems accurately.

2.3.3 Network Architectures

In some circumstances, one neuron, even with many inputs, might not be sufficient for modeling complex patterns in the input variables. Therefore, many neurons operating in parallel are needed (Hagan, Demuth, Beale, & De Jesús, 2014). Suppose, for example, that the input vector **x** corresponds to data from one single training case for four variables, x_1, x_2, x_3, and x_4.

These inputs are in Figure 2.1 connected with all the six neurons in the first hidden layer. The connection can be expressed as a matrix multiplication of a 4-dimensional input vector and a 6×4 matrix of weights, operated on by a nonlinear transformation *f*:

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]} \tag{11}$$

$$\mathbf{a}^{[1]} = f^{[1]}(\mathbf{z}^{[1]}) = f^{[1]}\left(\mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}\right) \tag{12}$$

The superscript indicates the specific layer, with the input layer is defined as layer 0, and the first hidden layer as layer 1 ($h^{[1]}$).

In a neural network each element of the *n*-dimensional input vector **x** is connected to each of the *q* neurons in $h^{[1]}$. As a result, the weight matrix for this layer $\mathbf{W}^{[1]}$ has *q* rows and *n* columns, and the bias vector $\mathbf{b}^{[1]}$ corresponds to a *q*-dimensional vector.

In general, it is common that the number of inputs to a layer (not only the first one) are different from the number of neurons in that layer. Also, it is not necessary that all layers have the same activation function. The same activation function is necessary for all neurons in a layer.

When a neural network includes more than one hidden layer, each layer has its own individual weight matrix $\mathbf{W}^{[k]}$, bias vector $\mathbf{b}^{[k]}$, and output vector $\boldsymbol{a}^{[k]}$. For example, the second layer ($h^{[2]}$) showed in Figure 2.1 includes five neurons which are fully connected to the six input neurons. Therefore, the matrix $\mathbf{W}^{[2]}$ will be a 5×6 weight matrix, and the vector bias $\mathbf{b}^{[2]}$, and the vector $\boldsymbol{a}^{[2]}$ will have five dimensions.

The process of forwarding the outputs of layer *k* as inputs for layer $k + 1$ can be written as:

$$\mathbf{a}^{[k+1]} = f^{[k+1]}\left(\mathbf{W}^{[k+1]}\mathbf{a}^{[k]} + \mathbf{b}^{[k+1]}\right) \tag{13}$$

for $k = 0,1, \ldots, K$, where K is the number of layers in the network and $f^{[k]}$ is the activation function in layer k.

The neurons of the first layer receive external inputs

$$\mathbf{a}^{[0]} = \mathbf{x} \tag{14}$$

the starting point for Equation 13. The outputs of the neurons in the last layer are the network outputs

$$\mathbf{y} = \mathbf{a}^{[K]} \tag{15}$$

and the output of the neural network shown in Figure 2.1 can be calculated as

$$\mathbf{y} = \mathbf{a}^{[3]} = f^{[3]}\left(\mathbf{W}^{[3]} f^{[2]}\left(\mathbf{W}^{[2]} f^{[1]}\left(\mathbf{W}^{[1]}\mathbf{a}^{[0]} + \mathbf{b}^{[1]}\right) + \mathbf{b}^{[2]}\right) + \mathbf{b}^{[3]}\right) \tag{16}$$

Training algorithms. A training algorithm or learning rule refers to a procedure which adjusts the weights and biases of a neural network. There are two types of learning rules, supervised and unsupervised.

In *supervised learning*, the network is trained by providing a set of input data (the training set)

$$(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), (\mathbf{x_3}, y_3), \ldots, (\mathbf{x_m}, y_m)$$

where $\mathbf{x_i} = [x_{i1}, x_{i2}, \ldots, x_{in}]$, for input elements i = 1, 2, …, m, is a n-dimensional input vector to the neural network; y_i is the corresponding target output; and m is the sample size of the training sample. The vectors $\mathbf{x_i}$ are combined into a matrix $\mathbf{X_{m,n}} = [\mathbf{x_1}^T, \mathbf{x_2}^T, \ldots, \mathbf{x_m}^T]^T$, and the scalars y_i are combined into a vector $\mathbf{y} = [y_1, y_2, \ldots, y_m]^T$.

After applying the inputs, the network outputs $\boldsymbol{y}$ are compared to the targets in **x**. The training algorithm modifies the weights and biases of the network to reduce the difference (error) between the network output and the target.

In *unsupervised learning*, the weights and biases are modified in response to network inputs only, since there are no target outputs available.

2.3.4 Autoencoders

Autoencoders are unsupervised learning methods on neural networks. Architecturally, the simplest form of autoencoder is a non-recurring neural network with an input layer, an output layer, and one or more hidden layers that connect them (Figure 2.7).

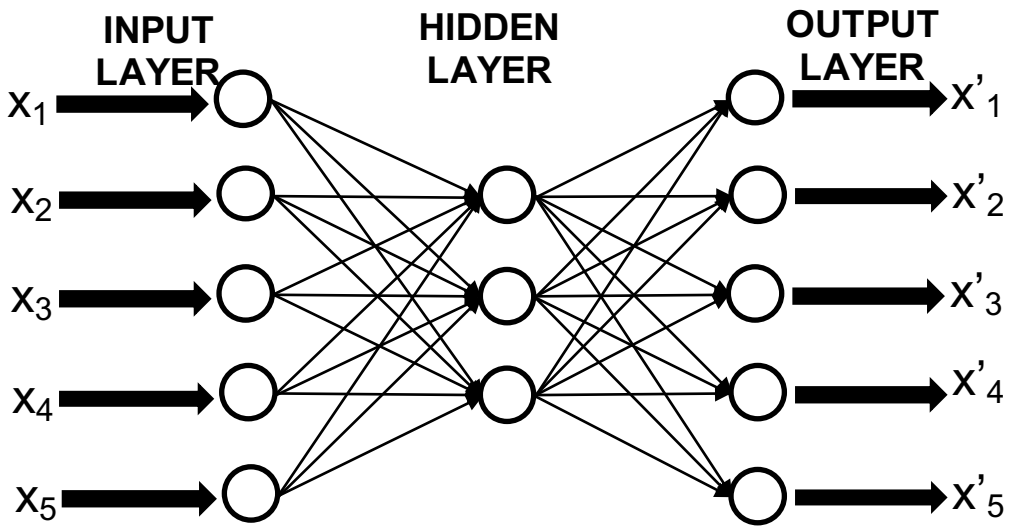

Figure 2.7. Autoencoder with a single hidden layer.

The number of hidden neurons is less (or more) than the number of input neurons. It is different from other neural networks, because the output layer has as many nodes as the input

layer, and instead of training to predict some target vector **y** given the matrix input **X**, an autoencoder is trained to reconstruct its own inputs **X**.

The basic idea of an autoencoder is to have an output layer with the same dimensionality as the inputs. The idea is to reconstruct each element $\mathbf{x_i}$ exactly by passing it through the network. An autoencoder replicates the data from the input to the output. Although reconstructing the data by a trivial copying of data forward from one layer to another, this is not possible when the number of units in the middle layers are constricted. In other words, the number of units in each middle layer is typically fewer than the input (or output). As a result, these units hold a reduced representation (a regression) of the data, and the final layer can no longer reconstruct the data exactly. A loss function of this neural network uses the sum-of-squared differences between the input and the output. The algorithm forces the output to be as similar as possible to the input, my minimizing the loss function.

It is common (but not necessary) for a *K*-layer autoencoder to have a symmetric architecture between the input and output where the number of units in the *k*-th layer is the same as that in the (K - k + 1) layer. Furthermore, the value of *K* is often odd, as a result of which the (K + 1)/2 layer is often the most constricted layer. In this expression, *K* includes the input layer **x** = $a^{[0]}$, and therefore the minimum number of layers in an autoencoder would be three: the input layer **x**, the constricted layer $a^{[1]}$, and the output layer $a^{[2]}$. That is, the autoencoder will take an input matrix **X**, reduce it, and return it back up to its original size. By analyzing how well the model rebuilt the data to the original input-space size, it is possible to determine what falls within a threshold of pattern acceptability and what doesn't. The units that cannot be reconstructed correctly receive a large reconstruction error score, which can then be used to

classify each case as an anomaly or outlier. For the purposes of the present research, anomaly or outlier case in multi-item scale applications can be also classified as careless responses.

2.4 Simulation Study

This simulation study was conducted to compare the ability of the standardized log-likelihood l_z^p and the autoencoder to detect respondents providing careless responses to a scale with a small number of items. A one-factor confirmatory factor analysis (CFA) model with categorical ordered indicators was used to generate the data for this simulation (see Figure 2.8 for an example with six items).

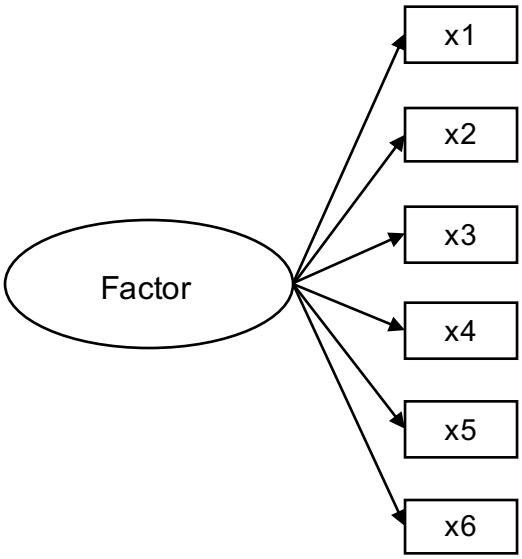

Figure 2.8. Example of a one factor CFA model with six items.

Threshold parameter values between item levels were chosen to obtain an item distribution that matched the skewness and kurtosis of a standard normal distribution. Table 2.1 shows the values that were used in this study. The levels chosen for each condition, and the reasons for the level selections, are described in the following subsection.

Table 2.1. Threshold parameter values for four and seven categories.

	Categories	
Threshold	*Four*	*Seven*
1	-1.25	-1.79
2	0.00	-1.07
3	1.25	-0.36
4		0.36
5		1.07
6		1.79

2.4.1. Design Characteristics

A full factorial design crossed each of six study factors to create 96 experimental conditions. The six factors were: 1) types of careless response (*random response behavior* and *non-differentiation of item direction changes* for mixed items with both positive and negative wordings); 2) scale lengths (6 and 12 items); 3) different percentages of careless responses in the sample (10%, 20%, and 30%); 4) percentage of careless responses (half items showing careless responding or all items); 5) factor loadings of the CFA model (0.4 and 0.6); and 6) number of response categories per item (4 and 7). For each condition, 1000 datasets were generated and analyzed.

Types of careless responding. Two strong forms of satisficing, random response behavior and non-differentiation of item wording direction changes, were examined. Among the two, non-differentiation of item wording direction changes only applies to mixed items with both positive and negative wordings. We have no prior expectation about which form of satisficing will produce better detection by either the standardized log-likelihood l_z^p or the autoencoder.

Scale lengths. Given that the purpose of this study is to evaluate how well the standardized log-likelihood l_z^p and the autoencoder perform in scales with a small number of items, six and 12 items were chosen as suitable levels. These are considered in the literature as a

small number of items (Emons, 2008; Conijn et al., 2014). We expect higher levels of accurate detection of CR with more items, and with more items involved in careless responding.

Different percentages of careless responses in the sample. The three percentages of careless responses in the sample (10%, 20%, and 30%) were chosen because 1) if the proportion of careless response is too low, the negative impact on the data quality is likely to be minimal (thus 10% as the lowest percentage); 2) it is unlikely that careless response will be much larger than 30%; and 3) this range of percentages is also consistent with literature evaluating satisficing behaviors (e.g., percentage of respondents failed the trap questions in surveys (Curran et al., 2010; Johnson, 2005; Meade & Craig, 2012)). We expect that higher levels of careless responses will be more accurately detected by both the standardized log-likelihood l_z^p and the autoencoder.

Percentage of items subject to careless responses. Similar to Conijn (2013) and Emons (2018), careless responding behavior could occur to all the items (i.e., six of six or 12 of 12 items), or half of the items (i.e., three of six or six of 12 responses). We expect that more items involved in careless responding will lead to more accurate detection levels by both the standardized log-likelihood l_z^p and the autoencoder.

Factor loadings of the CFA model. Previous studies (Meijer, Molenaar, & Sijtsma, 1994; Meijer & Sijtsma, 2001) showed that the power of PFS depends on the discrimination power of the items, with higher discriminations (i.e., higher factor loadings) associated with higher detection rates. To investigate the effect of item discrimination, two loadings levels, low (0.4) and medium (0.6), were considered. It is very unlikely in practice to have a loading as high as 0.8 consistently across all the items. Higher loads should lead to more accurate detection of CR.

Number of response categories per item. For the number of categories per item, four and seven categories were chosen, because these are in the range most commonly used in surveys.

More categories may lead to more accurate detection levels for either the standardized log-likelihood l_z^p or the autoencoder.

2.4.2. Dependent Variables

Three indicators of the accuracy of the classifications of respondents provided by the standardized log-likelihood l_z^p and the autoencoder were analyzed: sensitivity; the false positive rate; and the total accuracy rate. Sensitivity is the proportion of true CR that were correctly classified as CR. False positive rate is the proportion of true non-CR that were incorrectly classified as CR. The total accuracy rate is the number of absolutely correctly classified instances, either CR classified as CR or non-CR classified as non-CR, divided by the total number of cases.

2.4.3. Data Generation Mechanism

Data for 1,000 replications were generated independently for each of the 2 scale lengths × 2 types of items × 3 percentages of CR in the sample × 2 percentage of CR responses × 2 factor loadings × 2 number of response categories per item = 96 experimental conditions as follows:

- A dataset of polytomous item-score vectors (i.e., six and 12 items) was simulated from a one-factor confirmatory factor analysis (CFA) model with categorical ordered indicators (see Figure 2.8). A latent factor score was generated for each respondent based on a standard normal distribution. Then, continuous latent item responses were generated using the latent factor score and the factor loadings (e.g., 0.4 or 0.6). Finally, these continuous latent item responses were recoded to fpur or seven ordinal categories based

on the threshold parameters presented in Table 2.1[1]. This data-generation process was implemented using the *lavaan* package in R. The sample size for each dataset was 1000[2], respondents.

- A simple random sample without replacement of 10%, 20%, and 30% of respondents from each dataset was chosen and responses were replaced by simulated careless response patterns. For random response the four or seven category response patterns were generated by randomly selecting category responses from a uniform distribution probabilities with probabilities P_{x_j}= 1/4 and P_{x_j}= 1/7, respectively.
- The second careless response pattern is non-differentiation of item direction. This careless response pattern was generated by reverse coding the original answers for one-third or one-half of the items. For example, for a six-item scale with four categories, one-third or two of the items were reverse coded by changing category 1 to category 4, and vice versa, and category 2 to 3, and vice versa. That is, the generation process began with a simple random sample selected without replacement of 10%, 20%, or 30% of respondents. For each, a subset of one-third or one-half of the items were randomly chosen based on the condition and responses to the selected items were reversely coded for each respondent separately.

This new dataset that included the randomly generated careless responses is denoted as the *manipulated or contaminated dataset*.

[1] After generation, the distributions of the categories of the items were confirmed that to be normally distributed. Sees Appendix 2.1.
[2] The Cronbach's alpha coefficients for the data generated based on the CFA models range from 0.50 to 0.86.

2.4.4. Identifying CR based on the Standardized Log-Likelihood l_z^p

The following procedure was applied to each of the manipulated datasets:

- The GRM was applied to the manipulated dataset to estimate item parameter values. For each item j with $m_j + 1$ categories, a discrimination a_j and set of category boundary locations δ_{jk} ($k = 1, 2, \ldots, m_j$) were estimated (see Equation 1) using the *mirt* R package.
- Using the estimated item parameters, the latent score θ_i and its standard error $SE(\theta_i)$ for each respondent ($i = 1, 2, \ldots, 1000$) in the manipulated dataset was computed.
- The PFS value of $l_{z,i}^p$ was computed for each respondent using responses and the estimated respondent parameters θ_i and $SE(\theta_i)$.

Following De la Torre and Deng (2008) and Rizopoulos (2018), a parametric bootstrap method was used to obtain a p-value to classify based on $l_{z,i}^p$ each respondent as a CR or not a CR in the manipulated dataset:

- Generate a new latent score estimate, $\theta_{new,i}$, from a normal distribution with mean θ_i, and standard deviation equal to SE(θ_i).
- Generate a new response pattern of polytomous item-score vectors based on the GRM using the item parameters α_j, δ_{jk}, and the latent score $\theta_{new,i}$.
- For the new response pattern obtained using $\theta_{new,i}$, compute the standardized log-likelihood $l_{znew,i}^p$.

For each respondent in the manipulated dataset, a p-value for its $l_{z,i}^p$ score was computed as

$$p_i = \frac{1 + \sum_{b=1}^{1000} I\left(l_{z_{new},i,b}^p \leq l_{z,i}^p\right)}{1 + 1000}$$

where $I(\cdot)$ denotes an indicator function equal to 1 if $(\cdot)$ is true, and zero otherwise. Following Tendeiro, Meijer, and Niessen (2016), respondents in the manipulated dataset for which $p_i < 0.05$ were classified as CR.

2.4.5. Identifying CR based on the Autoencoder

The autoencoder must be trained to learn the patterns of non-CR, ideally using a perfect dataset without CR. The autoencoder trained on the perfect or higher-quality data can then be applied to a real dataset to identify CR who have large reconstruction errors. Since there is no perfect data completely free of CR available in practice, one critical issue is how to obtain data similar to that of the non-CR.

The approach introduced here is five step process: 1) estimate the CFA parameters model from the original manipulated data, 2) simulate new data based on these estimated parameters, 3) train the autoencoder using this simulated new data to learn non-CR response patterns, 4) apply the trained autoencoder to the original manipulated data to identify CR as respondents with larger reconstruction error, and 5) remove from the manipulated data all respondents identified at some criterion level as CR.

This approach was repeated iteratively to the data, taking the non-CR data from step 5 as the input data for step 1 in a new iteration. The ideal number of iterations was not known. It may be that a larger number of iterations leads to better sensitivity, false positive rates, and total accuracy. Sensitivity, false positive rate, and total accuracy were used to evaluate whether one or more iterations was better in detecting CR in the data.

Figure 2.9 illustrates a four-iteration application of this process. Parameters of a one-factor ordinal CFA model were estimated from the contaminated data that included CR under the

assumption that the simulated data dominated by non-CR will give preliminary estimated CFA parameters close to the true parameter values. The autoencoder was thus trained using simulated data to learn non-CR response patterns. After training, the autoencoder was applied to identify potential CR in the manipulated data for the first time (first iteration). Those identified as CR were then removed from the manipulated dataset, resulting in a reduced dataset.

Assuming that the reduced data includes fewer CR than the original manipulated dataset, parameters for the one-factor ordinal CFA were estimated from the reduced data. These parameter estimates were used to generate a second training dataset. The autoencoder trained using this second training dataset was then applied to the original manipulated data to identify CR in a second iteration.

These steps were then repeated, creating a reduced dataset obtained by removing the identified CR. CFA model parameters were estimated and used to simulate a third training dataset. The autoencoder trained from the third training data set was applied to the initial contaminated data to identify CR in a third iteration. The whole process (remove CR, estimate CFA model parameters, simulate data, train the autoencoder, and then apply the trained autoencoder to the initial manipulated data) was repeated for a fourth iteration as well.

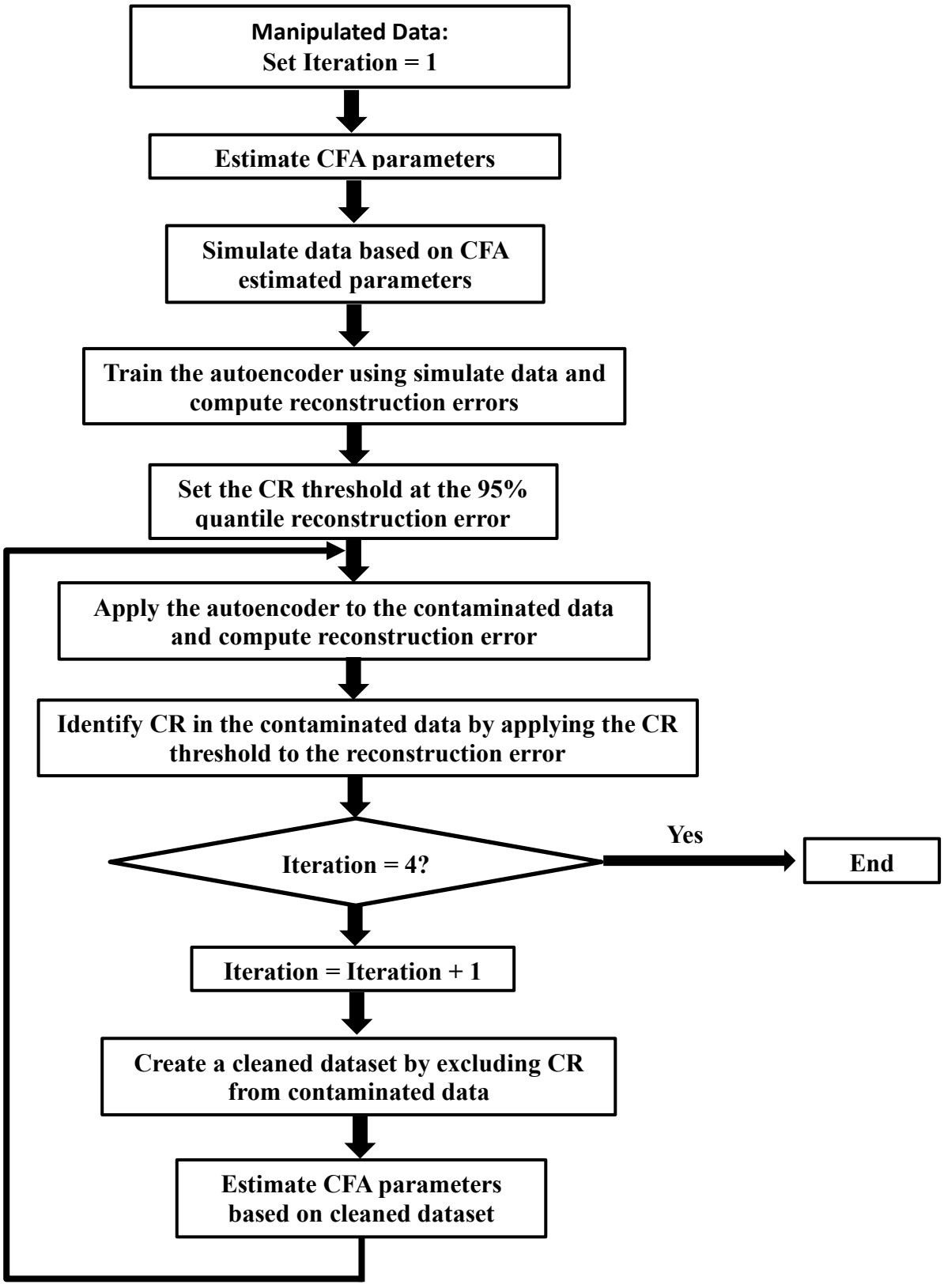

Figure 2.9. Flowchart of the autoencoder approach with up to four iterations.

The application of the autoencoder used one hidden layer with only one neuron. This structure mimics the structure of a one-factor model. The hyperbolic tangent activation function (tanh) was used to encode input data because Goodfellow, Bengio, and Courville (2016) recommend the function for neural networks. The activation function decoding the output data was the rectified linear unit activation function (ReLu), because the output data consisted of non-negative numbers corresponding to four or seven response categories, suitable for non-negative values generated by ReLu. At the end of the final iteration, sensitivity, the false positive rates, and the total accuracy rate were computed.

The R statistical software system (R Core Team, 2018) was used for all the steps in the simulations. The *mirt* package (Chalmers, 2012) was used to estimate the GRM parameters required to compute the standardized log-likelihood l_z^p. The *lavaan* package (Rosseel, 2012) was used to estimate the parameters of the CFA model and to simulate training data for the autoencoder. The CFA parameters were estimated using the Unweighted Least Squares Mean and Variance (ULSMV) estimator. The *keras* package (Allaire & Chollet, 2018) was used to specify the structure and compute the parameters of the autoencoder. All analyses were run in parallel using SLURM (Simple Linux Utility for Resource Management).

2.4.6. Example Illustration of Identifying CR based on the Standardized Log-Likelihood l_z^p and the Autoencoder

Consider two respondents, one non-CR and one CR, identified from the standardized log-likelihood l_z^p or the autoencoder, where for each respondent a five item satisfaction with life scale (SWLS) had been administered (see Appendix 4.1; Diener, Emmons, Larsen, & Griffin, 1985). Suppose as well that each item was rated using a seven-point response category ranging from 1 = "*Strongly disagree*" to 7 = "*Strongly agree*", with higher scores indicating higher

satisfaction. Finally, suppose the non-CR answered each of the five items as 1, 2, 2, 2, and 1, a consistent (similar) response pattern across the items, while the CR answered 4, 4, 1, 1, and 7, inconsistent responses across the five items.

Using the standardized log-likelihood l_z^p, the first respondent was identified as non-CR with an l_z^p value of 1.59, and p-value larger than 0.05. On the other hand, the CR was identified as CR with an l_z^p value of -2.52, and p-value less than 0.05.

The autoencoder method generated a reconstruction error for the non-CR of 0.23. However, the CR had a reconstruction error value of 42.63.

The results of the classification were evaluated using sensitivity, the false positive rate, and the total accuracy. Higher sensitivity indicates that a method can identify a larger proportion of "true" CR, with a sensitivity of 1.0 meaning that the method can identify successfully all CR in the data. The false positive rate is the proportion of the non-CR that are misidentified as CR; lower rates for a method are clearly more desirable. Following Emons (2008) and Conijn (2013), a 0.05 false positive rate is used as the cutoff criterion for the CR identification. Finally, total accuracy is the proportion of all respondents correctly identified as either CR or non-CR; a total accuracy of 1.0 denotes perfect identification.

2.5 Results

2.5.1 Autoencoder Iterations

Figure 2.10 shows sensitivity (2.10A) and false positive rates (2.10B) for the autoencoder at different numbers of iterations and by careless response types (i.e., random response and non-differentiation of item direction changes, for items that include both positive and negative wording). The overall pattern shows the largest increase in sensitivity from one iteration to the

next for iteration 1 to 2. From Iteration 2 to 3 and 3 to 4 the sensitivity increases only slightly. Careless response types have similar patterns as the overall results. The use of the autoencoder in the random response condition (using non-mixed item wordings where all items are worded in the same direction) has higher sensitivity than the non-differentiation of item direction in the mixed items, and higher increase in sensitivity across iterations. That is, the autoencoder works better to find CR that use random responses as opposed to those who do not differentiate positively and negatively worded items.

In Figure 2.10B, the false positive rate increases as the iteration number increases. This pattern is consistent for both careless response types. Similar to the sensitivity results shown in Figure 2.10A, the autoencoder has better false positive rates in random response than non-differentiation of item directions in scales with mixed item directions. The increase in false positive rate rises above 0.05 for both careless response conditions at the third iteration. At iteration 2, the overall false positive rate is slightly below 0.05 with the rate for the non-differentiation condition slightly above 0.05, and the rate for random response condition below 0.05. The autoencoder does not gain much benefit in sensitivity but suffers from the increased false positive rates after two iterations of the process.

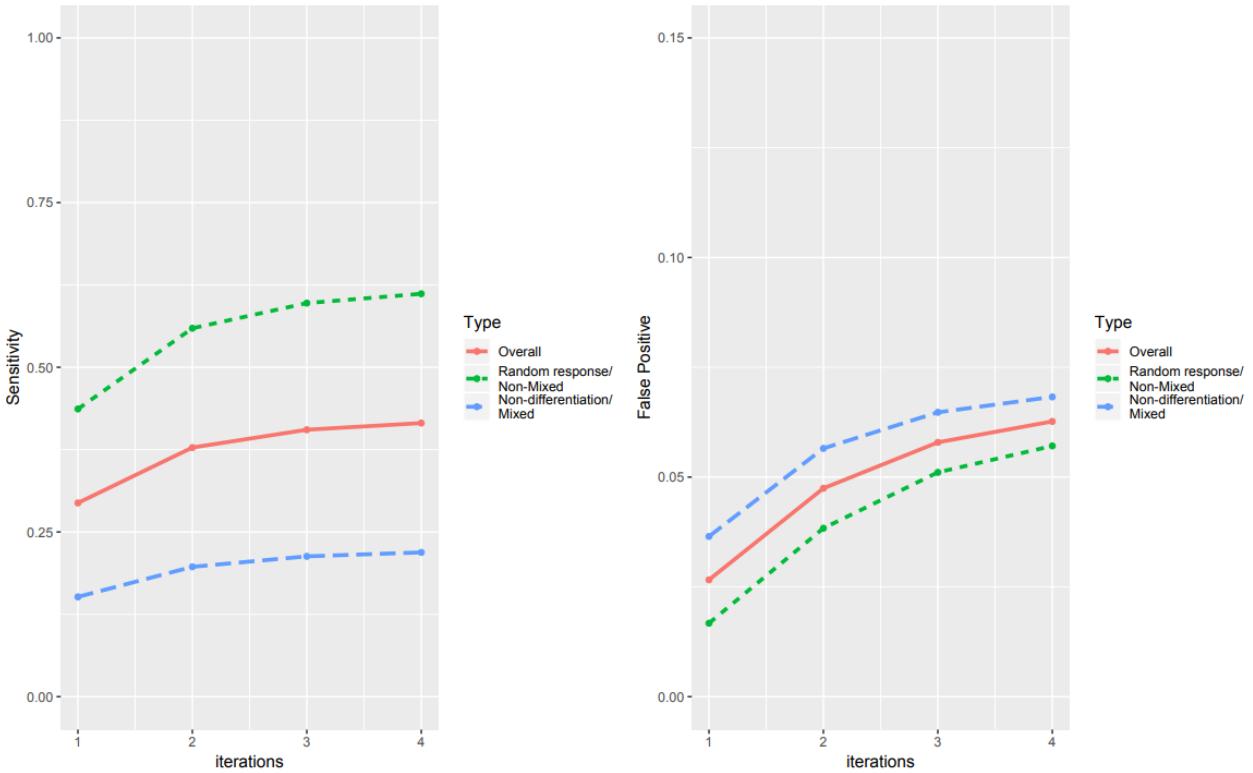

Figure 2.10. Sensitivity (2.10A, left) and false positive rate (2.10B, right) by autoencoder iterations.

Figures 2.11 shows sensitivity based on different numbers of iterations by number of items in the scale (six vs. 12), careless response types (random response and non-differentiation of item direction changes), and percentages of CR (10%, 20%, and 30%). Similar to the Figure 2.10 results, the autoencoder works better in random response, especially for the 12-item scales. The autoencoder also works better when the percentage of CR is smaller, where the 10% condition yields the highest sensitivity. As in Figure 2.10, the largest increase in sensitivity across the iterations is from iteration 1 to iteration 2, suggesting that two iterations of the autoencoder is a better than more iterations.

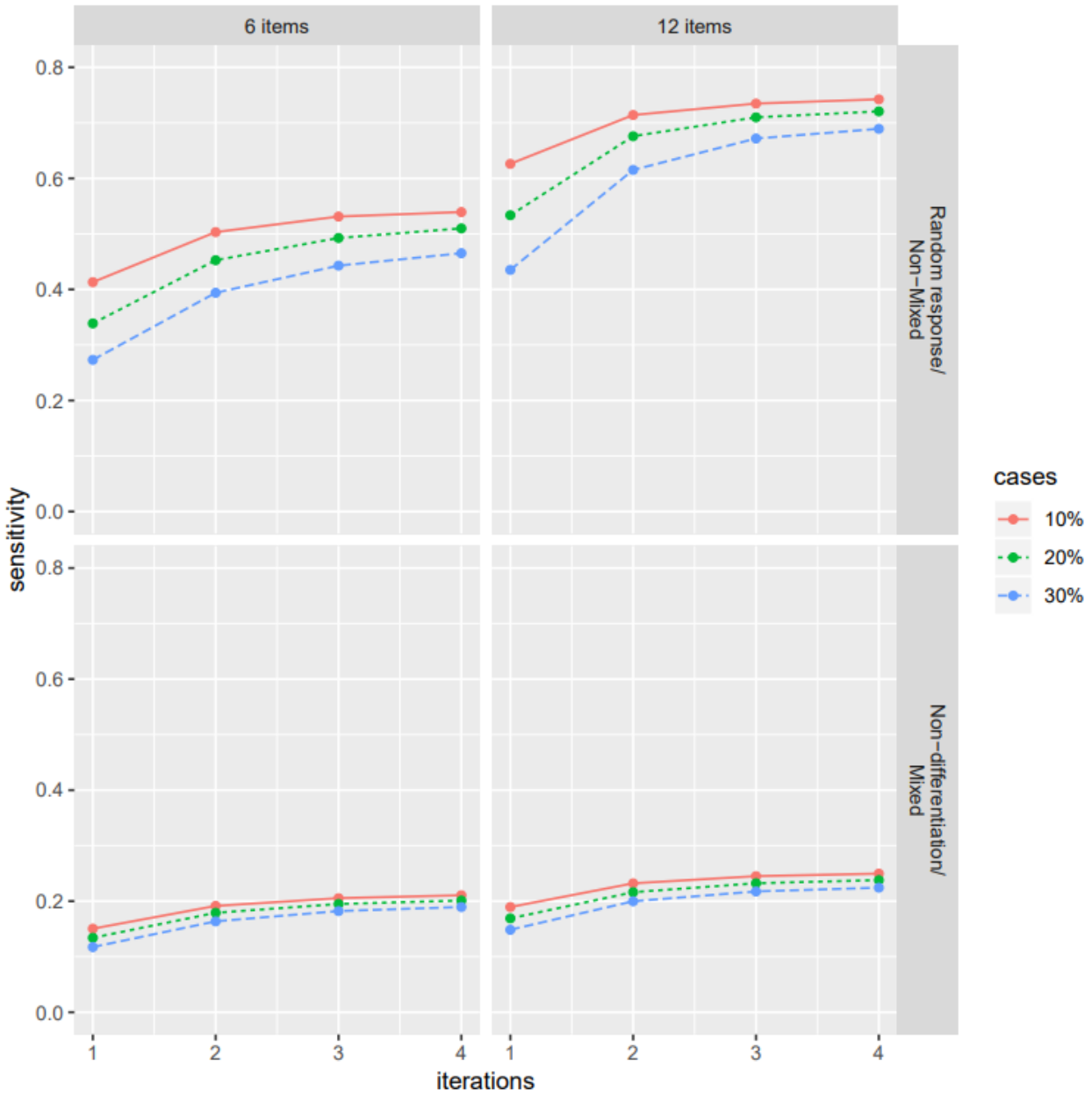

Figure 2.11. Sensitivity by autoencoder iterations for percentage of CR (10%: red curve; 20%: green curve; and 30%: blue curve), scale length (six or 12 items), and careless response type (random or non-differentiation).

Figure 2.12 shows the false positive rates for the autoencoder by number of iterations, number of items in the scale (6 vs. 12), careless response types (random and non-differentiation of item direction), and percentage of CR (10%, 20%, and 30%).The autoencoder works better in the random response condition, with smaller false positive rates. The 12 item condition has

slightly lower false positive rates than the 6-item condition. The false positive rates are the highest for the 10% CR condition, followed by 20% and 30% CR rates.

Across these findings, the increase in sensitivity often has the tradeoff of increasing the false positive rate. As in Figure 2.10, in the random response condition, the false positives rates go beyond 0.05 for 10% and 20% CR rates after the second iteration, suggesting again that two iterations of the autoencoder is for random responses. For non-differentiation of item direction, the false positive rates go beyond 0.05 for almost all iterations except the first one. However, for iterations 2 to 4 the second iteration shows the closest values to 0.05.

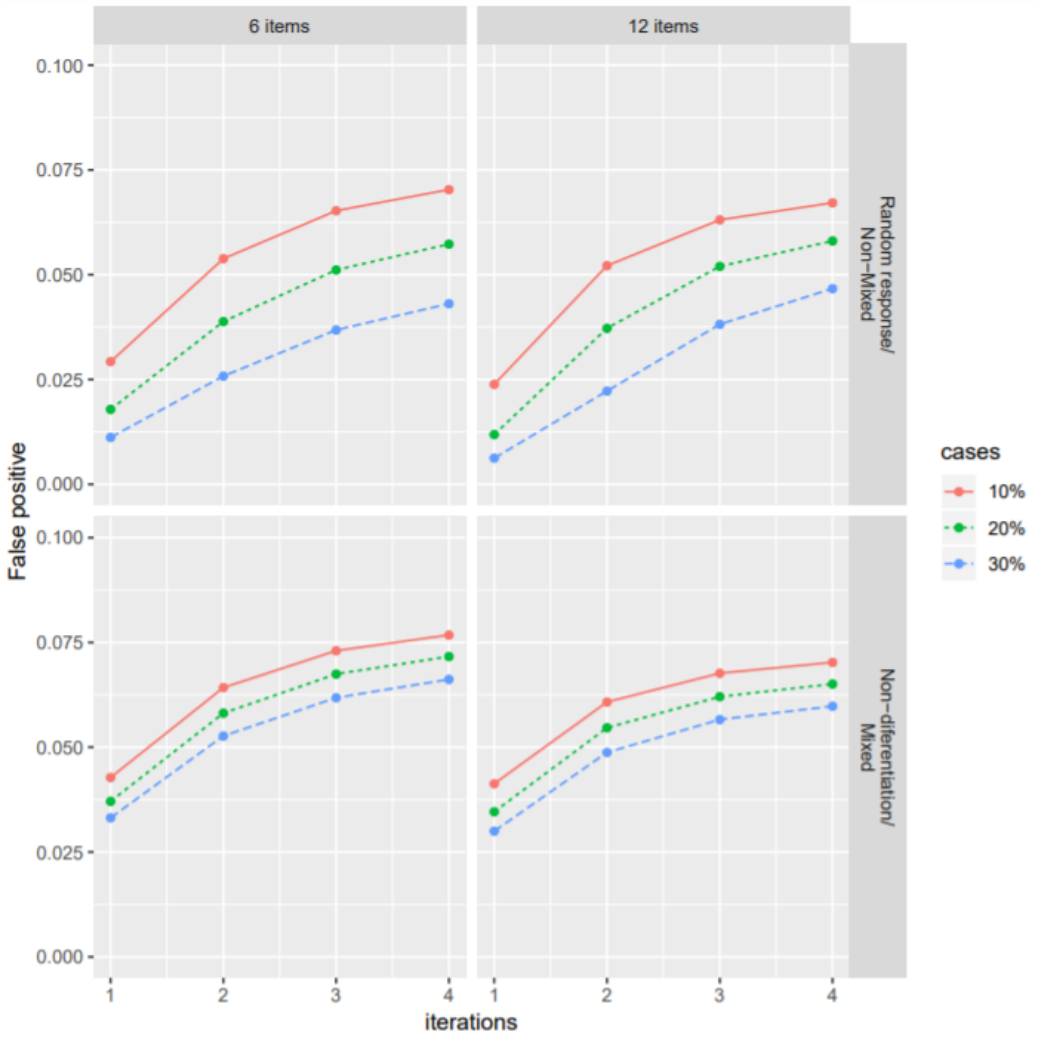

Figure 2.12. False positive rate by autoencoder iteration for percentage of CR (10%: red curve; 20%: green curve; and 30%: blue curve), six vs. 12 items, and random vs. nondifferentiated contamination type.

Figures 2.13 to 2.15 show the total accuracy rates by autoencoder iteration for percentage of CR for random response CR behaviors. In most of the conditions, the total accuracy rates remain similar across iterations, especially when the sample includes 10% or 20% CR. When the sample includes 30% CR, there is an increase of total accuracy rate from iteration 1 to iteration 2, and the rate remains stable for iterations 3 and 4. In most conditions, the total accuracy rate is the highest when 10% CR are present in the sample and the lowest when 30% CR are present.

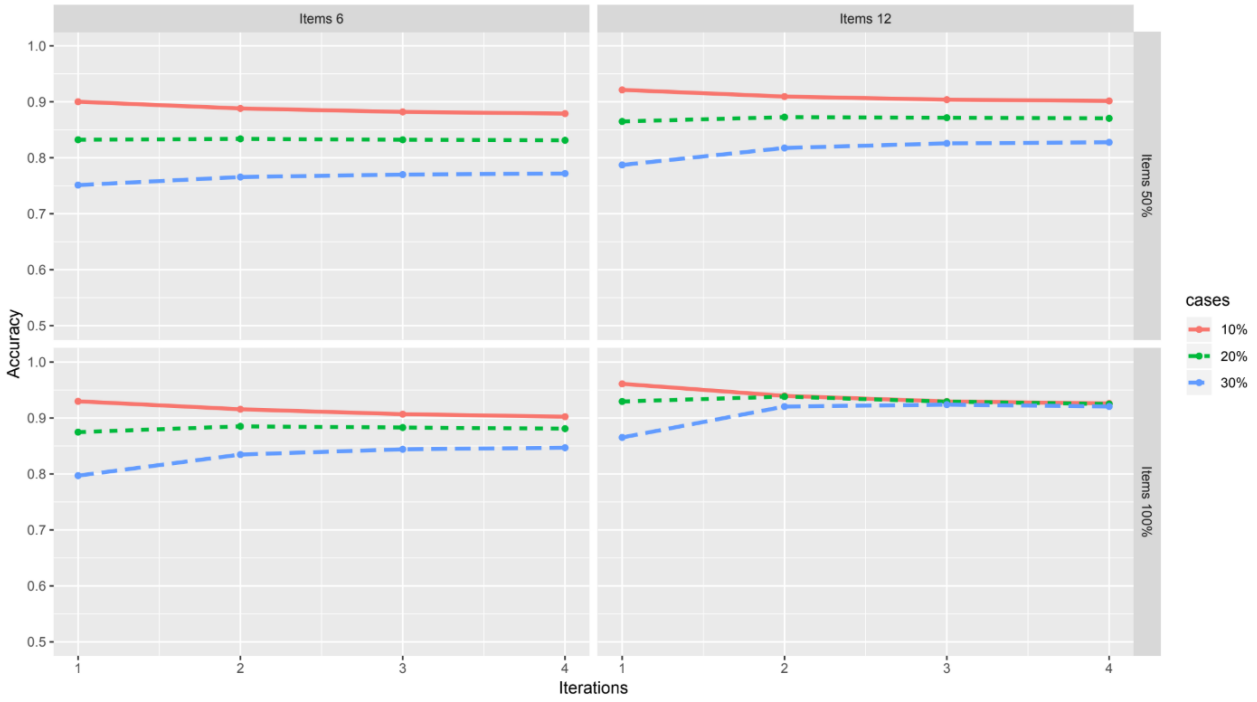

Figure 2.13. Total accuracy rate by autoencoder iteration for percentage of CR (10%: red curve; 20%: green curve; and 30%: blue curve), random response contamination, on scale length six items with half items (i.e., three) having careless response, and length 12 items with half items (i.e., six) having careless responses and with all items having careless responses.

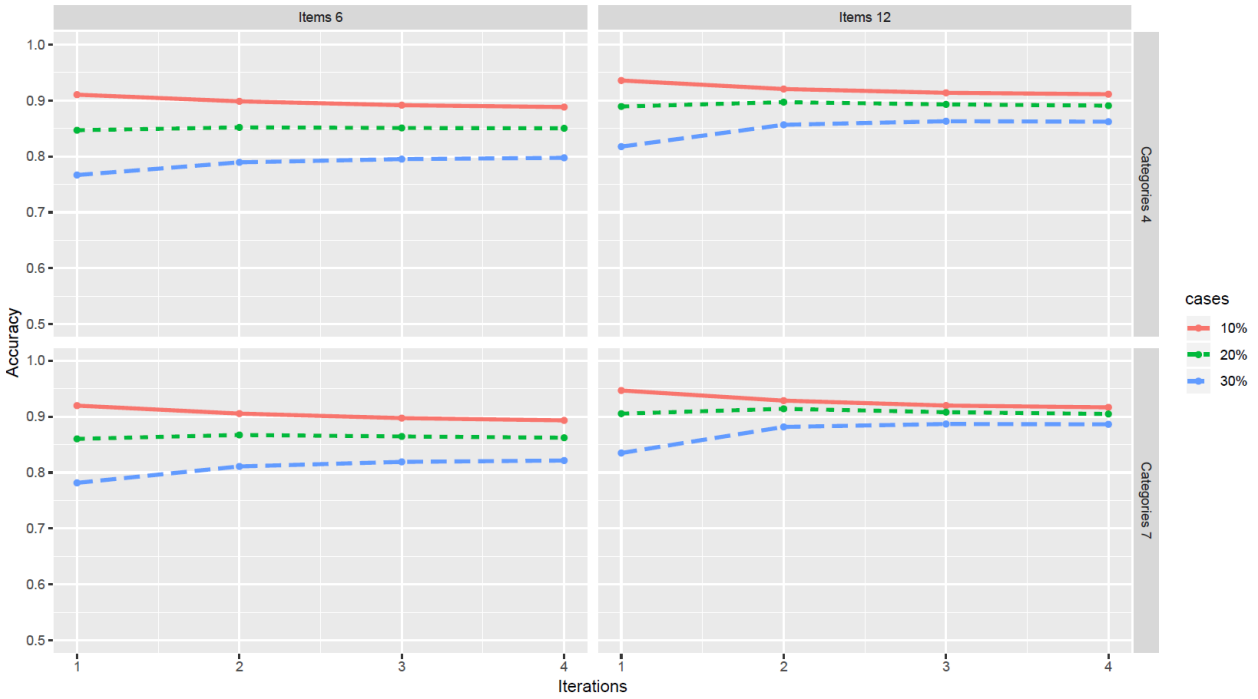

Figure 2.14. Total accuracy rate by autoencoder iteration for percentage of CR (10%: red curve; 20%: green curve; and 30%: blue curve), scale length of six or 12 items, with four vs. seven response categories.

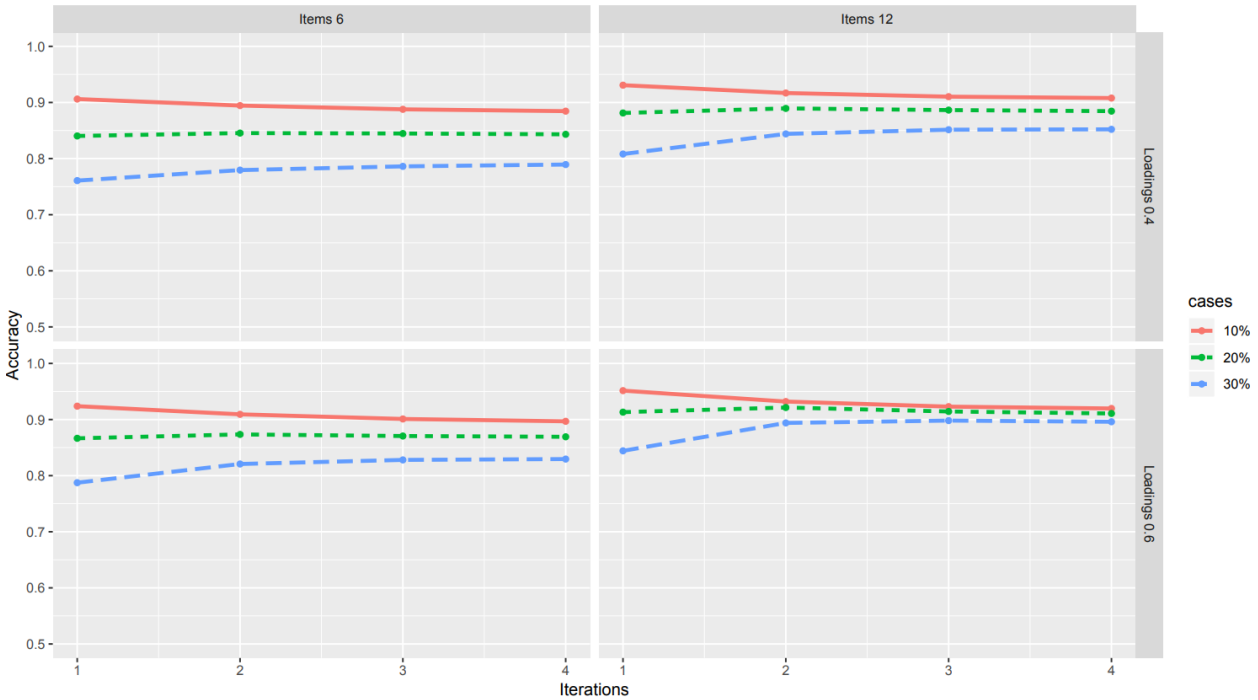

Figure 2.15. Total accuracy rate by autoencoder iteration for percentage of CR with random response contamination (10%: red curve; 20%: green curve; and 30%: blue curve), scale length of six and 12 items with low (0.4) and high (0.6) factor loadings

Figures 2.16 to 2.18 show the total accuracy rates by autoencoder iteration and percentage of CR for non-differentiation CR behaviors. In almost all the conditions, the total accuracy rates remain similar across iterations. Similar to the results of random response CR behaviors, the total accuracy rate is the highest for 10% CR in the sample and the lowest for 30% CR in the sample. Compared to the total accuracy for random response CR behaviors, the total accuracy is slightly lower for non-differentiation CR behaviors.

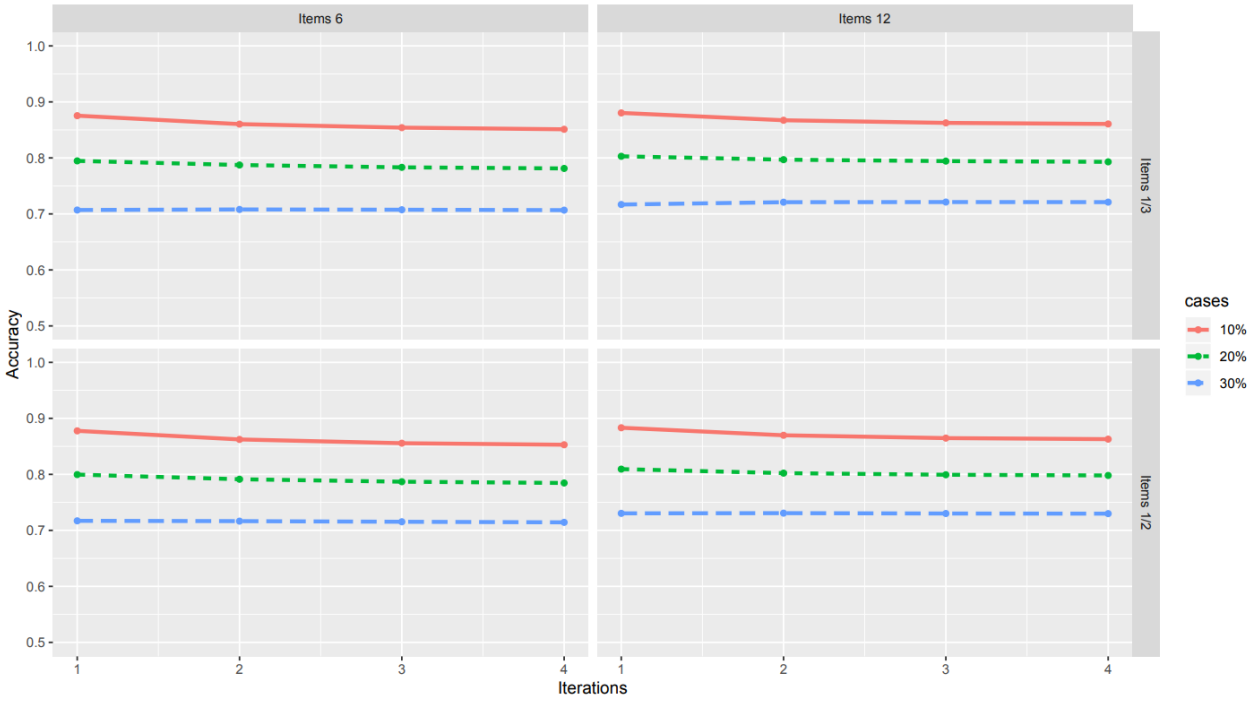

Figure 2.16. Total accuracy rate by autoencoder iteration for percentage of CR (10%: red curve; 20%: green curve; and 30%: blue curve) with non-differentiation careless response contamination by scale length of six or 12 items, and 1/3 or 1/2 contaminated items.

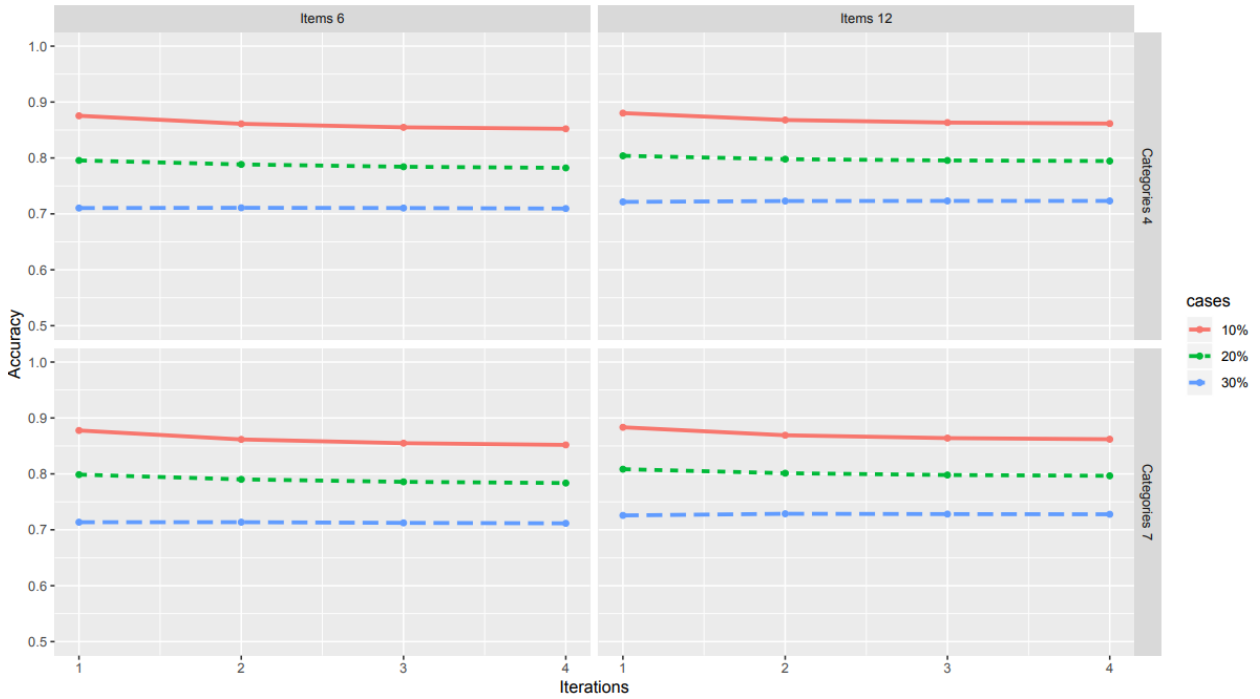

Figure 2.17. Total accuracy rate by autoencoder iteration for percentage of CR (10%: red curve; 20%: green curve; and 30%: blue curve) for non-differentiation careless response contamination by length of six or 12 items, with four or seven response categories.

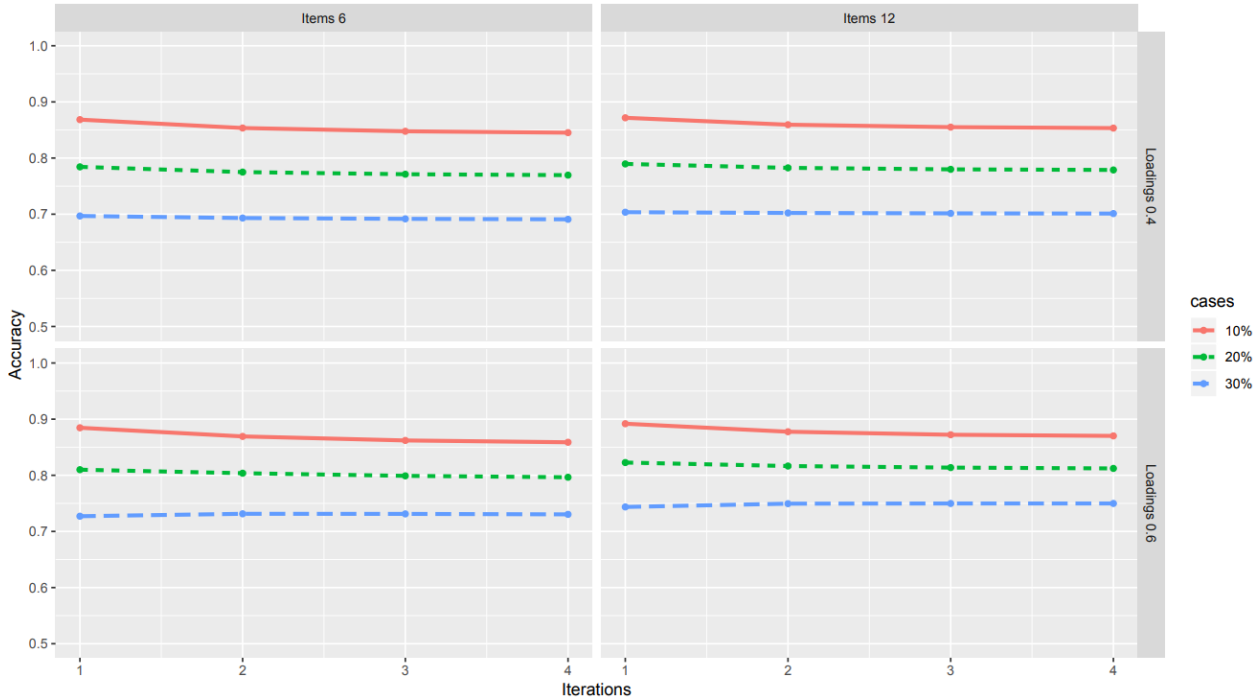

Figure 2.18. Total accuracy rate by autoencoder iteration for percentage of CR (10%: red curve; 20%: green curve; and 30%: blue curve), with non-differentiation careless response contamination by scale length of six or 12 items and factor loadings 0.4 or 0.6.

Based on these results, a two-iteration approach was chosen for the autoencoder method. The two-iteration approach has an obvious increase of sensitivity compared to a single iteration while maintaining a false positive rate below or near 0.05. Using three or more iterations has unacceptably large false positive rates. The total accuracy for two iterations is better than one iteration and similar but lower than that for three and four iterations, especially when the sample has 30% CR employing random response behaviors. For the other conditions, the total accuracy measure does not seem to favor any particular number of the iterations. Given the increased computation time and complexity when adding iterations, the two iteration approach seemed to be the most effective and efficient choice.

2.5.2. The Autoencoder and the Standardized Log-Likelihood l_z^p

Figure 2.19 shows a comparison of sensitivity level for random response behaviors across all experimental conditions for the two-iteration autoencoder method versus the standardized log-likelihood l_z^p method. The autoencoder clearly outperforms the standardized log-likelihood l_z^p, with higher sensitivity in all conditions. When the sample contains 10% CR, the sensitivity of both the autoencoder and the standardized log-likelihood l_z^p is higher than when the sample contains 20% CR, which is higher than when the sample contains 30% CR. In other words, the sensitivity decreases for both methods when the percentage of CR increases in the sample. Both methods work better in scales with 12 items, likely due to more information in the 12 item scales. Other factors related to higher sensitivity include higher factor loadings (0.6 versus 0.4), a higher percentage of contaminated items (100% versus 50%), and more response categories (seven versus four). In the condition for 12-item scales with seven response categories, a

moderate factor loading (i.e., 0.6) and all items contaminated, both methods work better with higher sensitivity than any other conditions.

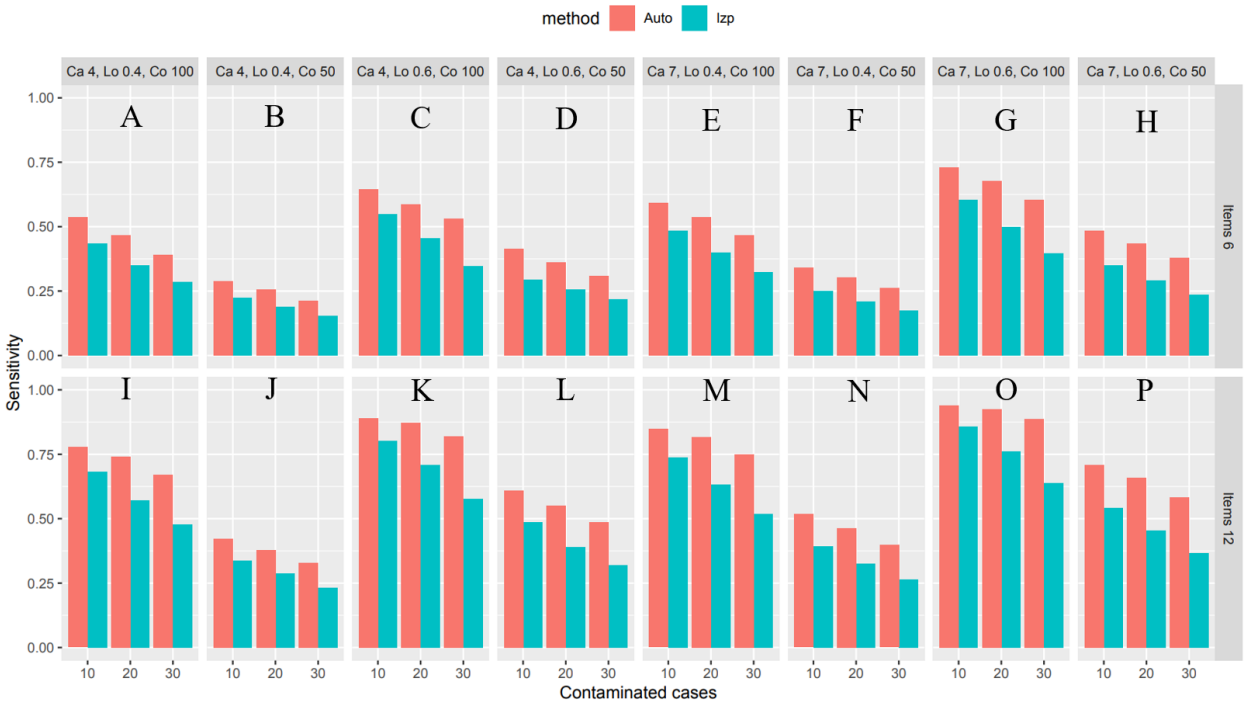

Figure 2.19. Sensitivity values by percentage of CR for random response behaviors by scale length, number of response categories, factor loadings, and percentage of contaminated items.

Figure 2.20 shows the false positive rate comparisons of the two methods for random response behaviors across all experimental conditions. Given the higher sensitivity of the autoencoder method (shown in Figure 2.19), it is surprising that the autoencoder also has higher false positive rates in all conditions compared to the standardized log-likelihood l_z^p. Despite higher false positive rates, most of the rates for the autoencoder are below or near 0.05, a cutoff criterion for CR identification in previous literature (Emons, 2008; Conijn, 2013). The false positive rates decrease as the percentage of CR in the sample increases.

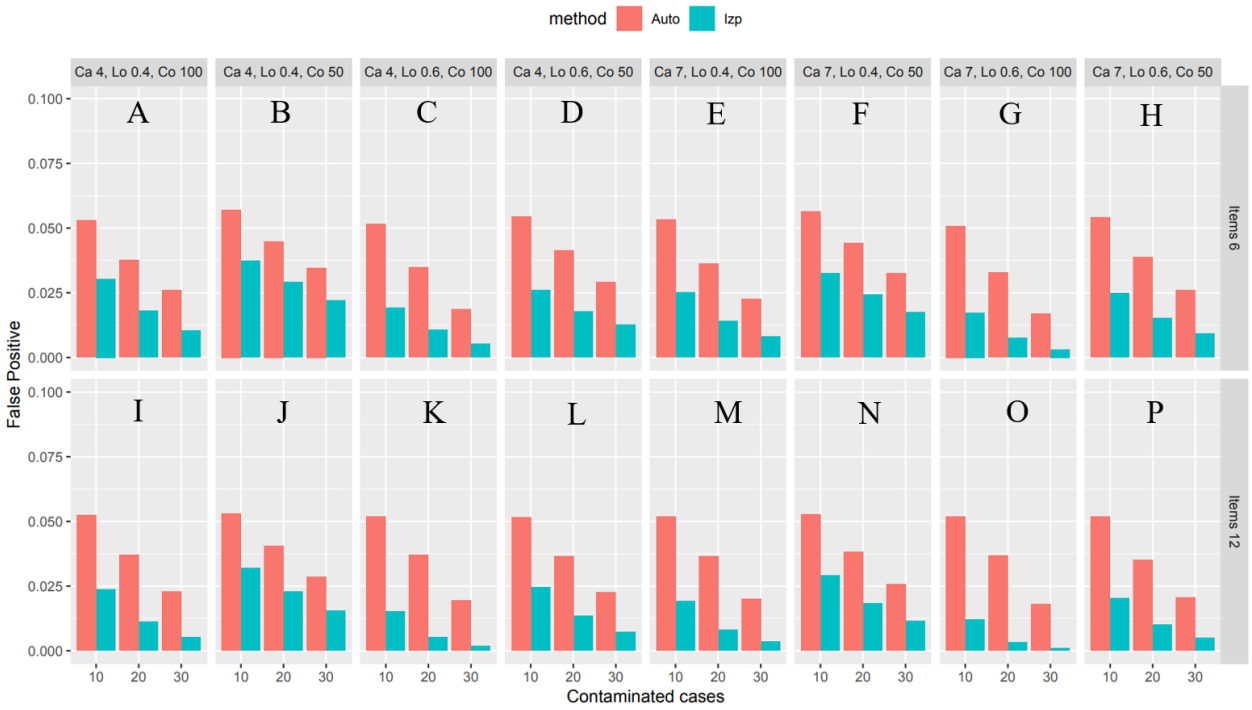

Figure 2.20. False positive rates by percentage of CR for random response behaviors by scale length, number of response categories, factor loadings, and percentage of contaminated items.

Figure 2.21 compares the autoencoder and the standardized log-likelihood l_z^p for non-differentiation careless response behaviors across all experimental conditions. Similar as the results for random response CR behaviors (Figure 2.19), the autoencoder clearly outperforms the standardized log-likelihood l_z^p with higher sensitivity in all conditions. Compared to random response behaviors, both methods have lower sensitivity values in identifying non-differentiation behaviors. When the sample contains 10% CR, the sensitivity of both the autoencoder and the standardized log-likelihood l_z^p is higher than when the sample contains 20% CR, and higher than when the sample contains 30% CR. In other words, the sensitivity decreases for both methods when the percentages of CR increase in the sample. Consistent with results for random response behaviors, both methods work better in scales with 12 items, higher factor loading, higher percentage of contaminated items, and more response categories.

Figure 2.21. Sensitivity values by percentage of CR for non-differentiation behaviors by scale length, number of response categories, factor loadings, and percentage of contaminated items.

Figure 2.22 shows the false positive rates of the two methods for non-differentiation behaviors across all experimental conditions. The false positive rates for non-differentiation CR behaviors are higher than those for random response behaviors, indicating that both methods work better in identifying random response than non-differentiation behaviors. The autoencoder has higher false positive rates in all conditions compared to the standardized log-likelihood l_z^p.

Figure 2.22. False positive rates by percentage of CR, for non-differentiation CR behaviors by scale length, number of response categories, factor loadings, and percentage of contaminated items.

2.6 Discussion

The two-iteration autoencoder works better in terms of increased sensitivity and acceptable false positive rates than the standardized log-likelihood l_z^p, with higher sensitivity across all conditions. Despite higher two-iteration autoencoder false positive rates compared to the standardized log-likelihood l_z^p, in most conditions the false positive rates for the autoencoder are below or near 0.05, the acceptable level as suggested in previous literature (Emons, 2008; Conijn, 2013). In total, combining all results across different conditions, the autoencoder works better in identifying CR than the standardized log-likelihood l_z^p.

Several important factors influence the performance of the two methods. Both methods work better in situations where CR employ random response strategies, rather than not paying attention to item direction change (from positive to negative items, or non-differentiation

behavior). This is likely associated with how the non-differentiation responses were generated. A subset of items in the scales were randomly selected and then reverse coded to generate contaminated non-differentiation responses. In a scale with seven response categories, if a respondent chooses the midpoint option (i.e., 4), his/her response remains the same after recoding. In other words, non-differentiation behavior does not contaminate data for those who select mid-response categories. As a result, the actual percentage of contaminated cases is lower than the percentages of cases who received recoding (10%, 20%, and 30%). In a similar fashion, the recoding of responses closer to the mid-point response (e.g., response option 3 and 5 in a seven-category scale) result in less inconsistency than the recoding of extreme responses (e.g., response option 1 and 7). Since both methods aim to detect data inconsistencies, non-differentiation with extreme responses is easier to detect than non-extreme responses. Future study could evaluate how the two methods perform when non-differentiation occurs in non-midpoint responses or in extreme responses only.

Sensitivity is the highest for both methods when the data has 10% CR, followed by 20% CR and 30% CR. This is likely due to the larger proportion of systematic patterns in the data with fewer contaminated cases. For a similar reason, scales with 12 items have higher sensitivity than scales with six items. Scales with better measurement properties (e.g., higher factor loadings) result in higher sensitivity. This is because in scales with higher factor loadings, the items are more highly correlated leading to higher consistency or less variability in the items. Both methods are then better able to identify inconsistent cases.

In summary, for both types of careless response behaviors, the autoencoder results in the highest sensitivity for 12-item scales with seven response categories, a moderate factor loading (i.e., 0.6), and fewer contaminated cases in the data (e.g., 10%). There is a tradeoff between

sensitivity and false positive rates. Conditions with higher sensitivity (identifying more "true" CR) also tend to have higher false positive rates (identify more "true" non-CR as CR). Fortunately, in most conditions, the false positive rates for both methods are at an acceptable ($\leq$ 0.05) level.

This is the first study that introduces and evaluates the use of both the standardized log-likelihood l_z^p and the autoencoder to identify satisficing behavior in survey research. It bridges theory and methods from psychometrics, neural networks, and survey methodology. Current survey literature uses quality indicators such as response time and straightlining behaviors to identify satisficing in multi-item questions with grid question formats. These methods, despite their wide uses, have several important drawbacks. Response time is influenced by respondent cognitive functioning and question difficulty. It is difficult to find a cutoff point in response time to identify careless response, and the cutoff points in the real world may differ across population groups (e.g., older respondents have a higher cutoff on response time for the same question than younger). Methods to detect straightlining do not work well in scales where the items are all in the same direction (all positively worded).

This chapter focused on the identification of two satisficing behaviors, random responses and non-differentiation of item direction. However, other types of satisficing behavior appear in multi-item scales, such as response styles and response order effects. Future study can further examine the use of the autoencoder to identify other types of careless responses.

This work is based on a simulation study, which may not reflect real world situations. Chapter 4 is based on web survey data from an online survey panel. There an application of the standardized log-likelihood l_z^p and the autoencoder to real-world data is examined.

Future studies should examine other person-fit statistics, especially those based on non-parametric methods. Since results here suggest that the two methods compared do not work well in situations where respondents do not distinguish scale direction changes in scales with both positive and negative item wordings, future study could evaluate other methods that work better for this type of careless responses, or to examine how the two methods can be improved in other contexts.

2.7 References

Allaire, J.J., & Chollet, F. (2018). *keras: R interface to 'keras'*. R package version 2.2.0.

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

Bethlehem, J. & Biffignandi, S. (2012). *Handbook of Web Surveys*. Hoboken, New Jersey: Wiley.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27*, 281-291.

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*, 573-591.

Chalmers, R.P. (2012). *mirt*: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29.

Conijn, J.M. (2013). *Detecting and Explaining Person Misfit in Non-cognitive Measurement*. Universiteit van Tilburg.

Conijn, J.M., Emons, W.H.M., & Sijtsma, K. (2014). Statistic lz-based person-fit methods for non-cognitive multiscale measures. *Applied Psychological Measurement, 38*, 122-136.

De la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*, 159-177.

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.

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

Felt, J.M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Frontiers in Psychology, 8*, 1-9.

Ferrando, P.J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences, 52*, 718-722.

Glas, C.A.W., & Meijer, R.R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27*, 217-233.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, Massachusetts: The MIT Press.

Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement, 76*, 912-l932.

Hagan, M.T., Demuth, H.B., Beale, M.H., & De Jesús, O. (2014). *Neural Network Design* [PDF file]. Retrieved on January 15, 2019, from http://hagan.okstate.edu/NNDesign.pdf.

Hauser, D.J., & Schwarz, N. (2015). It's a Trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open*, 1-6.

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

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

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

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

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

Krosnick, J. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567.

Krosnick, J., & Alwin, D.F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*, 201-219.

LaHuis, D.M., & Copeland, D. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods, 12*, 296-319.

Levine, M.V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

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

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

Meijer, R.R. (2002). Outlier detection in High-Stakes certification testing. *Journal of Educational Measurement, 39*, 219-233.

Meijer, R.R., Molenaar, I.W., & Sijtsma, K. (1994). Influence of test and person characteristics on non-parametric appropriateness measurement. *Applied Psychological Measurement, 18*, 111-120.

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

Nering, M.L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121-129.

Nunes, I., Hernane, D., Andrade, R., Bartocci, L.H., & dos Reis, S.F. (2017). *Artificial Neural Networks*. Sao Paulo, Brazil: Springer.

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

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reise, S. P., & Flannery, W.P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education, 9*, 9-26.

Rizopoulos, D. (2018). *Package 'ltm'* [PDF file]. Retrieved on April 8, 2018 from https://cran.r-project.org/web/packages/ltm/ltm.pdf.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*, 1-36.

Samejima, F. (2016). Graded response models. In *Handbook of Item Response Theory* (pp. 123-136). Boca Raton, FL: Chapman and Hall/CRC.

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

Tendeiro, J.N. & Meijer, R.R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement, 51*, 239-259.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*, 1-27.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge, UK: Cambridge University Press.

Turner, G., Sturgis, P., & Martin, D. (2014). Can response latencies be used to detect survey satisficing on cognitively demanding questions? *Journal of Survey Statistics and Methodology*, *3*, 89-108.

Van Der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*, 267–298.

Widhiarso, W., & Sumintono, B. (2016). Examining response aberrance as a cause of outliers in statistical analysis. *Personality and Individual Differences, 98*, 11-15.

Wise, S.L., & DeMars, C.E. (2005). Low examinee effort in Low-Stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

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

Woods, C.M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*, 189-194.

Woods, C.M., Oltmanns, T. F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment*, *20*, 159-168.

Zhang, C., & Conrad, F. G. (2013). Speeding in Web Surveys : The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*, 127-135.

Zickar, M.J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

Appendix 2.1

Average distributions of the categories for the first item across all the conditions with four and seven categories.

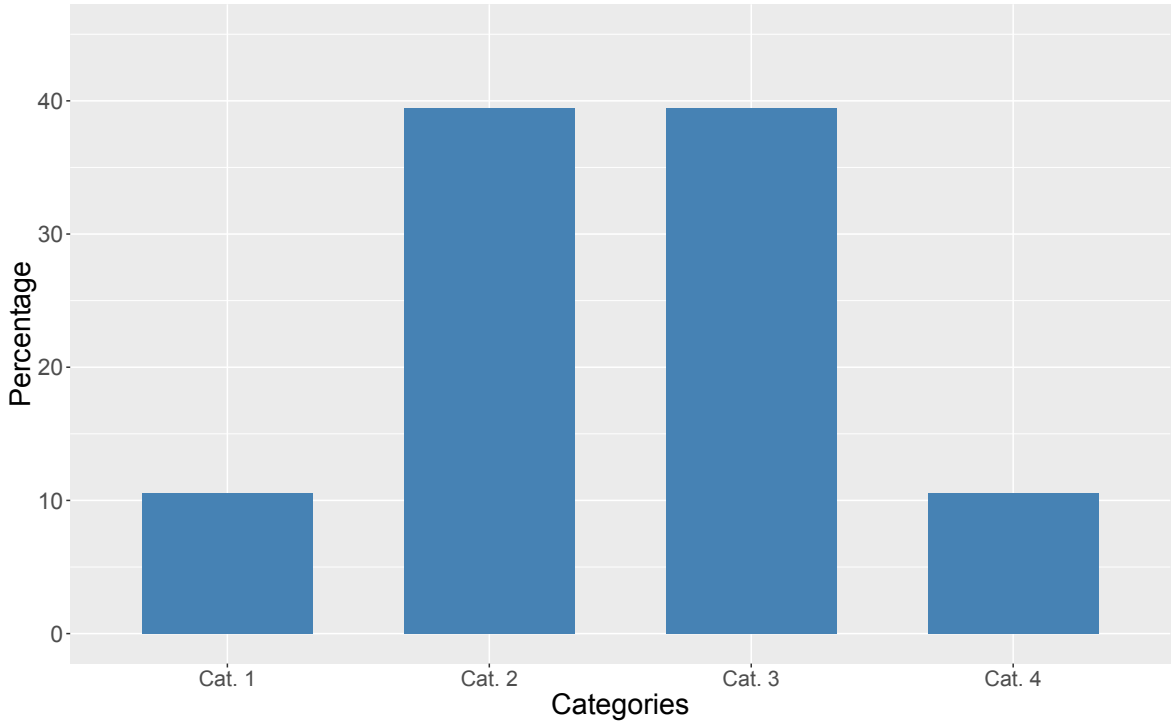

Figure 2.23. Average distributions of the categories for the first item across all the conditions with four categories.

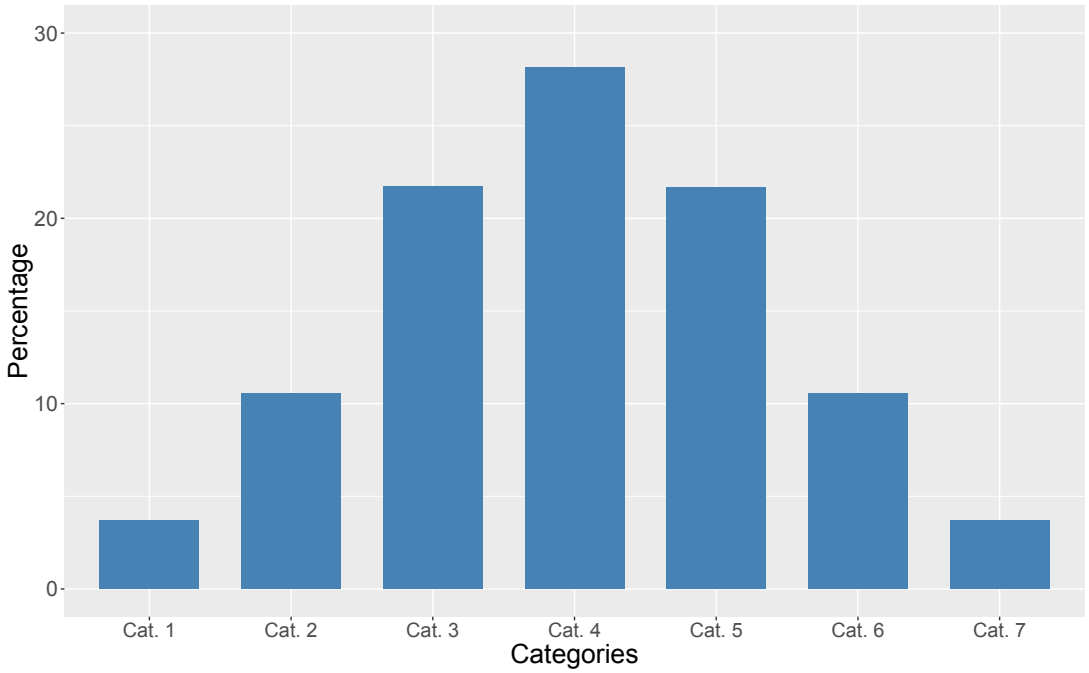

Figure 2.24. Average distributions of the categories for the first item across all the conditions with seven categories.

CHAPTER III

Imputing Careless Respondent Data: A Comparison Between the Standardized Log-Likelihood Person-Fit Statistic l_z^p and the Autoencoder

3.1 Introduction

Multi-item scales, often appearing in grid question formats, are commonly used in surveys to assess a variety of constructs including respondents' attitudes, behaviors, health (e.g., mental health scale), wellbeing, and personality. Previous literature studying satisficing behavior related to multi-item scales mainly focus on quality indicators like straightlining and speeding (Schonlau & Toepoel, 2015; Zhang & Conrad, 2013). Two other methods that can be used to identify satisficers or careless respondents (CR) have received little to no attention in previous survey literature – 1) the person-fit-statistic widely discussed in psychometric literature and 2) the autoencoder method recently developed in the engineering field. In Chapter 2, the standardized log-likelihood l_z^p and autoencoder approaches are compared by assessing capacity to identify CR in a simulation study. Findings showed that the autoencoder works better in identifying CR in scales with a small number of items, as well as in a number of other conditions. This chapter centers around the question of how to best deal with data provided by CR in analysis.

Once individuals have been identified as CR, researchers need to decide what to do with the responses of those subjects. The literature proposes two main approaches. The first is "complete data analysis" (Anduiza & Galais, 2017) where the researcher analyzes responses of

all subjects, regardless of whether the subject is classified as a CR or a non-CR. The assumption underlying this alternative is that the responses affected by careless responding will not affect the results importantly. The second approach is "casewise deletion" where all CR data are excluded from analysis (Hauser & Schwarz, 2015; Oppenheimer, Meyvis, & Davidenko, 2009). This approach is advocated when the researcher observes results obtained using the entire sample are different from those results that exclude subjects classified as CR.

A limitation of the first approach (complete data analysis) is that, unless the researcher validates the "no-effect" assumption, conclusions could be biased. Researchers who want to adopt this approach are obligated to evaluate this assumption before conducting any analysis in a form of sensitivity analysis. However, it appears this sensitivity analysis is rarely done.

Casewise deletion has three main limitations. First, it reduces the sample size, with a consequent reduction in power to detect effects in hypothesis testing. Second, given that a questionnaire may have multiple scales, the subjects classified as CR for one scale may differ on classification for another scale. CR exclusion becomes a complex decision when CR could overlap across all scales, in only one particular scale, or in several scales. Analyses using different scales, if different exclusion rules are used, will include different sample sizes. Third, the approach of casewise deletion assumes that a respondent having careless response behavior for one scale will have similar behavior in all other scales. It is possible that respondents may have careless responses for some but not all scales. Krosnick (1991) suggested that respondents may report truthfully and in a diligent way at the beginning of the survey and, due to increased fatigue, may satisfice at the end of the survey. Excluding all CR for any scale in the dataset will lose the good responses, particularly those at the beginning of the survey.

A third approach (Little, & Rubin, 2002) from the survey literature has not been considered: delete careless responses, multiply imputing the deleted responses and then apply multiple imputation combining rules to obtain estimates and test statistic values from the imputed dataset. The potential advantages of the "delete and multiply impute" approach over "complete data analysis" and "casewise deletion" are that 1) it may increase statistical precision compared to "casewise deletion" approach and 2) it can reduce the bias compared to "complete data analysis".

This chapter compares different approaches to the treatment of CR for those with random response behavior. The "delete and multiply impute" approach is compared to the two commonly used approaches. Comparisons of the three different approaches examine properties of estimates of a key parameter in a Confirmatory Factor Analysis (CFA) model – the correlation between two latent variables. Following Chapter 2, the standardized log-likelihood l_z^p and the autoencoder are employed to detect CR in simulated data, and the relative bias and relative root mean square error of the CFA correlation estimate are computed under "delete and multiply impute," "complete data analysis," and "casewise deletion" approaches. The research hypothesis is that "casewise deletion" and the "delete and multiply impute" approaches outperform the "complete data analysis" approach, and the "delete and multiply impute" approach enables researchers to retain the original sample size and more statistical information.

3.2 Method

3.2.1. Design Characteristics

For the simulation study, a two-factor CFA model with categorical ordered manifest variables was used to generate data for this simulation. Figure 3.1 presents an example of a two-

factor CFA model with 12 items, six associated with each latent factor, and a hypothesized correlation between factors. Each manifest variable j has $m_j + 1$ possible levels, with $m_j = 3$ or $m_j = 6$ for $j = 1,2, \ldots, 12$.

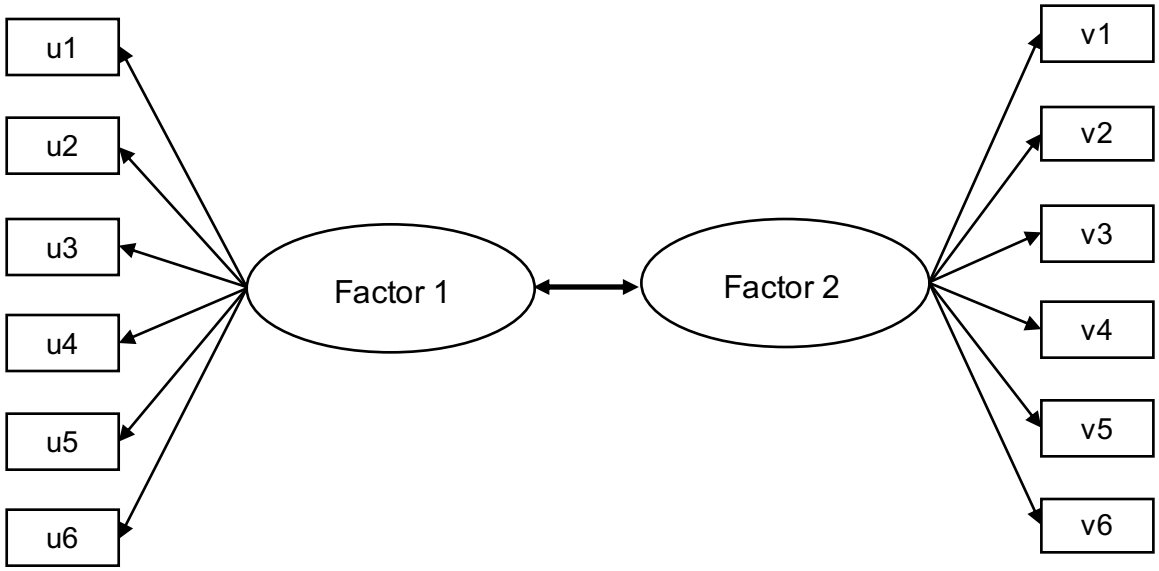

Figure 3.1. Example of a two factor CFA model with 12 items.

The threshold parameter values were chosen to obtain an item distribution that matched the skewness and kurtosis of a standard normal distribution. Table 3.1 shows the threshold values used in this study for the four and seven category alternatives.

Table 3.1. Threshold parameter values for four and seven categories.

	Categories	
Threshold	*Four*	*Seven*
1	-1.25	-1.79
2	0.00	-1.07
3	1.25	-0.36
4		0.36
5		1.07
6		1.79

A factorial design crossed each of six study factors of interest: 1) scale lengths (six and 12 items); 2) percentage of CR in the sample (10%, 20%, and 30%); 3) percentage of careless responses across items (one-half and all items); 4) factor correlation (0.30 and 0.50); 5) factor loadings (0.4 and 0.6); and 6) number of response categories per item (four and seven). In total, there were 96 conditions. For each condition, 1000 datasets were generated.

Scale lengths. One purpose of this study is to evaluate how well the standardized log-likelihood l_z^p and the autoencoder perform in a scale with a small number of items. Six and 12 items were chosen as the two levels for scale length for each scale. Previous literature claims these are a small number of items for adequate detection (Emons, 2008; Conijn et al., 2014).

Percentage of CR in the sample. The three percentages of CR in the sample (10%, 20%, and 30%) were chosen to assess a range from low where the impact on data quality is likely to be minimal (10%) to high where more than 30% is unlikely in practice. This range of percentages is also consistent with literature evaluating satisficing behaviors where the percentage of respondents failing a trap question in surveys falls in such a range (Curran et al., 2010; Johnson, 2005; Meade & Craig, 2012).

Percentage of careless responses. Similar to Conijn (2013) and Emons (2018), CR employing careless responding behavior are applying it to all the items (e.g., 12 responses from 12 items), or only half of the items (e.g., 6 responses from 12).

Factor correlations. Correlation levels between the latent variables were $r = .30$ and $r = .50$, corresponding to moderate and high correlation based on Cohen's (1988) conventions to interpret effect size.

CFA factor model loadings. Previous studies (Meijer, Molenaar, & Sijtsma, 1994; Meijer & Sijtsma, 2001) showed that the power of person fit statistics like the standardized log-

likelihood l_z^p depends on the discrimination power of the items. Higher discrimination (i.e., higher factor loadings) are associated with higher detection rates. Two loading levels, low (0.40) and medium (0.60) discrimination were considered. Higher loadings (e.g., 0.80) are very unlikely to appear in practice consistently across all items.

Number of response categories per item. For the number of categories per item, four and seven categories were chosen mainly because these are the lower and upper limits for most commonly used scales in surveys.

For each of these factors, and combinations of factors in the 96 experimental conditions, it is difficult to know which of the two detection methods and three data handling techniques for each will have lower variance and biased estimates. We expect that estimated correlations from the "delete and multiply impute" technique will most often have lower variance and bias compared to the other two techniques, regardless of the detection method employed.

3.2.2 Data Generation Mechanism

Data for 1000 replications for each of the 96 conditions described above were generated based on the following procedure:

- A dataset of polytomous item-score vectors (i.e., 12 or 24 items in total across two scales) was generated based on a two correlated-factor CFA model with categorical ordered indicators. Each latent factor is measured through one multi-item scale (as in Figure 3.1). This data-generation process followed three steps. First, latent factor scores were generated for each respondent based on a standard normal distribution for each factor. Second, continuous latent item responses were generated using latent factor scores and the factor loadings (e.g., 0.4 or 0.6). Finally, these continuous latent item responses were

recoded to ordinal categories based on the threshold parameters presented in Table 3.1. This data-generation process was implemented using the *lavaan* package in R. The sample size for each dataset was 1,000 respondents.

- A simple random sample without replacement of 10%, 20%, or 30% from each dataset was chosen. The responses to the second scale were replaced using simulated careless response patterns generated by randomly selecting category responses with uniform distribution probabilities (i.e., P_{x_j}= 1/4 and P_{x_j}= 1/7 for each category level, respectively). The dataset that included the randomly generated careless responses is denoted as the *manipulated dataset*. It is important to note that only one of the two scales were modified by careless responses.

3.2.3 Identifying CR based on the Standardized Log-Likelihood l_z^p.

The following procedure was applied to each of the manipulated datasets:

- The graded response model (GRM) (see Chapter 1) was applied to the data from the second scale to estimate item parameter values. For each item *j* with m_j + 1 categories, a discrimination α_j and set of category boundary locations δ_{jk} (k = 1, 2, …, m_j) were estimated (see Equation 1, Chapter 2) using the *mirt* R package.
- Using the estimated item parameters, the latent score θ_i and its standard error $SE(\theta_i)$ for each respondent in the manipulated dataset was computed.
- For each respondent in the manipulated dataset, the $l_{z,i}^p$ was computed using their responses and the respondents parameters θ_i and $SE(\theta_i)$.

Following De la Torre and Deng (2008) and Rizopoulos (2018), a parametric bootstrap method was used to obtain the *p*-value that was used to classify each $l_{z,i}^{p}$ as CR or non-CR for 1,000 replications for each respondent in the manipulated dataset:

- A new latent score estimate, $\theta_{new,i}$, was generated from a normal distribution with mean θ_i, and standard deviation equal to SE(θ_i) obtained above.
- A new response pattern of polytomous item-score vectors based on the GRM using the item parameters, α_j, δ_{jk}, and the latent score $\theta_{new,i}$ was generated.
- For the new response pattern, and using $\theta_{new,i}$, the standardized log-likelihood $l_{znew,i}^{p}$ was computed (see Chapter 2 for a description of this procedure).

For each respondent i in the manipulated dataset, a *p*-value for the standardized log-likelihood $l_{z,i}^{p}$ score was computed as the proportion of the number of the 1000 $l_{znew,i}^{p}$ values which were no larger than $l_{z,i}^{p}$:

$$p_i = \frac{1 + \sum_{b=1}^{1000} I\left(l_{z_{new},i,b}^{p} \leq l_{z,i}^{p}\right)}{1 + 1000}$$

where $I(\cdot)$ denotes the indicator function equal to 1 if $(\cdot)$ is true and zero otherwise. Following Tendeiro, Meijer, and Niessen (2016), respondents in the manipulated dataset for which $p_i < 0.05$ were classified as CR.

3.2.4 Identifying CR based on the Autoencoder

The autoencoder method was described more in detail in Chapter 2. As shown in Chapter 2, the ideal number of iterations to apply the autoencoder method in terms of sensitivity and the false positive rate was two. In this chapter, we used the two-iteration autoencoder to identify CR.

3.2.5 Estimation of CFA Parameter Models

The CFA parameter estimates were computed then under the following five conditions:

- "Complete data analysis": CFA parameters were estimated using all respondents included in the manipulated dataset, regardless of CR status.
- "Casewise deletion" with the standardized log-likelihood l_z^p identification method: CFA parameters were estimated after excluding all CR identified by the standardized log-likelihood l_z^p.
- "Casewise deletion" with autoencoder identification method: CFA parameters were estimated after excluding all CR identified by the autoencoder.
- "Delete and multiply impute" with the standardized log-likelihood l_z^p identification method: CFA parameters were estimated after deleting and multiply imputing CR responses to the second scale for those identified as CR by the standardized log-likelihood l_z^p.
- "Delete and multiply impute" with the autoencoder identification method: CFA parameters were estimated after deleting and multiply imputing CR responses to the second scale for those identified by the autoencoder.

3.2.6 Multiple Imputation Method

The deleted responses from the standardized log-likelihood l_z^p and the autoencoder were multiply imputed using a predicted mean matching sequential regression multivariate imputation (Raghunathan, Berglund, & Solenberg, 2018). The imputation model included all the responses from the first and second scales. The number of imputed datasets was 20. Multiply imputed CFA parameter estimates were obtained using Rubin's combining rules.

Multiple imputation (MI) is a method used to handle missing data. As discussed more in detail in Rezvan, Lee and Simpson (2015) and Little and Rubin (2002), MI proceeds with replicating the incomplete dataset multiple times and replacing the missing data in each replicate with plausible values drawn from an imputation model. The statistical analysis of interest is then performed on each completed dataset separately. Finally, a single MI estimate (and its standard error) is calculated by combining the estimates (and standard errors) obtained from each completed dataset using "Rubin's rules" (Little & Rubin, 2002; Rezvan, Lee, &, Simpson, 2015).

Comparing to single imputed methods, the advantage of MI is that it takes into account the uncertainty associated with the imputed values. The estimated variance of the overall MI estimate allows for within-imputation (i.e. the uncertainty in the estimate within each completed dataset) and between-imputation (i.e. the uncertainty between the estimates across the completed datasets) variability (Little & Rubin, 2002; Rezvan, Lee, &, Simpson, 2015).

3.2.7 Performance Measures

Two empirical measures, the relative bias and the empirical relative root mean square error were used to compare the different methods. Let ρ_m denote the general estimate of the true

correlation $\rho_{1,2}$ for method *m* between the two latent variables, the empirical relative bias based on *K* simulated datasets is:

$$\text{RelBias}(\rho_m) = \frac{\text{Bias}(\rho_m)}{\rho_{1,2}} = \frac{\text{E}(\rho_m) - \rho_{1,2}}{\rho_{1,2}} = \frac{\sum_{i=1}^{K} \frac{\rho_{m,i}}{K} - \rho_{1,2}}{\rho_{1,2}}$$

The empirical relative root mean square error based on *K* simulated datasets is:

$$\text{RelRMSE}(\rho_m) = \frac{\text{RMSE}(\rho_m)}{\rho_{1,2}} = \frac{\sqrt{\sum_{i=1}^{K} \frac{(\rho_{m,i} - \rho_{1,2})^2}{K}}}{\rho_{1,2}}$$

3.3 Results

Figure 3.2 presents the relative bias for the 10% CR condition with different treatment of CR ("complete cases", "casewise deletion", and "delete and multiply impute") where CR were identified using the standardized log-likelihood l_z^p or the autoencoder identification methods. In an ideal situation where no bias is present, the height of the bars in Figure 3.2 would be close to zero. As shown in Figure 3.2, almost all relative biases are negative (the latent variable correlation is under-estimated), as would be expected when real data are disrupted by random response.

Scale length (i.e., six or 12 items) appears to be an important factor for the treatment effects. Figure 3.2 shows that for six-item scale conditions, the casewise deletion and the delete and multiply impute approaches are not superior to the complete data analysis using the full sample, except for scales with seven response categories, better psychometric properties (0.6

factor loading), and having all items contaminated. On the other hand, for the 12-item scales, in most conditions, the casewise deletion and the delete and multiply impute approaches reduce relative bias compared to the complete data analysis approach.

Another factor that influences the treatment effects is the percent of contaminated items, that is, whether a CR employs satisficing behaviors to half or all of the items. The casewise deletion and the delete and multiply impute approaches result in more or similar relative bias than the complete data analysis approach where only half of the items are contaminated. When all the items are contaminated in the 12-item condition, both the casewise deletion and the delete and multiply impute approaches reduce relative bias more than the complete data analysis approach. Psychometric properties also influence the treatment effects. The relative biases are smaller for both the casewise deletion and the delete and multiply impute approaches in scales with higher factor loadings (0.6) than weaker (0.4).

Comparing the CR identification methods, for six-item scales, the delete and multiply impute CR data with CR detected using the autoencoder identification produces more or similar relative bias for estimated correlations compared to the standardized log-likelihood l_z^p identification. For 12-item scales and all items contaminated, the delete and multiply impute for autoencoder identification yields similar or smaller relative bias than when the standardized log-likelihood l_z^p is used to identify CR. The most promising treatment effects of both the casewise deletion and the delete and multiply impute approaches occur in the condition whit 12-item scales with seven response categories, better factor loading (0.6), and all items contaminated (Figure 3.2O). In this condition, the autoencoder performs better than the standardized log-likelihood l_z^p, with almost zero relative bias.

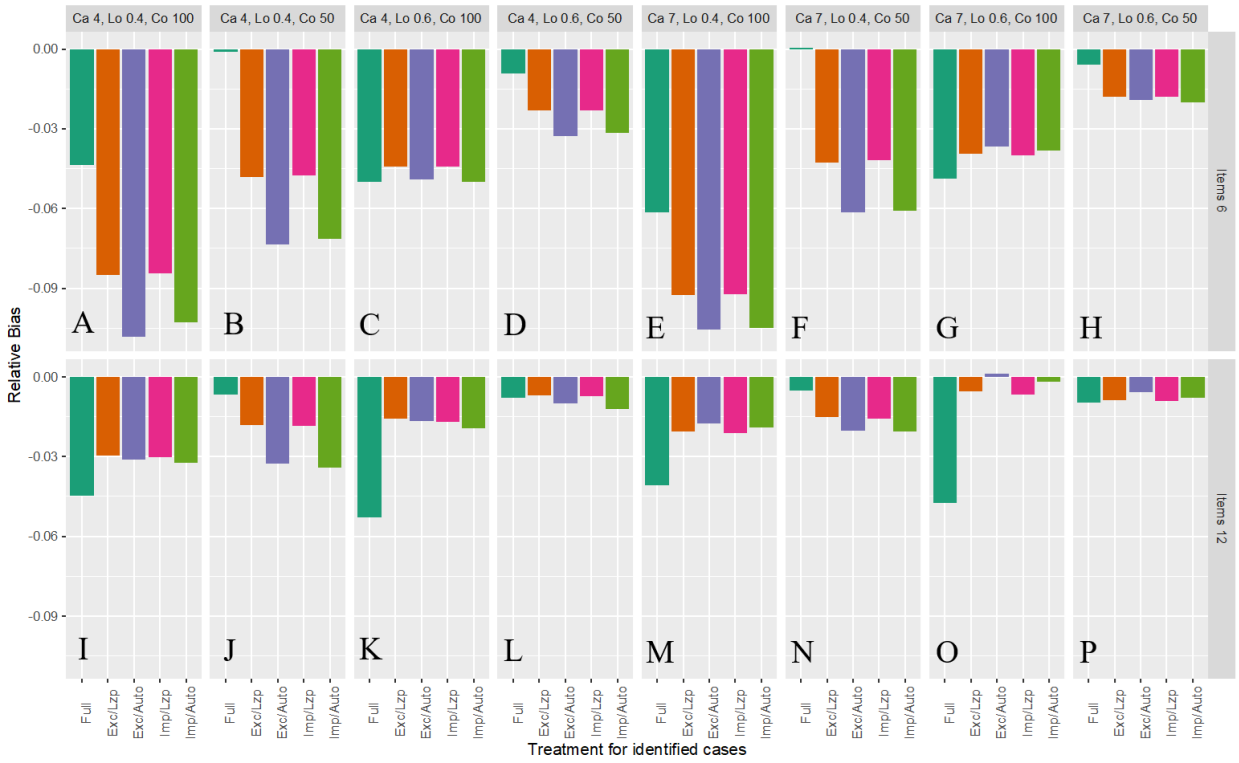

Figure 3.2. Relative bias for the 10% CR condition with different treatment of CR ("complete case" denoted as "*Full*"; "casewise deletion for l_z^p or the autoencoder", denoted as "*Exc/Lzp*" or "*Exc/Auto*"; and "delete and multiply impute for l_z^p or the autoencoder", denoted as "*Imp/Lzp*" or "*Imp/Auto*") by scale length (six or 12 items), number of response categories (denoted as *Ca* 4 or 7), factor loadings (denoted as *Lo* 0.4 or 0.6), and percent of contaminated items (denoted as *Co* 50 or 100).

Figures 3.3 and 3.4 present the relative bias for the 20% and 30% CR condition, respectively. Similar patterns of relative bias across conditions were found in Figures 3.3 and 3.4 as those found in Figure 3.2. Not surprisingly, as the percentage of CR increases in the dataset, the relative bias increases. This is true not only for the complete case or full sample analysis for both detection methods but also for the other four approaches. For 12-item conditions with all items contaminated (Charts I, K, M, and O in Figures 3.3 and 3.4), the differences between the autoencoder and the standardized log-likelihood l_z^p approaches appear to be larger when the data include 20% or 30% CR. This suggests that when more CR employ satisficing behaviors in all

the items in the 12-item scales, the autoencoder has an advantage in reducing relative bias. This is true for both the casewise deletion and the delete and multiply impute approaches.

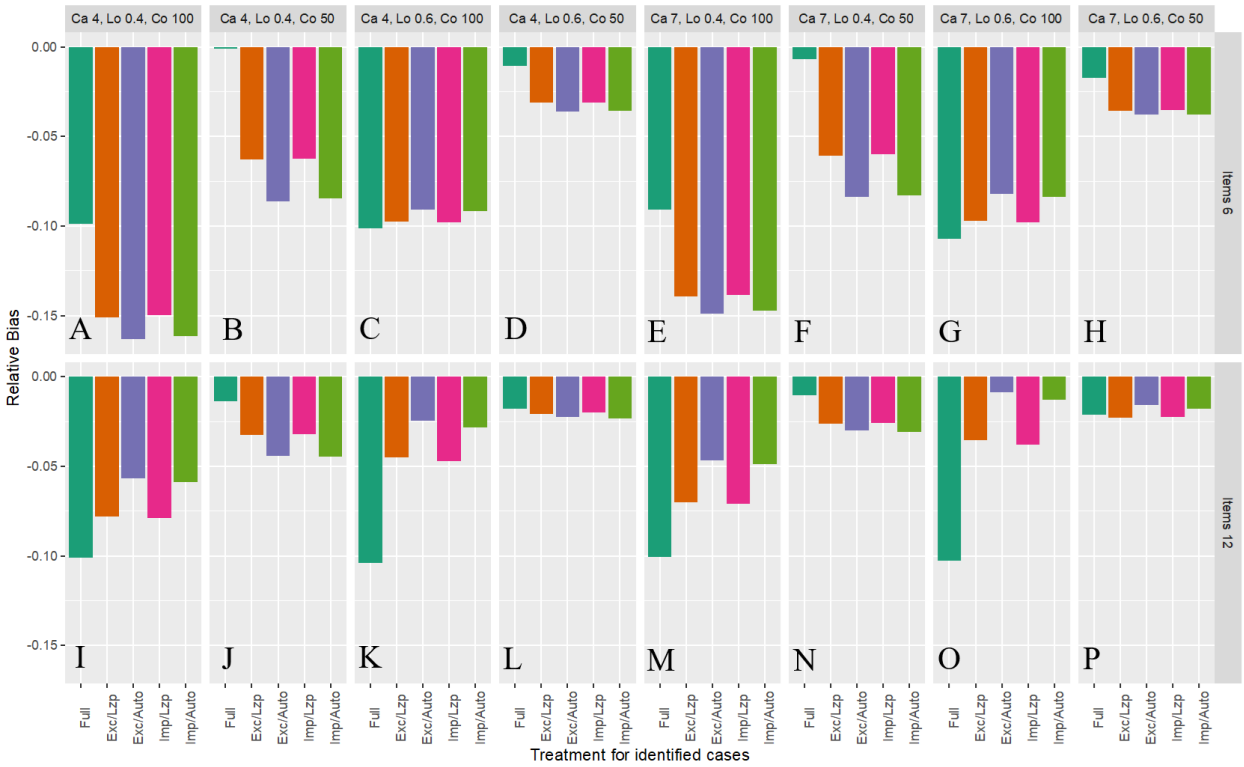

Figure 3.3. Relative bias for the 20% CR condition with different treatment of CR ("complete case" denoted as "*Full*"; "casewise deletion for l_z^p or the autoencoder", denoted as "*Exc/Lzp*" or "*Exc/Auto*"; and "delete and multiply impute for l_z^p or the autoencoder", denoted as "*Imp/Lzp*" or "*Imp/Auto*") by scale length (six or 12 items), number of response categories (denoted as *Ca* 4 or 7), factor loadings (denoted as *Lo* 0.4 or 0.6), and percent of contaminated items (denoted as *Co* 50 or 100).

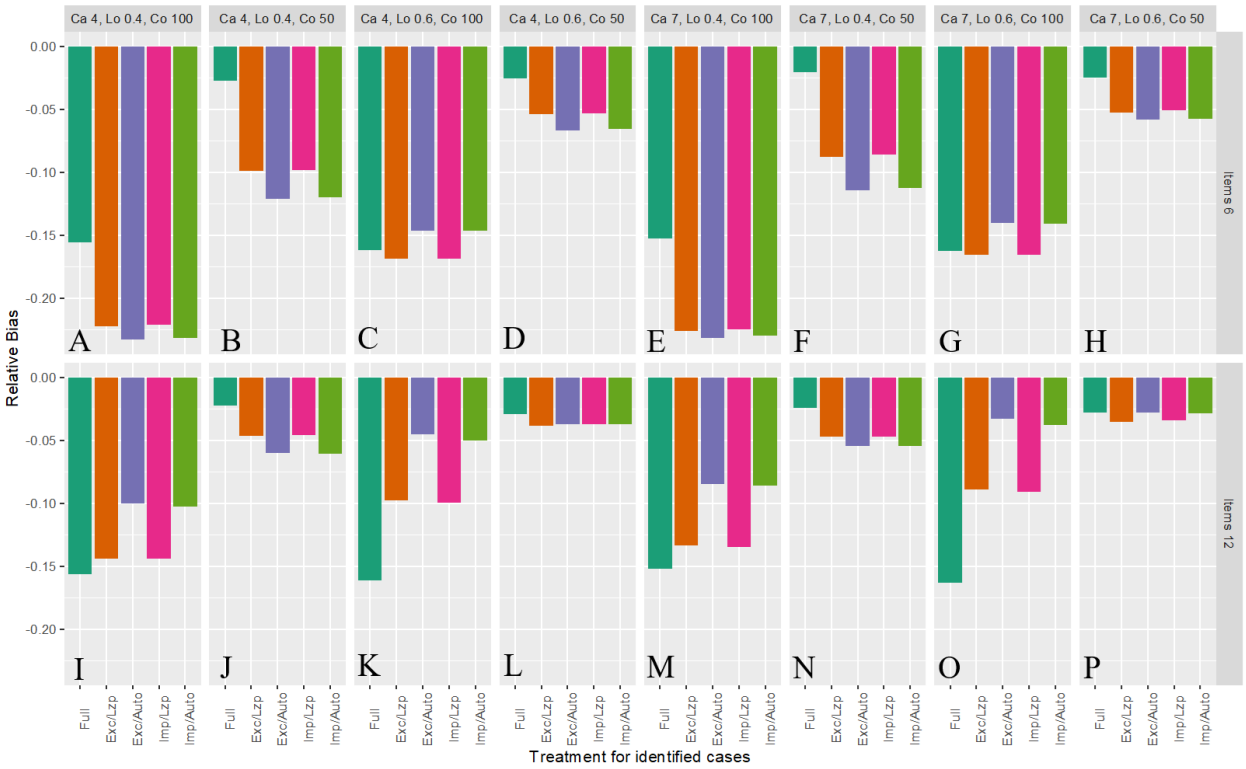

Figure 3.4. Relative bias for the 30% CR condition with different treatment of CR ("complete case" denoted as "*Full*"; "casewise deletion for l_z^p or the autoencoder", denoted as "*Exc/Lzp*" or "*Exc/Auto*"; and "delete and multiply impute for l_z^p or the autoencoder", denoted as "*Imp/Lzp*" or "*Imp/Auto*") by scale length (six or 12 items), number of response categories (denoted as *Ca* 4 or 7), factor loadings (denoted as *Lo* 0.4 or 0.6), and percent of contaminated items (denoted as *Co* 50 or 100).

Figures 3.5 to 3.7 present the relative root mean square error (RMSE) for the conditions 10%, 20% and 30% CR, respectively. Lower values of relative RMSE indicate higher quality estimates of the correlation coefficients. Similar to Figures 3.2 to 3.4, five different treatment methods of CR data are compared. Similar to the results for the relative bias, the casewise deletion and the delete and multiply impute approaches perform better in the 12-item scales, especially when more CR are included in the dataset. In the 12-item conditions, the casewise deletion and the delete and multiply impute approaches using the autoencoder or the standardized log-likelihood l_z^p yield similar results whether there are 10% or 20% of CR are in the data. When the data include 30% CR, the two identification methods have a larger difference.

In Figure 3.6, when all the items are contaminated, the casewise deletion and the delete and multiply impute approaches with both CR-identification methods reduce RMSE compared to the complete case data analysis using the full sample. In addition, the autoencoder clearly outperforms the standardized log-likelihood l_z^p in these conditions. For example, in Figure 3.6O, both casewise deletion and the delete and multiply impute treatment methods using the standardized log-likelihood l_z^p reduced the relative RMSE, compared to the complete data analysis, by about 40%, while the methods using the autoencoder reduced the relative RMSE by about 60%.

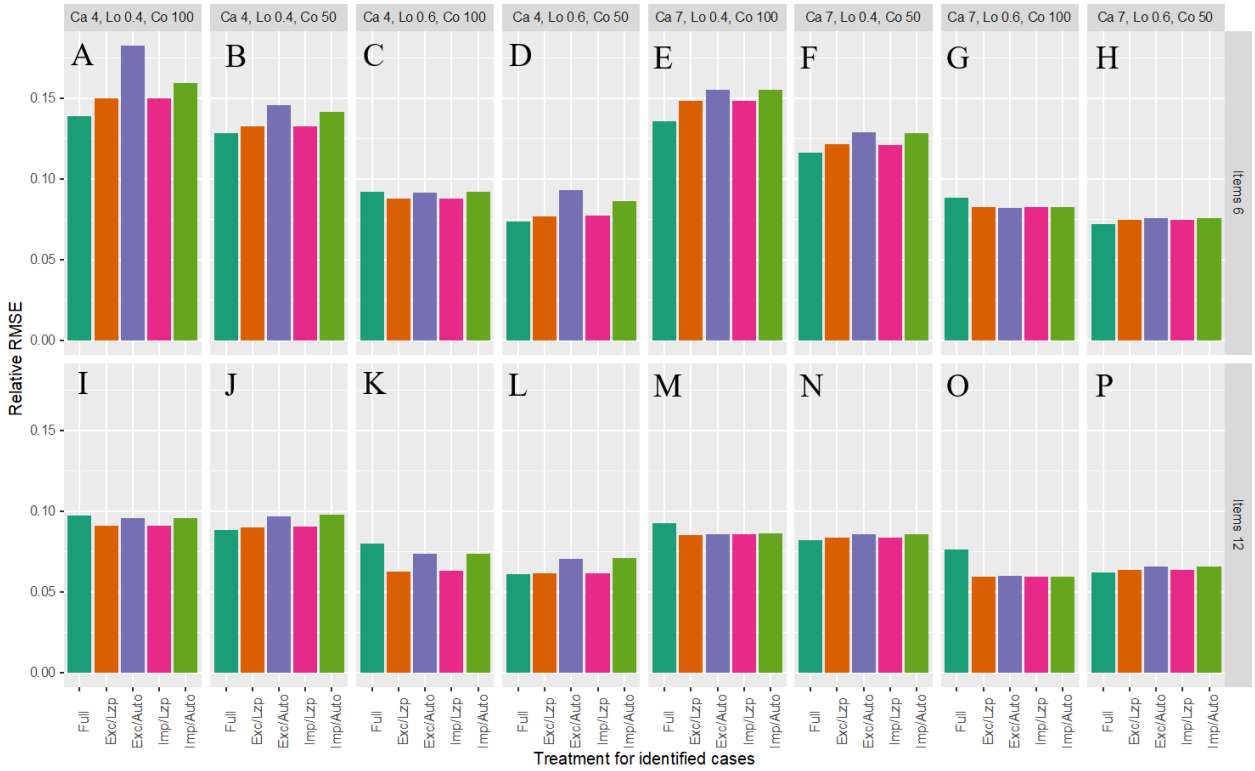

Figure 3.5. Relative root mean square error (RMSE) for the 10% CR condition with different treatment of CR ("complete case" denoted as "*Full*"; "casewise deletion for l_z^p or the autoencoder", denoted as "*Exc/Lzp*" or "*Exc/Auto*"; and "delete and multiply impute for l_z^p or the autoencoder", denoted as "*Imp/Lzp*" or "*Imp/Auto*") by scale length (six or 12 items), number of response categories (denoted as *Ca* 4 or 7), factor loadings (denoted as *Lo* 0.4 or 0.6), and percent of contaminated items (denoted as *Co* 50 or 100).

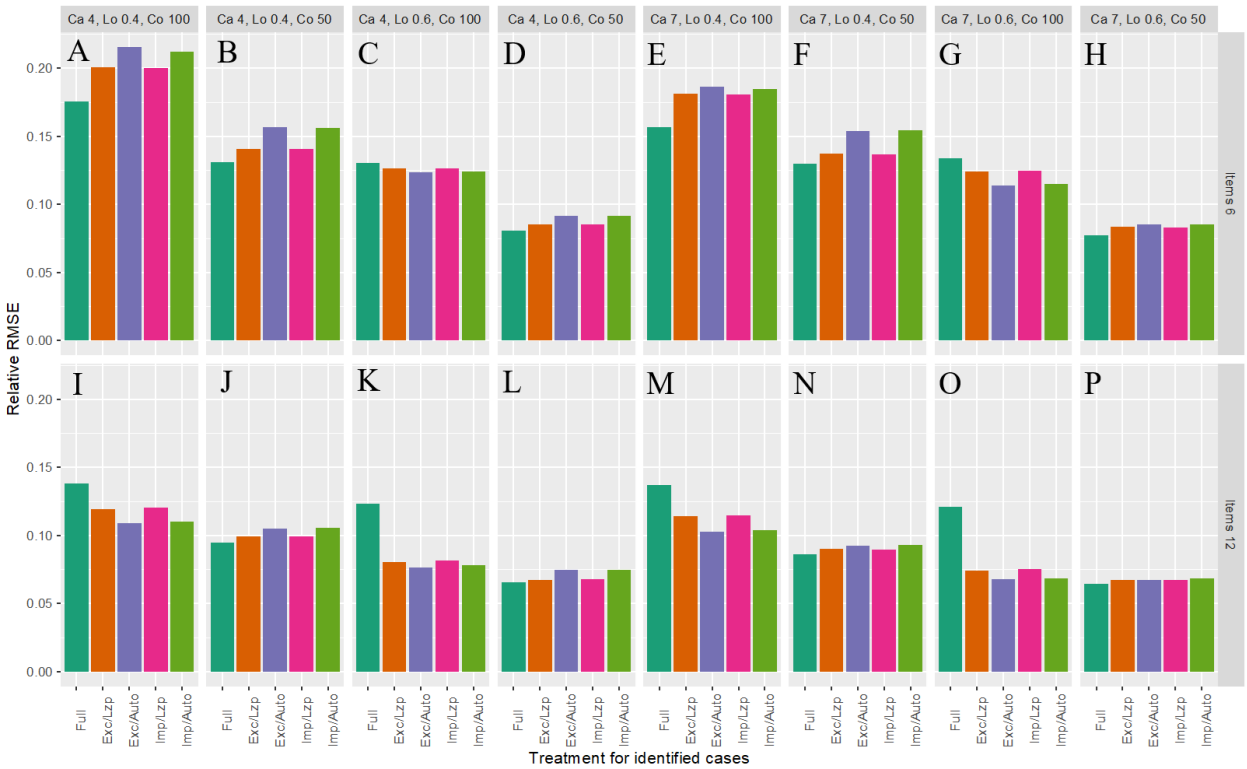

Figure 3.6. Relative root mean square error (RMSE) for the 20% CR condition with different treatment of CR ("complete case" denoted as "*Full*"; "casewise deletion for l_z^p or the autoencoder", denoted as "*Exc/Lzp*" or "*Exc/Auto*"; and "delete and multiply impute for l_z^p or the autoencoder", denoted as "*Imp/Lzp*" or "*Imp/Auto*") by scale length (six or 12 items), number of response categories (denoted as *Ca* 4 or 7), factor loadings (denoted as *Lo* 0.4 or 0.6), and percent of contaminated items (denoted as *Co* 50 or 100).

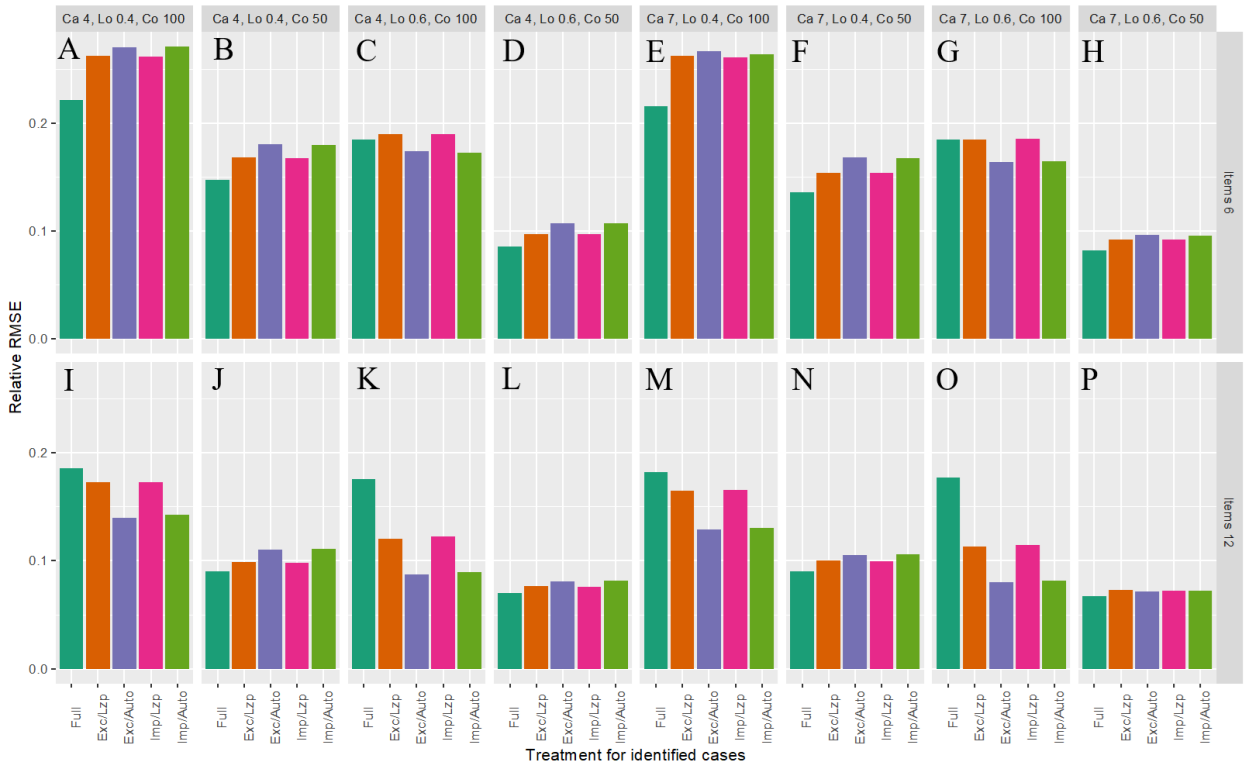

Figure 3.7. Relative root mean square error (RMSE) for the 30% CR condition with different treatment of CR's ("complete case" denoted as "*Full*"; "casewise deletion for l_z^p or the autoencoder", denoted as "*Exc/Lzp*" or "*Exc/Auto*"; and "delete and multiply impute for l_z^p or the autoencoder", denoted as "*Imp/Lzp*" or "*Imp/Auto*") by scale length (six or 12 items), number of response categories (denoted as *Ca* 4 or 7), factor loadings (denoted as *Lo* 0.4 or 0.6), and percent of contaminated items (denoted as *Co* 50 or 100).

3.4 Discussion

This research aimed to answer the question of how best to deal with CR data following identification across three different missing data approaches: "complete data analysis" (i.e., using the full sample), "casewise deletion", and the "delete and multiply impute" approaches. Several important factors influence the performances of these CR data treatment methods.

Scale length and the number of contaminated items impact the performances of casewise deletion and the delete and multiply impute CR data approaches. Specifically, the two approaches yield less relative bias and relative RMSE comparing to the "complete data analysis" approach when using CFA with a larger number of items and when CR employ careless response

behaviors in all of the items. Both casewise deletion and the delete and multiply impute approaches perform better in 12-item scales, likely due to more available information. As for the number of contaminated items, in the 50% contaminated condition where CR only employ careless responding behavior in 50% of the items, *all* of the identified CR data were deleted in the casewise deletion approach or the delete and multiply impute approaches. This is done because it is currently not possible to identify in which items CR behaviors occur. In other words, in the conditions with 50% contaminated items, the casewise deletion or the delete and multiply impute approaches not only deleted 50% contaminated items but also deleted the other good data. Comparing the two CR identification methods used in both casewise deletion and the delete and multiply impute approaches, the autoencoder clearly outperforms the standardized log-likelihood l_z^p for 12-item scales with all items contaminated, especially with larger proportions of CR in the data.

This chapter has important implications for survey research. Previous literature focusing on the identification of CR does not examine examine the best ways to deal with CR data after identification. Here, the delete and multiply impute approach is useful in dealing with CR data in CFA models with high factor loadings. The autoencoder method outperforms the l_z^p method under the delete and multiply impute approach to deal with CR data.

Both Chapters 2 and 3 share similar limitations. This chapter is based on a simulation study and may not well reflect real-world situations. However, despite the lack of a real data application, both the standardized log-likelihood l_z^p and the autoencoder were applied to real web survey data in Chapter 4, and results consistent with the simulation studies here were found there as well. The autoencoder outperforms the standardized log-likelihood l_z^p in scales with a small number of items and for data with good psychometric properties.

This chapter focuses exclusively on the identification of one satisficing behavior, random responses. For other types of satisficing behavior, such as response styles and response order effects (Jürges, 2007; Yan & Keusch, 2015), future studies can further examine the use of the autoencoder to identify CR of different types.

Due to computational complexity, this chapter fixes the sample size to 1000. It is unknown how these different approaches work with a smaller sample size. Future studies could examine the effects of sample size on the performances of these different approaches.

This chapter has identified several directions for future research. Future studies could examine the impacts of these approaches on parameters of the measurement models (e.g., loadings and thresholds) rather than a structural element such as a latent factor correlation. Future studies can also examine other person-fit statistics, especially those based on non-parametric methods.

Since this chapter suggests that the casewise deletion and the delete and multiply impute methods do not work well for CFA with low factor loadings, future study could evaluate other methods that work better in situations where the factor loadings of CFA are low.

The indicators of the CFA model were simulated using normal distributions, while in reality the distributions are likely to be skew. Future study could examine the performances of casewise deletion and the delete and multiply impute methods when using indicators from skewed distributions.

3.5 References

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum

Conijn, J.M., Emons, W.H.M., & Sijtsma, K. (2014). Statistic lz-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement*, *38*, 122-136.

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

De la Torre, J., \& Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, *45*, 159-177.

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

Hauser, D. J., & Schwarz, N. (2015). The war on prevention: Bellicose cancer metaphors hurt (some) prevention intentions. *Personality and Social Psychology Bulletin*, *41*, 66-77.

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

Jürges, H. (2007). True health vs response styles: Exploring cross-country differences in self-reported health. *Health Economics*, *16*, 163-178.

Krosnick, J. a. (1991). Response strategies for coping with the cognitive demands of attitude measures in survey. *Applied Cognitive Psychology*, *5*, 213-236.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York, NY: Wiley.

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

Meijer, R.R., Molenaar, I.W., & Sijtsma, K. (1994). Influence of test and person characteristics on non-parametric appropriateness measurement. *Applied Psychological Measurement*, *18*, 111-120.

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

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

Raghunathan, T., Berglund, P. A., & Solenberger, P. W. (2018). *Multiple imputation in practice: With examples using IVEware*. Boca Raton, FL: CRC Press.

Rezvan, P. H., Lee, K. J., & Simpson, J. A. (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, *15*, 30.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and Item Response Theory analyses. *Journal of Statistical Software*, *17*, 1-25.

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

Yan, T., & Keusch, F. (2015). The effects of the direction of rating scales on survey responses in a telephone survey. *Public Opinion Quarterly*, *79*, 145-165.

Zhang, C., & Conrad, F. G. (2013). Speeding in web surveys : The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*, 127-135.

CHAPTER IV

Trapped Respondents in Online Surveys: Detection and Adjustment Methods

4.1 Introduction

The use of web surveys (i.e., self-administered online surveys using computers or mobile devices) has increased substantially in the past two decades. Web surveys have several important advantages over other modes of data collection, including cost-effectiveness, more accurate answers, reduced social desirability bias, ability to conduct survey experiments involving complex randomization, ability to include visual design, and increased efficiency in data editing and data management (e.g., Couper, 2000; Benfield and Szlemko 2006; Kreuter, Presser, & Tourangeau, 2008). Web survey respondents are, however, likely to be less motivated and engaged in the web surveys than respondents to surveys administered by interviewers, being more likely to break off, provide more incomplete data, or employ more satisficing behaviors.

Survey cognitive theory suggests that respondents in general go through four required cognitive processing steps in answering survey questions: comprehending the question, retrieving relevant information from memory, integrating information to arrive at a judgment, and formulating and editing a response (Tourangeau, Rips, & Rasinski, 2009). Those who perform all four steps diligently for all survey questions are referred to as survey optimizers. However, not all respondents optimize in taking surveys. To reduce cognitive burden, some respondents skim instructions, respond in a haphazard fashion, or rush through questions.

Respondents with these behaviors are also known as survey satisficers (Krosnick, 1991), or careless respondents (CR).

There are different methods to evaluate satisficing behaviors in surveys, including quality indicators such as speeding and straightlining in grid questions. One method with increasing popularity for identifying respondents with satisficing behavior in web surveys is the *trap questions* (Oppenheimer, Meyvis, & Davidenko, 2009). Despite growing use, no studies have comprehensively compared the performance of trap questions with other methods to identify survey satisficers. There is also no consensus on how to deal with trapped respondents.

This chapter has three goals. The first is to examine the use of trap questions to identify CR. The second is to compare the standardized log-likelihood l_z^p and the autoencoder methods to identify CR (as discussed in Chapters 2 and 3) to a trap question method. The standardized log-likelihood l_z^p method is a person-fit statistic identifying the inconsistency of responses by comparing the model-expected responses to the actual reported responses (Der Flier, 1982). The autoencoder method, initially developed in engineering, can be used to identify anomaly and outlier cases (see Chapters 2 and 3 for more details) such as CR. The third goal is to explore how to best deal with trapped respondents, whether by excluding CR data or deleting and multiply imputing CR responses.

4.1.1 Survey Satisficing

Satisficing theory (Krosnick, 1991) was originally developed to explain consumers' decision making that did not maximize personal gain (Simon, 1956). The concept was proposed as an alternative to the common economic model as a means to describe conventional decision-making behaviors. Rather than expending the effort required to maximize the positive outcomes

of their decisions, Simon suggested that people expend only the effort necessary to make a satisfactory or acceptable decision – a strategy Simon called satisficing.

Krosnick (1991) later adapted Simon's theory to the field of survey research, creating a framework within which a variety of specific undesirable respondent behaviors might be better understood. When respondents perform all four cognitive processing steps specified in Tourangeau, Rips, and Rasinski (2009), for each survey question presented, they are said to be optimizing and not satisficing (Krosnick, 1991).

Respondents failing to carefully consider the meaning of questions or how to answer them may shift response strategies. Instead of performing the four steps necessary to optimize, they may choose to perform one or more of these steps in a cursory fashion (also known as weak satisficing) or even skip them altogether (i.e., strong satisficing). This shift from an optimal to an unsatisfactory decision strategy that gives satisficing theory its name.

4.1.2 Satisficing in Multi-Item Scales

Multi-item scales in surveys often appear in grid question formats. They are commonly used to measure respondent attitudes, behavior, health (e.g., mental health scale), wellbeing, and personality. Previous research identified different satisficing strategies associated with multi-item scales, covering both weak and strong forms of satisficing. Weak satisficing associated with multi-item scales includes response order effects, in which the order of the response categories affects respondents' answers (Yan & Keusch, 2015), and response styles, the tendency respondents use to favor certain categories (e.g., extreme categories), regardless of question (Baumgartner & Steenkamp, 2001). Strong satisficing related with multi-item scales mainly includes two forms, random responses and straightlining. As suggested by the name, random

responses refer to respondents "blindly" answering questions by "randomly" choosing responses, putting no thoughs into assigning answers. Straightlining respondents select the same response option in the grid for all or most of the questions without distinguishing question content. Both forms of strong satisficing originate from respondents' inattentive responses and are referred as careless responding in the survey literature. In this chapter, the focus is on strong satisficing in multi-item scales.

Careless responding can lead to damaging consequences to survey data quality. Random and inconsistent respondents in multi-item questions can bias survey estimates and increase measurement error (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Huang, Liu, & Bowling, 2015; McGrath, Mitchell, & Hough, 2010). It may reduce internal consistency, and attenuate the estimate of relationships among variables. Careless responses can have serious impacts psychometrically, distorting factor structures of measurement models evaluating certain constructs using multi-item scales (e.g., in Confirmatory Factor Analysis, or CFA). Johnson (2005) found that the inclusion of careless responses in CFA can lead to estimated factor structures with unnecessary "method" factors. Woods (2006) concluded in a simulation study that the inclusion of careless responses generated additional method factors in CFA.

Given its potential adverse impacts on data quality, it is important to identify CR and examine methods to reduce the negative impacts of careless responses. In the survey literature, careless responses are identified through evaluation of response time and straightline response patterns.

Recently, in web surveys, the trap question, or instructional manipulation checks (IMC), has gained popularity for detecting CR. Trap questions provide simplicity in application and flexibility in use. Most trap questions have "a lure question with lure responses" (Liu &

Wronski, 2018). An instruction directs respondents to ignore the lure question and provide a specific response following the instruction (Hauser & Schwarz, 2015). For example, trap questions embedded in multi-item scales may ask respondents to skip one item embedded in the list of items or select a pre-defined answer for a particular item.

In Chapters 2 and 3, two innovative approaches to identify CR in multi-item scales were examined, the standardized log-likelihood l_z^p and the autoencoder. The standardized log-likelihood l_z^p is a psychometric method to identify satisficing behavior (referred to as *aberrant responses* in the psychometrics literature; Meijer & Sijtsma, 2001). The autoencoder method is an unsupervised neural network initially used to identify anomaly and outlier cases in engineering applications (Amunategui, 2018; Chartea, Charteb, García, del Jesus, & Herrera, 2018). Here, these two approaches are compared to the trap question method to evaluate properties of all three methods to identify CR in survey data collection.

4.1.3 Trap Questions

Since first introduced by Oppenheimer et al. (2009), studies have found that those who failed trap questions spent significantly less time on the survey, provided less consistent answers, and introduced more measurement error than those who passed the trap questions. Berinsky, Margolis, and Sances (2014) found that trapped respondents could add noise to the data and change the significance and effect sizes in results. They suggested that to better measure respondents' attention to the survey, multiple trap questions be used. Liu and Wronski (2018) evaluated how failure on a trap question is correlated with other data quality measures, including straightlining, response rounding (where respondents provide less-precise round numbers divisible by 5 or 10; Holbrook et al., 2014), lengths (number of characters provided) of responses

to open-ended questions, and shorter time to complete survey questions (Ansolabehere & Schaffner, 2015). They conclude that trap questions are a promising way to identify satisficing behavior.

The trap question may be used for more than detection of CR. It can also serve as an intervention method to improve participants' attention and awareness in later questions. Hauser and Schwarz (2015) used a trap question at the beginning of a survey to improve respondents' systematic thinking, leading to better performance on cognitive reflection and probabilistic reasoning tasks later in a questionnaire.

Anduiza and Galais (2017) evaluated the causes of respondents failing to answer trap questions following instruction. Using education as a proxy for respondent ability, they found that those with lower education level are more likely to fail trap questions. They also examined the effects of intrinsic motivation, measured by interest in the survey topic, and material motivation, whether respondents mention material incentives among the two main reasons for completing the survey. They found intrinsic motivation but not material motivation plays a significant role in the failure to pass trap questions.

The difficulty level of trap questions can differ greatly depending on the format and design of trap questions. Anduiza and Galais (2017) found that among many factors tested, the difficulty level of the trap questions plays a critical role in respondent failure to successfully pass the questions. Liu and Wronski (2018) reported the passing rate of various trap questions varied widely, from 27% to 87%.

4.1.4 Approaches to deal with Trapped Respondents

There is no consensus on when respondents fail the trap question whether to remove or keep their data for analysis. Hauser and Schwarz (2015) suggest excluding CR data from any analysis. Several studies have reported the benefits of excluding trapped respondents from the analysis, including enhancing the validity and reliability of findings (e.g., Oppenheimer et al., 2009; Goodman, Cryder, & Cheema, 2013).

Other researchers were concerned that removing the trapped respondents may widen the existing survey biases, especially in situations where the success of passing the trap question is related to certain characteristics of respondents (Anduiza & Galais, 2017; Berinsky et al., 2014). For example, in a political survey, those who fail trap questions may be more likely to have less interest in politics. Removing them from analysis will lead to over-representation of those interested in politics. Berinsky et al. (2014) suggest analyzing results separately for those who pass and fail. Anduiza and Galais (2017) compared the inclusion and exclusion of the trapped respondents by evaluating correlations with external benchmarks obtained from representative face-to-face surveys. They found that including all respondents in the analysis produced higher correlation with the external benchmarks.

Trap questions may decrease the rapport between the respondents and the survey researcher. The inclusion of trap questions in web surveys may reduce respondent motivation to participate. As a result, several researchers suggest using trap questions cautiously (Anduiza & Galais, 2017). But no prior studies have compared trap questions to other identification methods. In this chapter, we compare the standardized log-likelihood l_z^p and the autoencoder to a trap question and explore whether these two alternative methods can perform as well as or better than the trap question.

The literature review has no consensus on how to best deal with trapped respondents. Contradictory results on survey data quality were found concerning the impact of including or excluding trapped respondents. Little attention has been paid to the consequences of reduction in sample size, such as loss in statistical power, when CR data is completely excluded.

In this chapter, deletion of CR data combined with multiple imputation methods are used with new approaches to detect CR and trapped respondents. For example, deleting and multiply imputing all trapped respondent data, or imputing subsamples of trapped respondents identified using person-fit statistics such as the standardized log-likelihood l_z^p or the autoencoder method are also examined. We hypothesize that deleting data from trapped respondents and multiply imputing the deleted data changes the statistical significance of results. Further, trapped respondents may introduce measurement error to the data, resulting in underestimation of associations when trapped respondent data are included in analysis (e.g., McKay, Garcia, Clapper, & Shultz, 2018). This underestimation might be overcome through deletion of trapped respondent data but with the addition of multiply imputing the deleted data.

Finally, we also hypothesize that the negative consequences of including trapped respondent data in analysis may be concentrated in a subset of CR data, particularly that portion providing the most inconsistent answers. Deleting and imputing data from this subset of trapped respondents only may reduce the attenuation of associations in data analysis.

4.2 Method

4.2.1 Data

Data were provided by a former graduate student, Mengyao Hu, in the Program in Survey Methodology at the University of Michigan, based on her dissertation research (Hu, 2018). A

Qualtrics non-probability online panel survey was designed to examine the use of anchoring vignettes in health measurement (Hu, 2018). Data collection was conducted from September to December 2017. The sample used quotas for different racial/ethnic groups – non-Hispanic White, non-Hispanic Black and English-speaking Hispanic and Spanish-speaking Hispanic. The sample had equal proportions of males and females, respondents with high school education or less and greater than high school education, and 18 - 49 year old and 50 year old or older respondents.[3] The online questionnaire took an average of 15 minutes to complete.

Since language effects are out of the scope of this study, the sample in the analyses of this chapter includes only respondents from three different racial/ethnic groups who responded to the English questionnaire. The sample size is n = 866.

4.2.2 Measures

The web survey data include two multi-item scales.

Satisfaction with life (SWLS) was measured using a five-item scale (Diener, Emmons, Larsen, & Griffin, 1985) where each item has a seven-point response from 1 = "*Strongly disagree*" to 7 = "*Strongly agree*". Higher scores indicate higher satisfaction. The five items in the SWLS were as following:

In most ways my life is far from my ideal.

The conditions of my life are excellent.

I am satisfied with my life.

So far I have gotten the important things I want in life.

If I could live my life over, I would change a lot of things.

[3] Initial sample size and response rates for this study are not available.

Depression was measured using a six-item self-reported scale. Each item had a five-point response category ranging from 1 = "*None of the time*" to 5 = "*All of the time*", with higher scores indicating increased depression. The question wording of this scale was as follows:

*In **the last 30 days**, about how often did you feel …*

nervous?

hopeless?

restless or fidgety?

worthless?

that everything was an effort?

so depressed that nothing could cheer you up?

A trap question item designed to detect CR was embedded in the satisfaction scale: *For quality purposes, please select option "Slightly disagree".*

Other variables measured in the web survey included respondent age (years), gender (female = 1 vs. male = 0), education (six categories ranging from *less than high school* to *postgraduate degree*), and race / ethnicity (non-Hispanic White, non-Hispanic Black, and Hispanic). Response time (in seconds) was collected at the end of each page. For satisfaction and depression scales, question items for each scale were presented on a single page.

A straightlining indicator variable was constructed for the SWLS and Depression scales, where those who did not distinguish the question items and selected the same categories across all the items in the scale are coded as 1.

4.2.3 Statistical Analysis

A structural equation model (SEM) was fit to the data in which the association between the two scales (see Figure 4.1) as assessed in two latent variables, satisfaction with life (five items) and depression (six items) was examined. The satisfaction latent variable is the dependent variable and the depression latent variable and four demographic items are predictors. Model coefficient and fit indices were examined across different methods of detection of CR and techniques for handling trapped respondent or CR data in analysis.

Parameter estimates from the SEM were compared for 11 samples. Sample one consists of the full sample – no CR were removed. Sample two excluded all trapped respondent data. The third and fourth samples excluded CR identified using the standardized log-likelihood l_z^p and the autoencoder. Samples five and six considered speeding (those whose response time was in the fastest quintile of response time) and straightlining (straightline response for the SWLS scale). Samples seven, eight, and nine consisted of all respondents but data from CR identified by the trap question, the standardized log-likelihood l_z^p, or the autoencoder was deleted and multiply imputed. Finally, samples 10 and 11 consisted of all respondents but scale data from CR who failed the trap question and were identified also by the standardized log-likelihood l_z^p or the autoencoder method was deleted and multiply imputed.

The multiple imputation was based on predicted mean matching sequential regression multiple imputation (all variables, scales and demographic, were used in the sequential regression models) with 150 imputed datasets generated (Raghunathan, Berglund, & Solenberg, 2018). Multiple imputation parameter estimates and their standard errors were obtained using the Rubin's combining rules. The SEM was estimated for each of the 150 imputed datasets and combined using Rubin's combining rules (Raghunathan, Berglund, & Solenberg, 2018). All the

analyses described above were conducted using R 3.5.1 (R Core Team, 2018) and Mplus 8.1 (Muthén & Muthén, 2017).

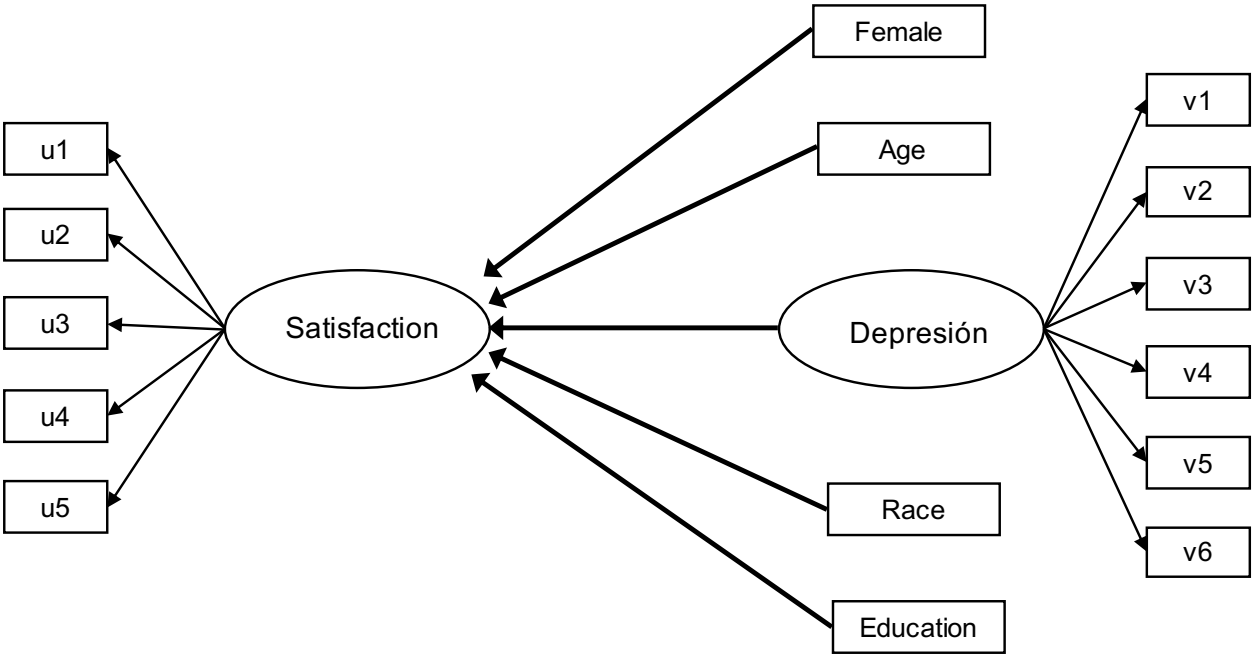

Figure 4.1. SEM model illustration.

4.3 Results

Of the 866 English-language respondents completing the Qualtrics survey, 204 (23.6%) failed the trap question. The first two columns of Table 4.1 describe the characteristics for the trapped and non-trapped respondents. Slightly more than half (51.7%) of the trapped respondents were female, compared to 48.5% of the non-trapped respondents. The mean age for trapped respondents was five years younger than those who passed the trap question. About half of the trapped respondents were Black, but only 33.4% of the non-trapped respondents were. The mean response time for trapped respondents was 36 seconds, while for the non-trapped respondents it was more than 50 seconds. About 25% of the trapped respondents employed a straightlining

strategy and 7.8% of the trapped respondents straightlined at the mid-response categories throughout the scale.

Table 4.1 also includes descriptive statistics for CR identified using other satisficing-detection methods, including the standardized log-likelihood l_z^p, the autoencoder, straightlining, speeding, and combinations of the trap question and the standardized log-likelihood l_z^p or the autoencoder method. The standardized log-likelihood l_z^p detected 81 CR. More than 50% of these identified CR were Black. They had a response time that was not much different from the non-trapped respondents (53.1 seconds). About 10% of the l_z^p-identified CR straightlined in the scales (i.e., selected same response options throughout the items in the scale).

The autoencoder method detected 200 CR, 47% of whom were female and 50% were Black. The mean response time for autoencoder detected CR was 48.2 seconds, and 17% employed a somewhat higher rate of straightlining. Neither the standardized log-likelihood l_z^p nor the autoencoder detected straightlining behavior at the mid-point response categories. A total of 58 respondents were identified as "straightliners". The straightliners' mean response time was faster at 20.3 seconds, 34.5% straightlined at the mid-response categories. A total of 174 respondents were identified as "speeders". About 20% of the speeders straightlined and 9% of the speeders straightlined at the mid-point response categories.

Finally, results are also shown for combinations of detection methods. The combined trap question and standardized log-likelihood l_z^p detection identified 35 CR. About one fifth (22.9%) of them straightlined. The combined trap question and autoencoder detection identified 85 CR. More than one third (35.3%) of them showing straightlining behaviors. Among all these groups, there were surprisingly few important education differences. However, all CR groups except "speeders" had fewer respondents with higher education.

Among the 58 “straightliners” in the sample, 86.2% of them were identified as CR in the trap question method, 13.8% were identified by the standardized log-likelihood l_z^p, and 56.9% were identified by the autoencoder. Both the standardized log-likelihood l_z^p and the autoencoder did not detect those who straightlined at the mid-response categories. Among the 174 speeders, 55.7% of them were identified using the trap question method, 14.9% of them were identified by the standardized log-likelihood l_z^p, and 37.9% of them were identified by the autoencoder.

Table 4.1. Female, age, race, education, response time, straightlining, and straightlining at mid-point of the SWLS scale for non-trapped respondents and CR identified using the trap question, the standardized log-likelihood l_z^p, the autoencoder, straightlining, speeding, and combinations of the trap question and the standardized log-likelihood l_z^p or the autoencoder method.

	Non-Trapped *N= 662* *% / (M)*	*Trapped* *N = 204* *% / (M)*	*l_z^p* *N = 81* *% / (M)*	*Auto* *N = 200* *% / (M)*	*Straight-Lining* *N = 58* *% / (M)*	*Speeders* *N = 174* *% / (M)*	*Trapped + l_z^p* *N = 35* *% / (M)*	*Trapped + Auto* *N = 85* *% / (M)*
Female	51.7	48.5	48.1	47.0	50.0	49.4	31.4	42.4
Age	(48.2)	(43.4)	(43.9)	(44.8)	(41.8)	(37.8)	(39.3)	(39.7)
Race								
White	35.2	30.9	25.9	26.5	22.4	29.9	20.0	21.2
Black	33.4	49.5	54.3	50.5	46.6	44.8	60.0	60.0
Hispanic	31.4	19.6	19.8	23.0	31.0	25.3	20.0	18.8
Education								
Less than high school graduate or less than GED	3.8	5.9	8.6	6.5	5.2	3.4	8.6	8.2
High school graduate or GED	45.2	49.5	56.8	55.0	50.0	46.6	60.0	52.9
Some college (no degree obtained)	14.5	14.7	14.8	14.0	15.5	12.6	14.3	14.1
Associate's degree	8.3	4.9	3.7	5.5	8.6	7.5	2.9	4.7
Bachelor's degree	16.5	15.7	8.6	10.5	12.1	16.7	5.7	11.8
Postgraduate (master's/professional/doctorate) degree	11.8	9.3	7.4	8.5	8.6	13.2	8.6	8.2
Response Time (secs.)	(52.2)	(36.0)	(53.1)	(48.2)	(20.3)	(15.5)	(33.4)	(31.8)
% straightlining	1.2	24.5	9.9	16.5	100.0	22.4	22.9	35.3
% straightlining at midpoint	0.6	7.8	0.0	0.0	34.5	9.2	0.0	0.0

To evaluate whether trapped respondents also exhibit other satisficing behavior, we examined whether trapped respondents spent less time answering questions and exhibiting more straightlining behavior. Table 4.2 presents results from an estimated logistic regression model predicting the log odds of being trapped. The model includes interaction effects of sex, age, and time with the straightlining indicator. The Likelihood Ratio Test comparing this *Full model* with a *Null model* that excluded all the interactions showed the null model cannot be rejected ($\chi^2(3) = 0.293$, $p = .573$).

Younger respondents were more likely to be trapped, as were Non-Hispanic Black and White respondents compared to Hispanic respondents. Those who spent less time answering questions were also more likely to be trapped, as were those who used straightlining in answering questions.

Table 4.2. Logistic regression analysis predicting being trapped ($n = 866$).

	OR	*SE*
Female	0.975	0.029
Age	0.981*	0.008
Race / ethnicity (ref = Hispanic)		
Non-Hispanic White	1.069*	0.033
Non-Hispanic Black	1.153***	0.032
Time	0.999**	0.0003
Straightlining	1.979***	0.099
Education	0.989	0.009
Female × Straightlining	1.071	0.106
Age × Straightlining	1.024	0.034
Time × Straightlining	0.996	0.003
Intercept	1.239***	0.043

*: $p < .05$; **: $p < .01$; ***: $p < .001$.

Trap question detection was compared to the other methods in terms of the SEM model outcome. Table 4.3 presents the model fit indices for the tested SEM models based on the full sample, sample without trapped respondents, and the imputed samples. In total, there were five

imputed sample conditions. In the first three conditions, only the trap question (i.e., *Trap* in Table 4.3), the standardized log-likelihood l_z^p, or the autoencoder (i.e., *Auto* in Table 4.3) was used to identify CR for whom scale data was multiply imputed. In the last two imputed conditions, two detection methods were combined, the standardized log-likelihood l_z^p with the trap question (l_z^p *& Trap* in Table 4.3) and the autoencoder with the trap question (*Auto & Trap* in Table 4.3).

In the *Trap* imputation condition, scale data of all 204 trapped respondents were deleted and multiply imputed, while for the l_z^p imputation condition, scale data were deleted and multiply imputed for 81 respondents. Under the *Auto* condition 200 respondents were classified as CR. For the l_z^p *& Trap* condition, scale data for 35 respondents were deleted and multiply imputed, and for the *Auto & Trap* condition, scale data for 85 respondents were deleted and multiply imputed.

Not surprisingly, the χ^2 fit statistics are significant for the full sample and non-trap subsample. Other model fit criteria including RMSEA, CFI and TLI suggest that the models fit the data well in all the conditions. Both CFI and TFL are above 0.90 for all sample conditions, and RMSEA is below 0.08 for all as well. The autoencoder and *Auto/trap* conditions have the best model fit indices, with lowest RMSEA and highest CFI and TLI among all conditions.

Table 4.3. SEM fit indices for full sample, non-trap subsample, and samples with different imputed scale data.

	Without Imputation		*With Imputation (m = 150[a])*				
	Full sample	*Non-trap subsample*	*Trap*	l_z^p	*Auto*	l_z^p & *Trap*	*Auto & Trap*
Chi2 (93)	224.439***	176.125***	190.607[b]	210.284	183.079	211.437	183.079
RMSEA	0.040	0.037	0.035	0.038	0.033	0.038	0.033
CI 95% RMSEA	(0.034, 0.047)	(0.028, 0.045)	----[b]	----	----	----	----
CFI	0.918	0.943	0.947	0.932	0.956	0.928	0.956
TLI	0.902	0.933	0.938	0.920	0.948	0.915	0.948
Cases imputed	0	0	204	81	200	35	85
N	866	662	866	866	866	866	866

[a]: Number of imputed datasets.
[b]: Mplus does not provide significance and confidence intervals when using imputed datasets.
*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

Tables 4.4 and 4.5 presents the SEM coefficients for the different samples. The model based on the 662 non-trapped respondents only (Model 1) is considered as the benchmark model, as no CR are included in the model. Model 2 is based on the full sample (sample size = 866) which includes both trapped and non-trapped respondents. Comparing Model 2 with Model 1, the size of coefficients changed in the model which included trapped respondents (Model 2) compared to Model 1, where the absolute value of the coefficient for depression in Model 2 is 20% smaller than in Model 1. The statistical significance at the 0.05 level of the coefficients also changed for Blacks, failing to achieve significance when trapped respondents are included in the sample (Model 2). The change in the absolute value of the regression coefficient is about 40%. If one conducts analysis using the full sample, race - ethnicity does not have an effect on satisfaction, while if non-trapped respondents only are examined, Blacks are less satisfied with their life than English-speaking Hispanics.

Table 4.4. SEM results based on full sample and non-trap subsample.

Predictors	*Model 1 Non-trap respondents*		*Model 2 Full sample*	
	B	*SE*	*B*	*SE*
Depression	-0.676***	0.026	-0.539***	0.026
Female	0.017	0.084	-0.028	0.066
Age	0.072**	0.024	0.061***	0.018
White	0.092	0.100	0.107	0.086
Black	-0.196*	0.096	-0.117	0.078
Education	0.177***	0.029	0.149***	0.023
Cases imputed	0		0	
N	662		866	

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

Table 4.5. SEM results for samples with different imputed cases (imputed cases identified by trap question, l_z^p and autoencoder).

	Model 3 Trap		*Model 4 l_z^p*		*Model 5 Auto*		*Model 6 l_z^p/trap*		*Model 7 Auto/trap*	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Depression	-0.686***	0.029	-0.576***	0.027	-0.660***	0.029	-0.571***	0.026	-0.660***	0.029
Female	0.030	0.080	-0.039	0.070	-0.030	0.077	-0.035	0.069	-0.030	0.077
Age	0.082***	0.024	0.061**	0.020	0.087***	0.022	0.064***	0.020	0.087***	0.022
White	0.096	0.097	0.086	0.087	0.084	0.092	0.091	0.089	0.084	0.092
Black	-0.202*	0.095	-0.138	0.083	-0.202*	0.092	-0.120	0.081	-0.202*	0.092
Education	0.165***	0.026	0.146***	0.024	0.152***	0.026	0.150***	0.024	0.152***	0.026
Cases imputed	204		81		200		35		85	
N	866		866		866		866		866	

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

Models 3 to 5 in Table 4.5 examine whether using the l_z^p and autoencoder methods to detect CR for deletion and multiply imputation can be used as an alternative to trap questions. In Model 3, we deleted and imputed responses to the satisfaction scale for all the 204 trapped respondents. Results of the SEM in Model 3 are similar as those in Model 1, indicating that the imputation of all trapped respondents provides almost equivalent results as removing all trapped respondents.

In Model 4, the standardized log-likelihood l_z^p identified 81 CR and their responses to the satisfaction scale were then deleted and multiply imputed for analysis. This method does not appear to provide any additional benefit compared to Model 2 which included all trapped respondents. There the trap question outperforms the standardized log-likelihood l_z^p in identifying CR.

After deleting and imputing CR scale data, Model 5 results were similar to Models 1 and 3. Imputation based on the autoencoder method yields model estimates nearly the same as the trap question method.

To examine whether removing a subset of trapped respondents can provide similar results as Model 1 and Model 3, two other delete and multiply impute strategies were used in Models 6 and 7. In Model 6, the l_z^p results in Model 4 was used as an additional criterion to identify the most influential CR in trapped respondents: those who failed the trap question and are also identified as CR (the intersection of the l_z^p and the trap question method). After their responses were deleted and multiply imputed, in Model 6, results were similar as in Model 4. In this application, the combination of the l_z^p and trap question methods to detect CR was no different than using all respondents (Model 2).

Finally, in Model 7, the autoencoder and trap question were combined to identify a subset of 85 of the 204 trapped respondents. After deletion and multiple imputation of the scale items, the results of this approach are similar to Models 1, 3 and 5. It appears unnecessary to delete all trapped respondents. The autoencoder performs well in identifying the most influential CR cases among trapped respondents.

4.4 Discussion

In this chapter, many results were consistent with previous literature. For example, those who were trapped were more likely to speed and employ straightlining strategies.

The trap question method performs the best in terms of detecting straightlining at the mid-point category. This is because in a multi-item scale with ordinal response categories, the mid-point is often the neutral category, which remains the same even when some items in the scales are reversely worded. Those with neutral opinions or perceptions will likely choose the mid-point categories throughout the items, making the straightlining at mid-point categories a reasonable response patterns. The standardized log-likelihood l_z^p and the autoencoder, which detect inconsistent patterns, cannot disentangle the true neutral response patterns with CR who satisficed and straightlined at the mid-point categories. On the other hand, the autoencoder method performs better in detecting random response patterns than the trap question and the standardized log-likelihood l_z^p methods. As for the model estimation and model fit of the three methods, the model fit indices and model results suggest that the use of the standardized log-likelihood l_z^p turns out to be no better than keeping all trapped responses in the analysis. On the other hand, as expected, the autoencoder method provides equivalent results to those of the trap

question, indicating that the autoencoder is a useful tool to identify careless response in multi-item scales, and can be used as an alternative to the trap question method.

These results in this chapter are consistent with Chapters 2 and 3 in which the autoencoder outperforms the l_z^p in identifying CR. Specifically, in Chapter 2, we found the autoencoder can identify larger proportions of CR than the standardized log-likelihood l_z^p, and in Chapter 3 imputation based on the autoencoder identification shows less bias compared to imputation based on the l_z^p CR identification method.

The use of the l_z^p or autoencoder to identify the most-concerning subset among the 204 trapped respondents showed contrasting results. The l_z^p (Model 4 in Table 4.5) does not yield any additional gains from the "do nothing" approach (Model 2 in Table 4.4). Combining the autoencoder and trap question methods enables identification of a subset of the trapped respondents that may have the most-influential impacts on data quality. Deleting and multiply imputing the data of this subset in analysis in Model 7 led to similar results (e.g., similar model fit indices and regression coefficients) and smaller standard errors for almost all regression coefficients. This suggests that it is desirable not to exclude all trapped respondents in analysis. Researchers can consider using the autoencoder to identify the most-influential subset of trapped respondents to delete and multiply impute in analysis.

These results support previous findings that trap questions are effective in identifying CR in web surveys. The autoencoder can also be used as an alternative method to the trap question. For researchers who are concerned about the potential negative effects of trap question method such as increasing respondents' confusion, reducing respondent motivations, and changing respondent behavior, using the autoencoder to identify CR may be an appealing alternative approach.

This research showed that imputation for trapped respondents is a useful tool to increase statistical power as well as provide valid results in SEM for multi-item scales. Researchers also do not need to delete all responses of trapped respondents. Researchers can consider combining the autoencoder and trap question methods to detect CR and retain more than half of the trapped respondents in the analysis and based on results observed in this application may achieve results that are consistent with deleting all trapped respondent data. Future web surveys that include trap questions in multi-item scales can consider imputing the subset of trapped respondents also detected as CR by the autoencoder method as illustrated in this study.

This chapter uses data from a nonprobability-based sample. It is unknown how these results can be generalized to the population. Despite the nonprobability sample, this study is still a valuable first step in evaluating alternative methods to trap questions and examining ways to deal with trapped respondents in web surveys. Future studies should replicate these detection and deletion/imputation methods using probability samples.

This study only examined one five-item scale with seven response categories each. It is unknown whether the performance of the combined autoencoder and trap question methods can differ with large scales or different numbers of item response categories. Future studies could address these issues.

As discussed in previous literature, the failure rate for trap questions has a very wide range (e.g., from 20% to 80%) depending on factors including difficulty of trap questions and sample composition. In this study, the specific sample had about 24% trapped respondents. It is unknown whether these methods will perform differently with different percentages of trap question CR in the sample.

There exist multiple types of person fit indices, this dissertation examined the parametric l_z^p because this statistic has been recommended based on previous simulation studies. Future studies could further evaluate other person fit indices, especially those based on non-parametric methods.

This study used the autoencoder to help to identify the most-influential subset of the trapped respondents. Future studies should examine the use of other quality indicators (e.g., straightlining and speeding, if paradata were available) to identify other such subsets.

Finally, this study compared the results of these different methods to identify CR on an SEM. Future studies should examine the performance of these methods for different models.

4.5 References

Amunategui, M. (2018). A practical look at anomaly detection using autoencoders with H2O and the R programming language. In M. Roopaei & P. Rad [eds.], *Applied Cloud Deep Semantic Recognition. Advanced Anomaly Detection*. Boca Raton, FL: CRC Press.

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

Ansolabehere, S. & Schaffner, B.F. (2015). Distractions: The incidence and consequences of interruptions for survey respondents. *Journal of Survey Statistics and Methodology, 3*, 216.-239.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research, 38*, 143-156.

Benfield, J.A. & Szlemko, W.J. (2006). Internet-based data collection: Promises and realities. *Journal of Research Practice, 2*, 1-15.

Berinsky, A.J., Margolis, M.F., & Sances, M.W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science, 58*, 739-753.

Charte, D., Charte, F., García, S., del Jesus, M. J., & Herrera, F. (2018). A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. *Information Fusion*, *44*, 78-96.

Couper, M.P. (2000). Review: Web surveys: a review of issues and approaches. *Public Opinion Quarterly, 64*, 464-494.

Diener, E., Emmons, R.A., Larsen, R.J. & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71-75.

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.

Hauser, D.J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "Tricky" tasks. *SAGE Open*, 1-6.

Holbrook, A. L., Anand, S., Johnson, T. P., Cho, Y. I., Shavitt, S., Chávez, N., & Weiner, S. (2014). Response heaping in interviewer-administered surveys: Is it really a form of satisficing? *Public Opinion Quarterly, 78*, 591-633.

Hu, M. (2018). *Anchoring Vignettes for Health Comparisons: The Validity of a Multidimensional IRT Model Approach and Design Improvements Using Visual Vignettes*. (Unpublished doctoral dissertation). University of Michigan, Ann Arbor, Michigan.

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

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

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

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Surveys: the effects of mode and question sensitivity. *Public Opinion Quarterly, 72*, 847-865.

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

Liu, M., & Wronski, L. (2018). Trap questions in online surveys: Results from three web survey experiments. *International Journal of Market Research, 60*, 32-49.

McKay, A.S., Garcia, D.M., Clapper, J.P., & Shultz, K.S. (2018). The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Computers in Human Behavior, 84*, 295-303.

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

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

Muthén, L.K. & Muthén, B.O. (1998-2017). *Mplus User's Guide* (8th Edition). Los Angeles, CA: Muthén & Muthén.

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

Raghunathan, T., Berglund, P.A., & Solenberger, P.W. (2018). *Multiple Imputation in Practice*. Boca Raton, FL: CRC Press.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*, 129-138.

Tourangeau, R., Rips, L.J., & Rasinski, K.A. (2009). *The psychology of survey response*. Cambridge: Cambridge University Press.

Van Der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*, 267-298.

Peer, E., Vosgerau, J. & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*, 1023-1031.

Woods, C.M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*, 189-194.

Yan, T. & Keusch, F. (2015). The effects of the direction of rating scales on survey responses in a telephone survey. *Public Opinion Quarterly, 79*, 145-165.

CHAPTER V

Conclusion

This dissertation focused on multi-item response scales, which are widely used in surveys to assess a variety of constructs including respondents' attitudes, behavior, health (e.g., mental health scale), wellbeing and personality. Multi-item scales often appear in grid question formats with the same response options for a set of survey question items. Previous literature has extensively discussed ways to identify satisficing behavior in these grid question scales, including the detection of response order effects, response styles, straightlining, and speeding. In web surveys, one method which has obtained increasing popularity to identify careless respondents, is the use of trap questions, also known as instructional manipulation checks (Hauser & Schwarz, 2015; Oppenheimer, Meyvis, & Davidenko, 2009).

Despite the large survey literature examining satisficing behavior in grid questions, there are two methods overlooked in the survey literature. One method is the person-fit-statistic, which has been extensively discussed in psychometric literature. This method identifies the inconsistency of responses by comparing the expected responses (based on the psychometric model) to the actual reported responses (van der Flier, 1982). One of the most popular person-fit statistic is called the standardized log-likelihood l_z^p, which has been proven to be a useful tool in multi-item scales with a large number of question items to identify patterns of inconsistent responses (Conijn, Franz, Emons, de Beurs, & Carlier, 2019). The other method is the

autoencoder method, which was initially developed and used in engineering to identify anomalies and outlier cases (Chen, Sathe, Aggarwal, & Turaga, 2017).

The purpose of this dissertation was to expand the existing literature on survey satisficing in multi-item scale grid questions in three directions, namely the identification of careless respondents in multi-item scale questions using the aforementioned two innovative methods – the standardized log-likelihood l_z^p and the autoencoder; the approaches to treat identified careless respondents; and the use of these two methods as an alternative to trap questions in web surveys. This dissertation has three primary objectives. The first objective was to compare the use of the standardized log-likelihood l_z^p and the autoencoder in identifying careless respondents in a multi-item scale with a small number of items (Chapter 2). This dissertation is the first study that introduces both methods into the survey field and applies them in the identification of satisficing behaviors. The second objective was to examine the use of multiple imputation to deal with data of the identified careless respondents, comparing results to two previously adopted approaches – excluding all careless respondents' data or keeping all careless respondents' data (Chapter 3). The third objective was to evaluate the use of the standardized log-likelihood l_z^p and the autoencoder as an alternative to trap questions and explore how to best deal with trapped respondents (Chapter 4). The ultimate goal was to provide some practical evidence as well as scientific contributions to improving future identification of careless respondents and treatment of their data.

Chapter 2 examines the optimal way of applying the autoencoder method to identify CR and compares the performances of the standardized log-likelihood l_z^p and the autoencoder in identifying CR in a multi-item scale with a small number of items. More specifically, Chapter 2 conducts a simulation study based on a full factorial experiment design with six factors: the

number of question items in the multi-item scale (6 vs. 12); the number of response categories (4 vs. 7); the quality of the scale (items with medium loadings vs. low loadings); CR types (*random response behavior* and *non-differentiation of item direction changes*, which is specifically for items include both positive and negative wordings); proportion of careless respondents in the simulated dataset (10%, 20%, and 30%); and proportion of the items that careless respondents employed satisficing behavior (half vs. all items). To evaluate the optimal number of iterations for the autoencoder method, Chapter 2 also examined the performances of the autoencoder method using different number of iterations. It is found in Chapter 2 that the autoencoder with two iterations works the best with increased sensitivity and acceptable false positive rates. The autoencoder with two iterations can identify more CR (higher sensitivity) in all conditions, compared to the standardized log-likelihood l_z^p. These findings highlight the promising use of the autoencoder method to identify CR in multi-item scales.

Chapter 3 extends the research in Chapter 2, and aims to answer the natural follow-up question after identifying careless respondents: what to do with their responses? Specifically, Chapter 3 compares three approaches in treating data of CR, including using the full sample or "complete data analysis", excluding all CR data and deleting and imputing CR data. Results of this chapter show that the quality of the Confirmatory Factor Analysis (CFA) model impacts the performances of excluding and imputing CR data approaches – specifically, the two approaches yield less relative bias and relative RMSE comparing to the "complete data analysis" approach when using CFA with high factor loadings but not when using CFA with low factor loadings. Consistent with what we found in Chapter 2, in CFA with high factor loadings, the exclusion or imputation of CR data using the autoencoder identification outperforms the approaches when using the standardized log-likelihood l_z^p identification of CR. It shows the great potential of

using multiple imputation combined with the autoencoder method to deal with data of CR in multi-item scales with high factor loadings in CFA models. This provides valuable information to the researchers who would like to decide what method to use for CR identification.

Chapter 4 studies whether the standardized log-likelihood l_z^p and the autoencoder can be used as an alternative to the trap question and whether it is possible to not delete (and impute) all CR data. Data for Chapter 4 are based on a web survey conducted through a Qualtrics online panel. Chapter 4 used a Structural Equation Model (SEM) and compared model results for the standardized log-likelihood l_z^p, the autoencoder and the trap question method. Chapter 4 also examined different approaches of identifying the most concerning subset of trapped respondents and compared results of imputing only the identified subset with results of deleting/imputing all trapped respondents' data. Results of that chapter suggest that the autoencoder method may provide equivalent results as the use of trap questions, indicating that the autoencoder is a useful tool to identify CR in multi-item scales, and can be used as an alternative to the trap question method. In addition, we found in this chapter that it is may be possible to not exclude all trapped respondents in analysis if researchers consider using the autoencoder to identify the most-concerning subset of trapped respondents. Chapter 4 is most similar to the simulation of non-differentiation CR behaviors in Chapter 2. Note that the difference between the autoencoder and the standardized log-likelihood l_z^p seems to be greater in Chapter 4, comparing to Chapter 2. This is likely due to the distribution of the response categories. The distribution of response categories in the simulation study (i.e., Chapter 2) is normal (see Appendix 2.1). Thus, when creating CR responses in Chapter 2, few of them are non-differentiation cases at the extreme responses. On the other hand, in the real dataset in Chapter 3, the distribution is left skewed with the majority of respondents reporting higher levels of satisfaction with their life. If CR with extreme responses

did not pay attention to the directions of the item wordings, their responses would be more inconsistent than those reporting neutral or close-to-neutral life satisfaction. This inconsistency may be more likely to be detected in the autoencoder method. In order to evaluate the effects of response distributions, future studies could conduct simulation studies with different response distribution conditions. This chapter fills the research gap in trap question literature and shed light on how to best deal with trapped respondents.

This dissertation presented a first attempt to examine several previously unexplored issues related to the identification of satisficing behaviors in multi-item scales. In summary, the newly introduced autoencoder method outperforms the standardized log-likelihood l_z^p in the identification of CR in multi-item scales with a small number of question items. Consistent with previous literature (e.g., Liu & Wronski, 2018), the trap question method is found to be effective in detecting satisficing behaviors in web surveys such as speeding and straightlining. The autoencoder method is also found to be useful in the detection of CR including those who speed and employ straightlining strategy. Among the three methods – trap question, the standardized log-likelihood l_z^p and the autoencoder, the trap question method performs the best in terms of detecting straightlining at mid-point category. This is because in a multi-item scale with ordinal response options, the mid-point category is often the neutral category, which remains the same even when some items in the scales are reversely worded. Those with neutral opinions or perceptions will likely choose the mid-point categories throughout the items, making the straightlining at mid-point categories a reasonable response patterns. The standardized log-likelihood l_z^p and the autoencoder, which detect inconsistent patterns, thus cannot disentangle the true neutral response patterns with CR who straightline at the mid-point categories. On the other hand, the autoencoder method performs better in detecting random response patterns than the

trap question and the standardized log-likelihood l_z^p methods. The model fit indices and model results suggest that the autoencoder method can be used as an effective alternative to the trap question method. The combination of both trap question and the autoencoder methods can detect the most-concerning subset of trapped CR, making it possible to delete and impute only a subset of trapped respondents' data.

From a methodological perspective, the results of this dissertation create basic theoretical knowledge regarding the identification of CR using the standardized log-likelihood l_z^p and the autoencoder methods. This dissertation also provides directions for future research of these methods. From a practical perspective, this research provides researchers and survey practitioners more options to identify careless respondents other than trap questions. As multi-item scales are ubiquitous in all types of surveys, this research provides an initial guide on identifying and dealing with careless respondents to researchers from different fields who use multi-item scales.

References

Chen, J., Sathe, S., Aggarwal, C., & Turaga, D. (2017). Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining* (pp. 90-98).

Conijn, J. M., Franz, G., Emons, W. H. M., de Beurs, E., & Carlier, I. V. E. (2019). The Assessment and Impact of Careless Responding in Routine Outcome Monitoring within Mental Health Care. *Multivariate Behavioral Research, 54*, 593-611.

Hauser, D. J., & Schwarz, N. (2015). The war on prevention: Bellicose cancer metaphors hurt (some) prevention intentions. *Personality and Social Psychology Bulletin*, *41*, 66-77.

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

Van Der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*, 267-298.