The effects of careless survey responding on the fit of latent variable models: A simulation study

by

Nathaniel Mark Voss

B.A., Bethel University, 2014
M.S., Kansas State University, 2019

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

Surveys are one of the most popular, useful, and efficient methods for gathering data within both academic and organizational settings. Despite the many benefits afforded by surveys, research shows that a nontrivial number of people engage in careless responding (CR) such that their responses to surveys are not effortful, accurate or valid. This is problematic because the failure to account for CR can distort research findings, result in false theoretical conclusions and lead to precarious workplace decisions. With surveys, it is common to model responses with a latent variable framework and use fit indices to makes conclusions about how well the model represents the data. Research shows that CR can be associated with poor fit, good fit, and/or unrelated to fit. To better understand how CR affects fit, the primary goal of this study was to examine the consequences of CR on the fit of latent variable models using a comprehensive, realistic and rigorous simulation paradigm. A secondary goal was to better elucidate the nature of CR and specify how CR behavior manifests within survey responses. In Study 1, participants' survey response patterns were experimentally shaped. In Study 2, these results were used as a basis for the primary simulation. A total of 144 conditions (which varied the sample size, number of items, CR prevalence, CR severity, and CR type), two latent variable models (item response theory and confirmatory factor analysis), and six model fit indices were examined. Overall, the results of this study show that CR is frequently associated with deteriorations in model fit. These effects are, however, highly nuanced, variable, and contingent on many factors. Moreover, this study demonstrates that good fit is not necessarily indicative of careful responding nor is poor fit always emblematic of CR. Rather, model fit and CR/response validity are distinct issues that must be separately addressed. These findings can be leveraged by researchers to develop more accurate theories and practitioners to better manage survey data that is laden with CR.

The effects of careless survey responding on the fit of latent variable models: A simulation study

by

Nathaniel Mark Voss

B.A., Bethel University, 2014
M.S., Kansas State University, 2019

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Jin Lee, Ph.D.

Copyright

Abstract

Surveys are one of the most popular, useful, and efficient methods for gathering data within both academic and organizational settings. Despite the many benefits afforded by surveys, research shows that a nontrivial number of people engage in careless responding (CR) such that their responses to surveys are not effortful, accurate or valid. This is problematic because the failure to account for CR can distort research findings, result in false theoretical conclusions and lead to precarious workplace decisions. With surveys, it is common to model responses with a latent variable framework and use fit indices to makes conclusions about how well the model represents the data. Research shows that CR can be associated with poor fit, good fit, and/or unrelated to fit. To better understand how CR affects fit, the primary goal of this study was to examine the consequences of CR on the fit of latent variable models using a comprehensive, realistic and rigorous simulation paradigm. A secondary goal was to better elucidate the nature of CR and specify how CR behavior manifests within survey responses. In Study 1, participants' survey response patterns were experimentally shaped. In Study 2, these results were used as a basis for the primary simulation. A total of 144 conditions (which varied the sample size, number of items, CR prevalence, CR severity, and CR type), two latent variable models (item response theory and confirmatory factor analysis), and six model fit indices were examined. Overall, the results of this study show that CR is frequently associated with deteriorations in model fit. These effects are, however, highly nuanced, variable, and contingent on many factors. Moreover, this study demonstrates that good fit is not necessarily indicative of careful responding nor is poor fit always emblematic of CR. Rather, model fit and CR/response validity are distinct issues that must be separately addressed. These findings can be leveraged by researchers to develop more accurate theories and practitioners to better manage survey data that is laden with CR.

Table of Contents

List of Figures

List of Tables

Acknowledgements

The number of people I would like to thank is simply too gargantuan to realistically include here. What follows is my best effort. First, I would like to thank my advisor, Christopher Lake, for his amazing support over the years. It has been such a pleasure to work with you and I am very grateful for all that you have taught me. I would also like to thank my committee members, Jin Lee (who also served as my official advisor throughout the final stages of this process), Patrick Knight, Janis Crow, and Jichul Jang, for providing me with helpful feedback about this project. I am also grateful for all the great support that Kimberly Kirkpatrick has provided me throughout my latter years in graduate school. It has been a real pleasure working with you. I would be remiss if I did not also acknowledge all the support that my cohort (i.e., Kelsey Couture, Frank Giordano, Kevin Kenney, Tiffany Lawless, Angela Rose, Stacy Stoffregen), lab members (i.e., Cassie Chlevin-Thiele and Chi-Leigh Warren), and other department friends (e.g., Jordann Brandner and Ted Moser) have provided me. I am fortunate I was able to go through this process with so many great friends and colleagues. I am also very grateful for the constant support and encouragement provided by my parents, Mark and Debbie Voss. I would also like to thank my in-laws, Brian (i.e., “Padre”) and Sharon (i.e., “Mumsie”) Wilson, as well as the “Seattle gang” (i.e., Barb Voss, Carol Voss, Kathy Voss, and Fred Dowd), for the many ways they have supported me over the years. Most crucially, however, I would like to thank my incredible wife, Bethany Voss, for her insatiable encouragement and relentless support. It is an indisputable fact that I would not have made it this far without you. Thank you for believing in me and always having my back throughout this adventure.

Dedication

In memory of my Grandpa, James Voss. An inspiring psychologist, cherished colleague, and quintessential lifelong learner. It is an honor to follow in your footsteps.

Chapter 1 - Introduction

“Data! Data! Data!... I can’t make bricks without clay!”
–Sherlock Holmes
(in *The Adventure of the Copper Beeches* by Sir Arthur Conan Doyle)

Data-related topics (e.g., big data, machine learning, artificial intelligence, people analytics, etc.) are surging in importance within all facets of society. Indeed, skills in data-based techniques and technologies have quickly become essential for employee success in numerous jobs and professions (Bradford, 2018; Davenport & Patil, 2012; Marr, 2019; Tonidandel, King, & Cortina, 2015). Furthermore, the ability to make data-driven decisions is increasingly viewed as essential for enhancing organizational effectiveness and maintaining a competitive advantage. Despite the growing excitement surrounding the use of various data-related methods and the ability of data to transform organizational practices, it is critical that researchers and practitioners first ensure that the data they use are of high quality. If data are not of high quality, researchers and practitioners are likely to make false theoretical conclusions and dangerous, ill-informed workplace decisions on the basis of such data. While many modes of data collection exist, one of the most widely used, efficient and useful ways that data can be gathered is with the use of online surveys (Evans & Mathur, 2018). Unfortunately, research indicates that the quality of survey data is frequently undermined by careless responding (CR) such that peoples’ responses to surveys are often not effortful, accurate and therefore useful (Arthur Hagen, & George, 2020; Meade & Craig, 2012). Given the prominent use of surveys within both business and academic settings, the failure to properly address CR can result in inaccurate research conclusions and precarious decisions for both practitioners and researchers alike.

While research has documented some of the effects of CR as well as methods for detecting and deterring CR (e.g., Curran, 2016; Huang, Liu, & Bowling, 2015; Ward & Meade, 2018), many important issues remain unaddressed. First, because the ability of CR indices to accurately detect CR is quite limited (e.g., Hong, Steedle, & Cheng, 2020; Goldammer, Annen, Stöckli, & Jonas, 2020), it is not always clear what CR behavior actually looks like. While CR is often defined as random responding (e.g., DeSimone, DeSimone, Harms, & Wood, 2018; Meade & Craig, 2012), this assumption is acknowledged as a limitation since people are often incapable of engaging in or recognizing true, mathematically random behavior (see Nickerson, 2002).[1] Second, few studies have thoroughly examined the effects of CR on the fit indices of latent variable models, which are regularly used to evaluate the quality of one's research findings.

The lack of research on both of these topics is concerning. First, the paucity of research on the nature of CR is detrimental to both extant and future studies of CR. If CR is not equivalent to true random responding, this would undermine a principal assumption of CR as well as the results of prior studies that have relied on this assumption. This would also inhibit future studies by causing them to adopt inaccurate conceptualizations of CR. It would seem that a better understanding of the nature of CR is needed if one is to engage in meaningful inquiries about this construct. Thus, one of the primary goals of this project is to better explicate what CR behavior actually looks like and the extent to which it is analogous to true randomness. Second, the lack of research about the relation between CR and fit is problematic because model fit is often used as a basis for justifying the usefulness of one's statistical model. Good model fit, however, does not necessarily guarantee that the responses underlying the model are valid (i.e., the respondents did

[1] While the phrase "mathematically random" can have many connotations, in the context of this study, "mathematically random" or "true random" means that survey responses follow a uniform distribution such that each response option has an equal probability of being endorsed (i.e., each response option for a 5-point, Likert-type scale has a 20% probability of being endorsed).

not engage in CR). This underscores a larger conceptual issue in that model fit and response validity are separate notions. Thus, good fitting models may contain invalid (i.e., careless) responses and poor fitting models may contain valid (i.e., careful) responses. The former situation implies that good fit is no guarantee that one's results are meaningful. This possibility is particularly egregious when researchers and practitioners assume that good model fit is indicative of high-quality data, without ever considering the impact of CR. If model-data fit is *positively* related to CR, this would make the model appear more useful than it really is, as model fit would not be taking into account the invalid responses underlying the model. On the other hand, if CR and model fit are *negatively* related, this would suggest poor fitting models could be attributable to CR, in addition to other explanations typically offered for poor fit (e.g., model misspecification). Finally, it could also be the case that CR and fit are *unrelated*. If this is the case, this would corroborate the idea that CR (i.e., validity) and fit are unique issues and should therefore be treated separately. Rigorously examining these issues is critical for developing accurate theories and ensuring that practitioners make appropriate data-driven decisions.

Based on these considerations, this study has two overarching goals: (1) To examine the consequences of CR on model-data fit for both item response theory (IRT) and confirmatory factor analysis (CFA) latent variable models across a range of conditions via a highly realistic simulation paradigm and (2) compare multiple forms of CR to determine whether or not CR is in fact equivalent to true random responding. In achieving these goals, this study will concurrently (1) explicate a method for designing more realistic CR simulation studies, (2) delineate the differences between fit and validity and note the implications of this distinction for developing more accurate theories and (3) give practitioners practical recommendations and guidance on how to better manage survey data that is laden with CR.

Defining Careless Responding

Broadly, CR refers to when participants complete surveys haphazardly without regard to item content such that their responses do not reflect their true standing on the underlying latent construct being assessed, thereby undermining the validity of the measure (Borsboom, Mellenbergh, & Van Heerden, 2004; Meade & Craig, 2012). Although there are some conceptual differences in terminology and the manner in which CR is assumed to manifest, alternative terms, such as "insufficient effort responding" (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Huang et al., 2015), "excessive inattention" (Maniaci & Rogge, 2014), "low-quality data" (DeSimone & Harms, 2018), "random responding" (Credé, 2010), "protocol invalidity" (Johnson, 2005), "content nonresponsivity" (Nichols, Greene, & Schmolck, 1989), and "response invalidity" (Edwards, 2019) have been used to describe CR. Although CR may take many forms, definitions of CR do not typically specify the form that CR may take or the underlying causes for CR (see Huang et al, 2012).

Estimates of the prevalence of CR can range anywhere from 2% to 50% of a sample (Hong, et al., 2020; Johnson, 2005; Meade & Craig, 2012), though these estimates vary widely depending on how CR is operationalized and what CR thresholds are applied (see Brühlmann, Petralito, Aeschbach, & Opwis, 2020; DeSimone & Harms, 2018; Oppenheimer, Meyvis, & Davidenko, 2009; Osborne, & Blanchard, 2011). CR, thus, probably exists in every dataset to varying degrees. It is also an issue that generalizes across different countries and/or cultures and is, therefore, a very robust phenomenon (Grau, Ebbeler, & Banse, 2019; Nichols & Edlund, 2020). It is generally believed that CR is most problematic in low-stakes settings, which are contexts where there are few personal consequences to participant's survey responses (e.g., student/Mechanical Turk research studies, organizational surveys; Arthur et al., et al. 2020). This is in contrast to high-stakes settings, in which a participant's responses are linked to a desirable outcome (e.g., getting hired by a company). In such settings, CR is less likely to be a concern, though still capable of negatively impacting data quality and the conclusions that are derived from such data. CR, therefore, should be actively addressed regardless of the context.

CR can be contrasted to other kinds of survey response behaviors. For instance, survey satisficing occurs when respondents use minimal, but still sufficient, effort to complete surveys (Krosnick, 1991, 1999). Because some effort is still being exerted by respondents employing a satisficing strategy, however, this survey response behavior is distinct from CR. If satisficing is extreme, however, this would constitute CR since respondents are no longer exerting the sufficient level of effort needed for optimal survey responding. CR is also distinct from other response sets such as faking (e.g., Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Komar, Brown, Komar, & Robie, 2008) and response styles more generally (e.g., Grau et al., 2019; Van Vaerenbergh & Thomas, 2013; Weijters, Schillewaert, & Geuens, 2008). For instance, while faking is similar to CR in that it distorts participants' true standing on a construct, it is effortful in that participants are attempting to denote a particular level of the construct being assessed. Furthermore, response styles refer to a proclivity to endorse a particular range of the scale (e.g., acquiescence or displaying a tendency to agree with items), which may or may not be due to CR. Hence, for survey responses to be considered careless, respondents must make no, or very little, effort to specify their true standing on the construct but still nonetheless provide responses to the survey. This view of CR is expounded upon in more detail in the later section integrating CR with theories of survey responding.

Empirical Consequences of Careless Responding

CR has been shown to have many negative empirical consequences. One of the most prominent ways that CR affects empirical results is by altering the effect size of the relations between variables. For instance, many studies have demonstrated that, depending on certain features of a scale (e.g., the scale midpoint), even small amounts of CR can either augment or attenuate the effect size of the relationship between substantive variables (Credé, 2010; Holden, Marjanovic, & Troister, 2019; Huang et al., 2015; Huang & DeSimone, 2020; Kam & Meyer, 2015; Maniaci & Rogge, 2014; McGonagle, Huang, & Walsh, 2016). This occurs for correlation coefficients (Huang et al., 2015), regression coefficients (Maniaci & Rogge, 2014) and when relationships are estimated within a structural equation modeling framework (Kam & Meyer, 2015). This in turn can impact the significance of the relationship between variables and consequently increase or decrease Type I/Type II error rates. Extending this line of reasoning further, this situation can result in misleading and/or completely inaccurate theoretical conclusions concerning the relations between variables. CR has been shown to produce other undesirable psychometric consequences as well. For instance, the presence of CR can also distort factor loading patterns (Kam, 2019; Kam & Meyer, 2015), inter-item correlation estimates (DeSimone et al., 2018), eigenvalues (DeSimone et al., 2018; Huang et al., 2012), and survey reliability estimates (e.g., Cronbach's alpha; DeSimone et al., 2018). Taken as a whole, CR represents a considerable threat to data quality.

Causes of Careless Responding

Given the many negative empirical consequences of CR, understanding the causes of CR is important for better understanding when CR is likely to occur, how CR can be measured, and what steps can be taken to address it. Broadly, the causes of CR tend to focus on lack of respondent motivation to respond carefully or lack of respondent obligation to respond carefully (Ward & Meade, 2018). CR has also been conceptualized to be a function of survey features and environmental factors. For instance, Meade and Craig (2012) identified the presence of environmental distractions as one potential source of CR, which is particularly relevant when data is gathered in online samples (see Cheung, Burns, Sinclair, & Sliter, 2017). Survey length also impacts CR such that longer surveys are associated with higher rates of CR towards the latter portions of the surveys (Bowling, Gibson, Houpt, & Brower, 2020; Galesic & Bosnjak, 2009; Gibson & Bowling, 2019).

More recently, however, CR has been hypothesized to be a function of the respondent's personality. For instance, CR has been found to be negatively associated with conscientiousness, agreeableness and grade point average (Bowling et al., 2016; Maniaci, & Rogge, 2014; Ward, Meade, Allred, Pappalardo, & Stoughton, 2017). Other research has demonstrated that CR is sometimes positively related to undesirable individual difference variables such as self-interest motivations (e.g., Machiavellianism; McKay, Garcia, Clapper, & Shultz, 2018). These findings are corroborated by other research showing that CR is positively related to measures of implicit aggression (DeSimone, Davison, Schoen, & Bing, 2020). CR has also been shown to be related to procrastination, though these effects are small and contingent on how CR, as well as procrastination, are measured (Voss & Vangsness, 2020). Initial longitudinal research also shows that CR displays some temporal stability across time thereby demonstrating that CR may indeed have a substantive component (Bowling et al., 2016).

Some caution is needed in deriving firm conclusions regarding the causes of CR, though. First, since many studies rely on self-report measures of personality, and because self-report data may be laden with CR, information concerning the relation between CR and the self-report measures may not be trustworthy (i.e., the relations may be spurious; Bowling et al., 2016). Some studies have attempted to circumvent this issue, however, by either using acquaintance-reported measures of personality (e.g., Bowling et al., 2016) or behavioral measures of potential antecedents (e.g., procrastination; Voss & Vangsness, 2020). Second, studies examining potential causes of CR will, in most cases, employ multiple indicators of CR, which can yield inconsistent results. Thus, if a personality variable is related to one CR index but not another, a justifiable decision must be made about whether this constitutes a meaningful relation or not. Despite some of these limitations, there is growing evidence that CR likely represents a substantive factor that is a function of a respondent's personality.

Addressing Careless Responding

Prevention Strategies

When addressing CR, two dominant approaches are to either prevent CR from occurring by implementing interventions or to screen out participants suspected of engaging in CR and then run the required statistical analyses on the cleaned dataset. The first of these approaches is based on the assumption that CR is due to lack of participant motivation (see Ward & Meade, 2018) and that CR can be reduced by increasing respondent's motivation to respond carefully. One reason that respondents may lack motivation is because there is little interaction between survey administrators and survey respondents when surveys are completed online (Johnson, 2005). This lack of interaction reduces accountability which subsequently decreases the quality of data that is provided (Ward & Pond III, 2015). Some CR interventions have attempted to counteract this by administering surveys with either virtual protectors or in-person proctors. The results of such studies suggest that these techniques are only minimally effective in reducing CR, though (Francavilla, Meade, & Young, 2019; Ward & Pond III, 2015). In contrast, other studies have attempted to reduce CR by providing warning messages to respondents instructing them to respond carefully and/or informing them there are negative consequences for CR (Gibson, 2019; Huang et al., 2012; Meade & Craig, 2012; Ward & Pond III, 2015). While these interventions are not always successful, when they are, they do not yield consistent results across different CR indices thereby raising questions as to their overall utility. Other studies have attempted to reduce CR by leveraging different social psychological theories to influence respondents' behavior (e.g., social influence theory, cognitive dissonance theory, and social exchange theory; Ward & Meade, 2018). While these theories can sometimes be used to reduce CR, these effects are not consistent across different interventions or CR indices. Thus, overall, interventions to reduce CR tend to be largely ineffectual. For this reason, a more common approach to addressing CR is to screen for participants suspected of engaging in CR after the data has been collected.

Statistical Detection Strategies

The development of indices for detecting CR is perhaps the most well-researched topic within the CR literature (for reviews see Curran, 2016; DeSimone, Harms, & DeSimone, 2015; Edwards, 2019; Meade & Craig, 2012). Some of these methods focus on embedding items into a survey prior to the administration of the survey. For instance, *infrequency approaches* provide respondents an opportunity to endorse statements that are unlikely or impossible (e.g., "I work fourteen months in a year"; Huang, Bowling, Liu, & Li, 2014). Other *directed-response methods* (e.g., "Please select 'moderately inaccurate' for this item"; Huang et al., 2012) ask respondents to endorse a specific response option. Endorsing seemingly impossible statements or incorrectly responding to a directed-response item are presumed to indicate that respondents are engaging in CR (see also Abbey & Meloy, 2017; Kung, Kwok, & Brown, 2018). Another index of CR is *response time* such that unrealistically quick response times (e.g., response time < 1 second per item) are presumed to be the results of CR (Wood, Harms, Lowman, & DeSimone, 2017). *Self-report measures* that ask participants to report their survey interest in the survey topic and diligence completing the survey have also been proposed as methods for detecting CR (Meade & Craig, 2012). Like self-report measures more generally, this approach is limited in that it assumes participants are providing valid responses. Additionally, each of these aforementioned CR indices are limited in that they must be included prior to the actual administration of the survey. Because this is not always possible, an array of post-hoc indices have been introduced.

Post-hoc CR indices are designed to identify aberrant responses that lack consistency, are too consistent, or otherwise appear to be an outlier within the dataset. For instance, both the *inter-item standard deviation* (ISD) index (Marjanovic, Holden, Struthers, Cribbie, Greenglass, 2015) and *intra-individual response variability* (IRV) index (Dunn, Heggestad, Shanock, & Theilgard, 2018) measure a respondent's standard deviation of responses across a set of items. Hence, while the typical standard deviation (*SD*) statistic is a measure of between-person variability, the ISD and IRV indices are a measure of within-person variability. Higher ISD values (i.e., greater response variability) presumably indicate greater CR while lower values indicate (i.e., greater response consistency) indicate less CR. It is possible for lower values, however, to also represent CR (Curran, 2016). Indeed, for the IRV index, low values are assumed to be indicative of CR. For situations where responses are completely invariant (i.e., the *SD* cannot be computed due to the absence of response variability), the *longstring* index, which measures the longest string of items to which respondents provided identical responses (Meade & Craig, 2012), can be used.

Other consistency indices include the even-odd consistency, person-total correlation, psychometric synonyms, psychometric antonyms, and Mahalanobis distance indices. The *even-*

odd consistency index represents the within-person correlation between even and odd scale halves (Johnson, 2005) whereby lower values indicate lower response consistency (i.e., higher CR). The *person-total correlation* index measures how consistently a respondent responds to a set of items relative to everyone else who responded to the same set of items (Curran, 2016; see also Donlon & Fischer, 1968) such that lower values indicate that the respondent is not responding similar to the rest of the sample and may, therefore, be engaging in CR. The *psychometric synonyms* index represents the within-person correlation across positively correlated item pairs (Johnson, 2005; Meade & Craig, 2012). Like the person-total correlation index, the psychometric synonyms index is a correlational method that first involves computing a correlation matrix for a set of scale items. The highest correlated item pairs within a sample are then identified. The within-person correlation across the item pairs that were identified is then calculated for each respondent. This method assumes that if item-pairs are highly correlated at the sample level, they should also be highly correlated at the person level. High psychometric synonyms values thus indicate more careful responding while lower values indicate greater CR. The *psychometric antonym* index is very similar except that it is computed with negatively correlated item pairs. Finally, the *Mahalanobis distance* index is a multivariate outlier detection method that represents the distance between a person's responses to a set of items and the sample's mean responses to the same set of items (Mahalanobis, 1936).

In addition to these more "traditional" methods, a variety of IRT-based person-fit indices may also be useful for detecting CR (for reviews see Karabatsos, 2003; Meijer & Sijtsma, 2001; see also Beck, Albano, & Smith, 2019; Liu, Lan, & Xin, 2019; Nye, Joo, Zhang, & Stark, 2019; Patton, Cheng, Hong, & Diao, 2019). Some of the most common IRT person-fit indices that have been examined for detecting CR include standardized log-likelihood (l_z), Guttman errors (G), and the *U3* index. The standardized loglikelihood index (l_z) is a parametric IRT person-fit index that measures the likelihood of a respondent's response pattern given their standing on the underlying construct (Drasgow, Levine, & Williams, 1985). Guttman errors represent the number of instances where someone's response pattern is incongruent with the ordering of item steps (Guttman, 1944; Curran; 2016; Niessen, Meijer, & Tendeiro, 2016). This occurs when a respondent answers a difficult item correctly but an easy item incorrectly (described in more detail in Study 2). Lastly, *U3* represents the ratio of the log-odds of item steps passed (i.e., the log-odds proportion of items answered "correctly"). While these person-fit indices are computed in unique ways, they are each similar in that they attempt to identify aberrant response patterns within a broader IRT framework.

Research on the detection of CR has predominantly focused on comparing the utility of multiple CR indices (Beck, et al., 2019; DeSimone & Harms, 2018; Dupuis, Meier, & Cuneo, 2019; Goldammer et al., 2020; Hong et al., 2020; Huang et al., 2012; Niessen et al., 2016; Meade & Craig, 2012; Steedle, Hong, & Cheng, 2019). This is often accomplished by comparing the sensitivity/specificity of the indices as well as their Type I and Type II classification rates under various circumstances. Several broad conclusions can be derived from this research. First, there is no single index can be used to sufficiently detect the many different forms of CR. Instead, researchers must utilize multiple indices to detect CR and its various modes of expression (e.g., random responding vs. invariant responding). Second, CR indices do not have consistently high convergent validity in that CR indices are only minimally correlated. This at best suggests that different CR indices assess unique components of the CR construct and at worst that certain indices fail to adequately measure CR. Third, the empirical and psychometric impact of CR is a function of which indices are used to screen for CR as well as the threshold that is used to define what constitutes CR. That is, different CR indices, and different data screening thresholds, will identify different individuals as careless responders. Fourth, the impact of CR is also determined by features of the scale (e.g., number of factors) and the sample (e.g., sample size). Thus, knowledge of how to optimally identify and address CR requires a holistic consideration of the many aspects of the overall survey context. While statistical methods are the most common way to address CR, they are not without limitations. For instance, flagging respondents suspected of engaging in CR may incorrectly flag careful responders (Type I error) or fail to flag true careless responders (Type II error). Such post-hoc screening strategies also reduce the size of the final sample and, correspondingly, statistical power.

A Theory of Survey Response

While research on CR has been largely atheoretical (Bowling & Huang, 2018), CR can be better understood by integrating it with extant theories of survey response. One prominent theory of survey responding is Tourangeau, Rips, and Rasinski's (2000) cognitive theory (see also Tourangeau, 2018). According to this theory, survey respondents progress through a series of cognitive steps when responding to survey items. Specifically, respondents first comprehend the question (Step 1), generate a retrieval strategy and extract the necessary information (Step 2), map their judgment onto the response category based on the information retrieved from the previous step (Step 3), and then finally provide their response to the item (Step 4; see also Karabenick et al., 2007; Shulruf, Hattie, & Dixon, 2008). When these processes are each fully enacted, respondents are employing an optimizing strategy (Krosnick, 1991, 1999). If these steps are enacted in a suboptimal manner, respondents are employing a satisficing strategy (Barge & Gehlbach, 2012; Krosnick, 1999). This can entail either "weak satisficing", whereby participants go through each cognitive stage somewhat haphazardly, or "strong satisficing", whereby latter steps of the cognitive process are skipped entirely (Krosnick, 1999; Tourangeau, 2018). In each instance (i.e., optimizing, weak satisficing, and strong satisficing), respondents are providing valid responses in that responses are presumably being caused by the latent construct underlying the survey and therefore accurately denote the respondent's position on the construct being assessed (Borsboom et al., 2004). Of course, while satisficing can yield somewhat valid responses, response validity is highest when all cognitive steps are enacted.

When someone engages in CR, however, response validity is low or entirely absent. A conceptual integration of CR, Tourangeau et al.'s (2000) cognitive theory of survey response, survey satisficing, and response validity is presented in Figure 1. CR manifests when a sufficient number of cognitive steps are bypassed but respondents still nonetheless provide responses to the survey (e.g., Brower, 2020). Within the context of Tourangeau et al.'s (2000) theory, this may entail bypassing the comprehension step (Step 1) entirely. This is consistent with the view that CR occurs when respondents complete surveys without regard to item content (Meade & Craig, 2012). In such cases, responses likely maintain no degree of validity since they are not being caused by the underlying construct. It may also be possible for respondents to tentatively enact Step 1, but then for various reasons (e.g., boredom, distractions, disinterest, etc.) exit this step early and respond carelessly. Such responses would possess very low validity since the cognitive steps were insufficiently enacted so as to produce a minimally valid response. These two possibilities are illustrated by the bottom (dashed) arrows in Figure 1. Further support for these considerations can be derived from studies showing that careless responders often respond to surveys quicker than careful responders (e.g., Wood et al., 2017). Presumably, bypassing the cognitive processes needed to enable an optimal response decreases the amount of time that it takes to respond to survey items. As noted previously, respondents may proceed through all cognitive stages in a suboptimal manner or skip the latter stages altogether, which according to Krosnick (1999), would still represent a satisficing strategy as responses maintain some degree of validity. Responses are most valid, however, when respondents proceed through all cognitive steps in an optimal manner (as indicated in the arrows in the upper portion of Figure 1 connecting the four steps). Like various definitions of CR that have been proposed (e.g., Huang et al., 2012), this view of CR does not specify the form that CR may take nor the underlying causes of CR.

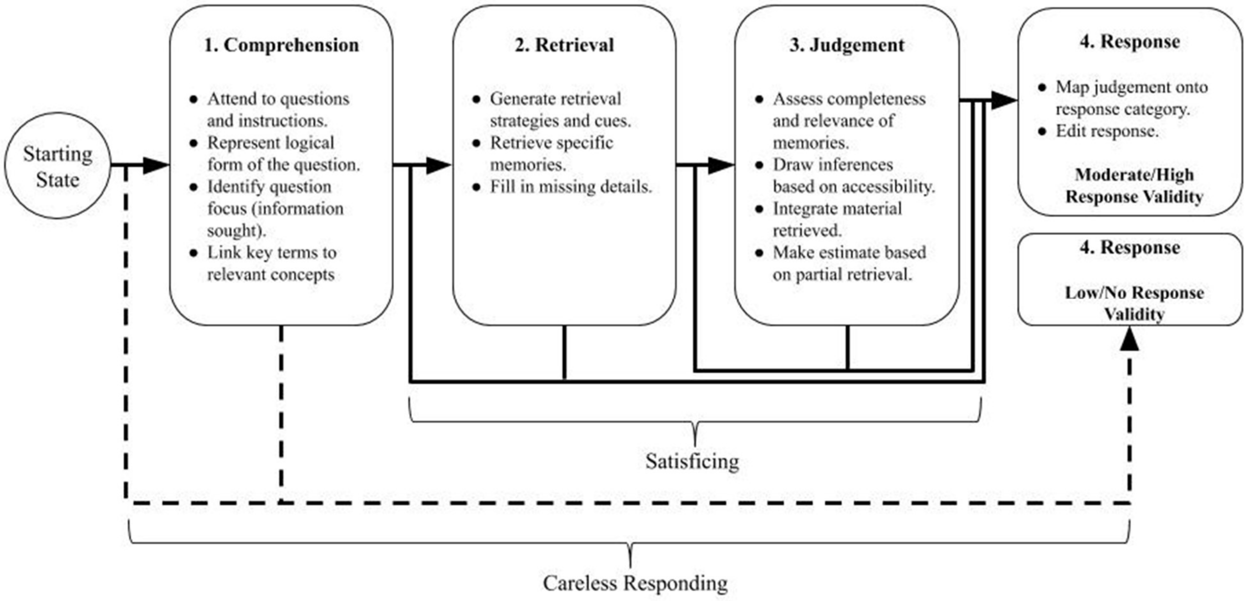

Starting State
1. Comprehension
• Attend to questions and instructions.
• Represent logical form of the question.
• Identify question focus (information sought).
• Link key terms to relevant concepts
2. Retrieval
• Generate retrieval strategies and cues.
• Retrieve specific memories.
• Fill in missing details.
3. Judgement
• Assess completeness and relevance of memories.
• Draw inferences based on accessibility.
• Integrate material retrieved.
• Make estimate based on partial retrieval.
4. Response
• Map judgement onto response category.
• Edit response.
Moderate/High Response Validity
4. Response
Low/No Response Validity
Satisficing
Careless Responding

Figure 1. Conceptual model integrating careless responding with survey response process

While this study will not explicitly test this model (but see Brower, 2020), this conceptual integration is nonetheless useful for better understanding the nature of CR. For instance, this model explicitly demonstrates how CR may function, shows where CR fits into the overall survey response process, and explains how different response behaviors may differentially impact response validity. This conceptual integration thus provides a more enhanced perspective of CR than is normally provided. For instance, many studies frequently rely on very vague definitions of CR, which can create confusion regarding the response behavior(s) that actually constitute CR. Because CR does not occur in isolation, but within the broader process of survey responding, knowledge of how CR is related to other important theoretical notions, such as response validity and survey satisficing, is important. This is especially true from a theoretical standpoint as these considerations imply that one of the primary negative consequences of CR is that it undermines various assumptions of latent variable theory.

Latent Variable Models

Latent Variable Theory

The concept of latent variables underlies most of modern psychological measurement (Bollen, 2002; Edwards, & Bagozzi, 2000; Markus & Borsboom, 2013). For example, there is frequently a distinction between psychological constructs and the methods by which constructs are measured (e.g., surveys; Binning & Barrett, 1989). When constructs are conflated with their mode(s) of measurement, or vice versa, inaccurate theoretical and practical conclusions follow (Arthur & Villado, 2008). Although different conceptualizations of latent variable models exist, latent variable models always presume that there exists a latent (i.e., unobservable) psychological attribute (i.e., theoretical entity) and this attribute serves as a causal determinant for a set of observed indicators (i.e., survey items; Borsboom, Mellenbergh, & Van Heerden, 2003). Because psychological attributes and their structure cannot be directly observed, their nature must be inferred via measurement. This typical framework is graphically illustrated in Figure 2 where a single latent variable is presumed to cause the variation in five indicator variables. For a set of indicators to denote something meaningful about a psychological attribute, the indicators must relate to the latent variable in a lawful manner (Markus & Borsboom, 2013). Typically, the relationship between latent variables and their indicators can be modeled as a generalized regression function (see Borsboom, 2008; Markus & Borsboom, 2013) such as, $E(x \mid \theta) = a\theta + \varepsilon$ where x represents an observed variable, θ represents the latent variable, a represents a standard regression parameter, and ε represents error. Applying this equation to a hypothetical conscientiousness scale, a respondent will provide a response (x) to each item based primarily on her conscientiousness level (participant's θ) and the properties of the item. Furthermore, such responses are assumed to be imperfectly measured and will therefore contain error (ε). Another common assumption of latent variable theory is that indicators (i.e., items) display local independence such that any relation between indicators only exists through their common relation to the underlying latent variable (Bollen, 2002).

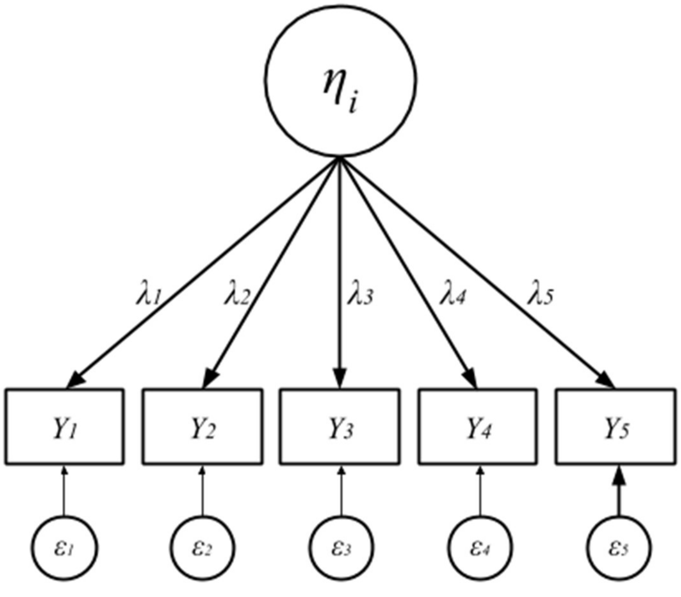


Figure 2. Visual representation of a reflective unidimensional latent variable model

When psychological data is gathered via surveys and analyzed, it is usually assumed the data can be modeled in accordance with the principles of latent variable theory (but see Edwards,

& Bagozzi, 2000; Markus & Borsboom, 2013). That is, the observed responses to a survey are presumed to denote something meaningful about the unobserved psychological construct. To model data in this manner, two common psychometric approaches are confirmatory factor analysis (CFA) and item response theory (IRT). Both CFA and IRT represent classes of statistical models that, while being somewhat different in their underlying computations, assumptions, and terminology, attempt to model the relationship between responses to a set of indicators and the latent construct underlying the indicators (referred to as theta [θ] within IRT frameworks]). When there is a correspondence between the data and statistical model, there is justification for regarding the indictors as measuring the latent, theoretical construct (Borsboom, 2008; Ropovik, 2015). Regardless of whether CFA or IRT is used to make this determination, decisions regarding the theoretical appropriateness of the model and data-model congruence are typically made with the use of model fit indices.

Model Fit and Careless Responding

Broadly, model-data fit concerns whether there exists a discrepancy between the observed values in a dataset and those values predicted by the statistical model (Greiff & Heene, 2017; Kline, 2016). A model with good fit implies that the statistical model accurately represents the data in question. Good model-data fit is, therefore, a necessary, but not sufficient component of a high-quality latent variable model. It is necessary because the absence of model-data fit suggests that a particular latent variable model does not accurately represent the given data. Demonstrating good model-data fit is not necessarily sufficient, however, because there are other aspects of latent variable model quality to consider in addition to fit (e.g., whether a different, model can better account for the data, the quality of the data underlying the model, etc.).

Regardless, assessing model-data fit is an important step in determining the accuracy by which a latent variable model can represent a given dataset.

Global fit statistics, such as χ^2 which is used often used for CFA models and the somewhat analogous *M2* used for IRT models, are used to specify the overall fit of an entire model. Fit can also be specified at a local level (i.e., item-level), which is particularly common within IRT frameworks. Fit indices can be further classified as absolute fit indices, which assess how well the specified model fits the sample data, or incremental fit indices, which usually compare the specified model to a null (H_o) model. Many different indices to assess both absolute fit (e.g., χ^2, *M2,* root mean square error of approximation [*RMSEA*] and standardized root mean square residual [*SRMSR*]) and incremental fit (e.g., Comparative fit index [*CFI*], Tucker-Lewis index [*TLI*]) have been proposed for both CFA and IRT models (Kline, 2016; Nye et al., 2019). Poor fitting models may be indicative of misspecification and are often rejected as incorrect since the relation between the indicators and latent construct cannot be justified given the data at hand. To make decisions about the quality of latent variable models, various cutoffs for commonly reported fit indices have been proposed (e.g., *RMSEA* < .08, CFI > .95; Hu & Bentler, 1999). Although useful heuristics, the use of cutoffs does not guarantee that the model is a good representation of the data (Nye & Drasgow, 2011). Nonetheless, model fit indices remain one of the primary methods for determining the quality and validity of latent variable models, and this has important theoretical and practical implications.

For example, the conclusions derived from model fit indices serve as an important basis for theoretical advancement. If a model displays poor fit, this may suggest that the relation between indicators and the latent construct is mis-specified and in need of revision. This may entail developing new items, refining the definition of the construct, or reassessing the dimensional structure of the psychological attribute, to name a few examples. From a practical perspective, poor model fit can have additional negative consequences. For example, if a survey is used to inform organizational policies, and the statistical model underlying the survey displays poor fit, any decisions made on the basis of the survey results could be misleading since the construct is not being properly assessed. It could also be the case that an entirely different construct is being assessed which will again lead to misinformed and unjustifiable decisions.

Initial research on the relationship between CR and model fit suggests that CR impacts model fit in inconsistent ways. For instance, some research has suggested that when careless responders are screened and removed from a dataset, the fit of estimated models improves (Arias, Garrido, Jenaro, Martínez-Molina, & Arias, 2020; Huang et al., 2012). Within simulation contexts, studies have also shown that the presence of CR deteriorates model fit under certain circumstances (Woods, 2006). Other studies, however, have found that the presence of CR is unrelated to model fit (Beck et al., 2019; Kam & Meyer, 2015; Kam, 2019; Schneider, May, & Stone, 2018; Steedle et al., 2019), even when up to half of the sample engages in CR (Liu et al., 2019). In contrast, other research has found that when careless responders are compared to careful responders, model fit can be counterintuitively better for the careless responders (Goldammer et al., 2020). Unfortunately, it is difficult to derive firm conclusions about the effects of CR on model fit since existing research relies on different definitions of CR, indices for measuring CR, research designs (e.g., experimental studies, simulations, etc.), and fit indices. Furthermore, features of these studies such as sample size, the measures used, and the number of respondents engaging in CR, are likely to vary from one study to the next. Thus, it remains unclear how CR affects model fit across different circumstances. Because a nontrivial number of respondents engage in CR and because model fit information is essential for ensuring high-quality data and making meaningful theoretical and practical advancements, knowledge of the relation between CR and model fit is essential. To study this issue in greater detail and with greater precision, Monte Carlo simulation studies are especially useful.

Monte Carlo Simulation

Monte Carlo simulation studies are well suited for situations where the true properties of participants and measures are difficult to know – which occurs frequently within naturalistic settings/studies – and the use of certain methods or analyses are impractical due to their complexity (e.g., manipulating many experimental conditions; Feinberg & Rubright, 2016; Harwell, Stone, Hsu, & Kirisci, 1996). For instance, to fully understand how CR impacts model fit would require conducting numerous studies with large sample sizes, each building on the other and examining hundreds of possible different scenarios. It would also require knowledge of who is truly engaging in CR and who is not truly engaging in CR, which is something that can never be known with certainty. Given that CR indices are not always able to accurately differentiate between careless and careful responders, studies that examine this issue with actual participants are unlikely to ever reach valid conclusions. This is not an issue with simulation studies, however, as the respondents that are engaging in CR can be specified a priori and therefore known with complete certainty.

Simulation studies are also advantageous for examining many scenarios simultaneously since each condition can be treated as a unique sample that can be replicated *n* times. This can be useful for avoiding potential research confounds such as sampling error. Additionally, to fully understand the relation between CR and model fit would require manipulating many variables in an extremely complex manner. For instance, this might entail ensuring that a certain proportion of the sample engages in CR, stipulating how much CR must occur per respondent, and specifying the type of CR to be exhibited. It is unlikely this incredibly detailed level of precision could be achieved within a naturistic research setting that utilizes actual participants. With a simulation study, however, complex manipulations such as this can be made with ease.

Consequently, in order to derive an accurate understanding of how CR and model fit are related, a simulation study will be conducted due to the control and precision that is afforded by this type of research design. In order to maximize the utility of this type of research design, various steps will be taken to ensure a high degree of realism. For example, as described below, CR will be modeled by adapting response patterns derived from real research participants. Attaining this high level of realism is essential for describing situations that are actually encountered and ensuring the results of the simulation are useful and generalizable (see Bulut & Sünbül, 2017; Harwell et al., 1996).

Study Overview and Research Questions

There are two primary goals of this simulation study: (1) To examine the consequences of CR on model-data fit for both IRT and CFA latent variable models across a range of conditions via a highly realistic simulation and (2) compare different forms of CR to determine if CR is in fact equivalent to true random responding. This will concurrently determine if different forms of CR are differentially related to model-data fit for IRT and CFA latent variable models. In achieving these goals, this study will also (1) explicate a method for designing more realistic CR simulations, (2) discuss the implications of CR for developing accurate theories in terms of accurately specifying relations between constructs and (3) provide practitioners with practical recommendations on how to better manage low-quality survey data.

As noted above, a major limitation of research on CR is that CR indices do not always accurately detect CR (see Hong et al., 2020; Goldammer et al., 2020; Niessen et al., 2016). Put another way, CR indices display low sensitivity and often yield Type I or Type II errors. Additionally, the relations between different CR indices are often low. Thus, a respondent that is classified as careless with one index does not guarantee they will be classified as careless by another index. Given these issues, it is not always clear what true CR behavior looks like. Although many studies assume that CR is analogous to true, mathematically random responding, such studies often view this assumption as tenuous (see DeSimone et al., 2018; Meade and Craig, 2012) since research indicates that people are incapable of detecting and engaging in true, mathematically random behavior (Kahneman & Tversky, 1972; Mosimann, Wiseman, & Edelman, 1995; Nickerson, 2002; Niessen et al., 2016; Rabin & Vayanos, 2010). Thus, it remains unclear the extent to random responding and CR are analogous. Another related limitation regarding previous CR simulation studies is that they cannot be informed by existing datasets since, due to the aforementioned issues regarding the detection of CR, it is not known who is actually engaging in CR within such datasets (DeSimone et al., 2018). A consequence of this is that researchers often make assumptions about the nature of CR that are not always grounded in real survey response behaviors (e.g., the assumption that CR is analogous to random responding). Because prior research often operationalizes CR as random responding (e.g., DeSimone et al., 2018; Huang et al., 2015; Meade & Craig, 2012), determining the extent to which CR and random responding are equivalent is critical. This is important for verifying the utility and generalizability of prior simulations and informing future studies on the nature of CR.

To rigorously examine this issue, the following simulation study will be informed by an initial study (Study 1) where real respondents' survey behaviors are shaped by experimentally manipulated stimuli. Hence, it will be possible to compare real participant's response behaviors to that of true mathematically random responses. To further add realism to this study, data for the simulation will be derived via a rigorous Monte Carlo IRT simulation procedure (described in detail in Studies 1 and 2 below). Taking these steps will ensure that all the data and parameters that are used to inform the simulation study are based on participants' actual survey behavior and not hypothetical data, parameters, and/or survey responses (Harwell et al., 1996). As noted previously, maintaining realism is useful for describing situations that researchers and practitioners actually encounter and ensuring that the simulation will have a high degree of utility. Formally, this project will answer the following research questions:

Research Question 1: What are the consequences of CR on the fit of both CFA and IRT latent variable models across a range of different conditions?

Research Question 2: To what extend are CR and mathematically random responding analogous response behaviors?

Chapter 2 - Study 1

The main goal of Study 1 was to shape respondents' survey patterns with experimentally manipulated stimuli and use this information (i.e., the estimated parameters and item responses) to inform the primary simulation study (Study 2). To accomplish this, participants were assigned to one of four possible conditions that contained instructions to engage in a particular response behavior. This entailed either a careful responding, careless responding, random responding, or a control condition. It was expected that participants in the careful condition would (1) respond more accurately, (2) provide more valid responses (i.e., responses would be caused by the underlying construct being assessed [conscientiousness]), and (3) engage in less CR within the survey compared to those in the careless and random conditions. No specific expectations existed regarding whether or not the careless and random conditions would differ, though separate conditions were included for both these response behaviors in order to examine this possibility. This approach, as well as the results of this study, are described in detail below.

Method

Participants

A total of 1,346 participants from Amazon's Mechanical Turk (MTurk) platform, which research indicates is a source of high-quality research data, (Buhrmester, Kwang, & Gosling, 2011), completed the Study 1 survey and were compensated $0.75. Approximately 300 participants per condition were desired to properly conduct the analyses needed to inform the simulation study. Because it was anticipated that some participants would not adhere to the study instructions, the number of participants recruited for each condition was slightly oversampled (data cleaning procedures described in detail below). Of these participants, 52% identified as men/male, 73% identified as white/Caucasian, 56% had a bachelor's degree or higher, 70% were employed full-time, and their average age was 37 (*SD* = 11.26).

Design and Procedures

Participants were randomly assigned to one of four conditions: (1) A careful responding condition, (2) careless responding condition, (3) random responding condition, or (4) control condition. Each of these experimental conditions contained instructions to respond either carefully, carelessly, or randomly. The control condition did not contain any instructions to respond in a particular manner. Similar to previous research examining strategies for reducing CR (e.g., Huang et al., 2012; Ward & Pond III, 2015), participants in the careful condition were asked to respond as accurately as possible and also warned that sophisticated statistical methods would be used to flag low-quality data. Participants were also reminded of the importance of having high-quality data for this research. Prior to proceeding with the survey, participants were also asked to check a box that stated, "I agree to complete the following survey as accurately as possible". Next, similar to previous research asking participants to elicit particular response behaviors (e.g., Zickar, Gibby, & Robie, 2004), participants in the careless and random conditions were instructed to respond as carelessly and randomly as possible, respectively (e.g., "Please respond to these statements as carelessly [randomly] as possible…"). Separate conditions were included for careless and random responding to clarify if carelessness and randomness would result in unique response patterns and thus whether they should be considered separately for the simulation. A summary of these instructional manipulations can be seen in Table 1.

Table 1. Instructions used within the four Study 1 conditions

Condition	Instructions
Control	You will now be presented with a list of statements that may or may not describe you. Please indicate for each statement how accurately it describes you.
Careful	You will now be presented with a list of statements that may or may not describe you. Please indicate for each statement how accurately it describes you. To ensure the quality of survey data, your responses will be subject to sophisticated statistical control methods. **Responding without much effort will be flagged for low-quality data.** Having **accurate information is very important for this research**. We really **appreciate your thoughtful responses!** If you agree to complete the following survey as accurately as possible, please indicate this by selecting the option from the dropdown menu below and proceeding with the survey!
Careless	You will now be presented with a list of statements that may or may not describe you. We ask that you **respond to these statements as carelessly as possible**. This **includes any attention-check items** that may be included. To be clear, we are not interested in how accurately these statements may or may not describe you. Please **provide us with very low-quality responses**. Please be assured, although we are asking you to respond carelessly, we will still approve this HIT assignment when you complete it. Thanks!
Random	You will now be presented with a list of statements that may or may not describe you. We ask that you **respond to these statements as randomly as possible**. This **includes any attention-check items** that may be included. To be clear, we are not interested in how accurately these statements may or may not describe you. Please **provide us with very low-quality responses**. Please be assured, although we are asking you to respond randomly, we will still approve this HIT assignment when you complete it. Thanks!

Note. Participants were randomly assigned to one of these four conditions. Statements were bolded as shown in order to direct participants' attention to the most essential part of the text and ensure successful manipulations.

For the primary survey, 60 conscientiousness items from the 300-item NEO International Item Personality Pool (IPIP; Goldberg et al., 2006) were used (see Appendix A for the list of these items). This measure contains six separate facets (self-efficacy, orderliness, dutifulness, achievement, self-discipline, and cautiousness) that comprise the conscientiousness construct. A 5-point, Likert-type format ranging from 1 (*Very Inaccurate*) to 5 (*Very Accurate*) was employed for all items. Two directed-response items (i.e., "Please select 'Very Accurate' for this item" and "Please select 'Moderately Inaccurate' for this item") were embedded into the survey in the same location for the four conditions. These directed-response items served as a way to assess CR and thus clarify if participants in the careful condition engaged in less CR than those in the careless and random conditions. After completing the main survey, all participants saw a separate screen where they were instructed to respond to all subsequent questions as accurately as possible. Specifically, this transition page stated, “Thank you for providing your responses to these statements! You will now be asked a few additional questions. We ask that you respond to these following questions as carefully as possible.”

Participants then completed two manipulation-check items. As noted above, it was expected that participants in the careful condition would respond more accurately, provide more valid responses, and engage in less CR when completing the survey compared to those in the careless and random conditions. No specific expectations existed for the control condition. A summary of these expectations can be seen in Table 2. For the first manipulation check item, participants were asked to indicate how accurately they completed the previous survey. Specifically, participants were asked, “How accurately did you respond to the previous statements (i.e., “to what extent were your responses accurate descriptions of you?)”. For the second item, participants saw a brief definition of conscientiousness and were then asked to indicate their level of conscientiousness based on the definition provided. The definition of conscientiousness provided to participants stated, “People with high levels of conscientiousness (conscientious people) tend to be very dependable, organized, and self-disciplined. Based on this definition of conscientiousness, please indicate how accurate the following statement is about yourself.” The statement that participants viewed was, “I am a conscientious person.” Responses to both these items were made on a sliding bar ranging from 0 (“Very Inaccurately” for manipulation check item 1 and “Very Inaccurate” for manipulation check item 2) to 100 (“Very Accurately” for manipulation check item 1 and “Very Accurate” for manipulation check item 2). This alternative format (i.e., the sliding bar) was used in order to differentiate these items from the previous survey and help facilitate accurate responses. Based on the above expectations regarding participants' response behaviors within each of these conditions, if participants in the careless and random conditions truly responded in a way that diverges from their true construct level, the participants in the carless and random conditions should provide lower accuracy ratings than those in the careful condition. Additionally, if the responses made by those in the careful condition are more valid, there should be a stronger correlation between the two measures of conscientiousness (i.e., the 60-item IPIP-NEO and the sliding bar) for those in the careful condition compared to those in either the careless or random conditions. Finally, it was expected that participants in the careful condition would answer fewer directed response items incorrectly than participants in either the careless or random condition (see Table 2).

Table 2. Summary of expected results for the manipulation checks

Condition	Expected Mean Accuracy Ratings	Expected Relation Between Conscientiousness Scores	Expected Responses to Directed-Response Items
Control	No expectations	No expectations	No expectations
Careful	High accuracy ratings	Positive relation	Accurate responses
Careless	Low accuracy ratings	No relation	Inaccurate responses
Random	Low accuracy ratings	No relation	Inaccurate responses

Note. The relation between conscientiousness scores refers to the expected relation between the scale-derived conscientiousness score (NEO-IPIP) and the sliding bar-derived conscientiousness score.

Item Response Theory Overview

In order to derive a set of realistic parameters that can be used to inform the subsequent simulation study, item response theory (IRT) was used to analyze the data from Study 1. This is necessary since the simulation study will employ an IRT-based Monte Carlo response generation procedure. IRT represents a class of statistical models that attempt to describe the relationship between a person's responses to a set of items and the latent construct (i.e., theta [θ]) that underlies the items. Specifically, IRT attempts to estimate a person's θ level given their responses to a set of items as well as the item properties (Embretson & Reise, 2000). Because the following study used a polytomous rating format and item response options are assumed to be ordinal and monotonic, the graded response model (GRM; Samejima, 1969), which is well-suited for assessing Likert-type data, is used. Formally, the GRM is defined as:

$$P_{ix}^{*}(\theta_j) = \frac{exp[\alpha_i(\theta_j - b_{ij})]}{1 + exp[\alpha_i(\theta_j - b_{ij})]} \tag{1}$$

Here, $P_{ix}^{*}(\theta_j)$ represents the probability of a person with a given theta (θ) level endorsing a response option within a given option boundary where *x* is the number of response options. This is sometimes defined as an "operating characteristic curve" (Embretson & Reise, 2000). In this equation, α_i denotes the item slope (or discrimination) parameter for item *i* and b_i denotes the category threshold parameters (or item difficulty) for item *i*. The α_i parameter represents the ability of an item to differentiate people at different levels of θ while the b_i parameter represents the point on the θ continuum where there is a 50% chance of selecting a given response option *or* the next-highest response option. There are always one fewer b_i parameters than there are response options. For example, a 5-point Likert scale would have the following category threshold parameters: b_1, b_2, b_3, and b_4. Formally, the probability of responding for each response option category, which is often referred to as a category response curve (CRC), is defined as:

$$P_{ix}(\theta_j) = P_{ix}^{*}(\theta_j) - P_{i(x+1)}^{*}(\theta_j) \tag{2}$$

For a 5-point, Likert-type scale (which is used within both Study 1 and Study 2), this can be further expressed for each response option as:

$$P_{i1}(\theta_j) = 1 - P_{i1}^{*}(\theta_j) \tag{3}$$

$$P_{i2}(\theta_j) = P_{i1}^{*} - P_{i2}^{*}(\theta_j) \tag{4}$$

$$P_{i3}(\theta_j) = P_{i2}^{*} - P_{i3}^{*}(\theta_j) \tag{5}$$

$$P_{i4}(\theta_j) = P_{i3}^{*} - P_{i4}^{*}(\theta_j) \tag{6}$$

$$P_{i5}(\theta_j) = P_{i4}^{*} - 0 \tag{7}$$

Again, the CRCs for the GRM represent the relation between θ and the probability of a person responding to each possible response option (see Embretson & Reise, 2000 for a more detailed overview of IRT).

Results

Manipulation Checks

To determine if the manipulations were successful and respondents' survey behaviors were shaped in the intended manner, the mean response accuracy ratings, correlation between the two measures of conscientiousness, and percentage of directed response items answered incorrectly were calculated for each condition. These results are summarized in Table 3. First, to compare accuracy ratings across all four conditions, a One-Way ANOVA was conducted with "Condition" serving as the independent variable and "Accuracy" serving as the dependent variable. The result of this analysis was significant, *F*(3, 1341) = 367.12, *p* < .001, indicating that there were differences in accuracy ratings across the four groups. To follow up on this, a Tukey HSD post-hoc test was conducted. This analysis revealed that participants in the careful condition had higher accuracy ratings (*M* = 96.04) than those in either the careless (*M* = 39.66, *p*

< .01) or random (*M* = 39.23, *p* <.01) conditions. The results also indicated that those in the control condition had higher accuracy ratings (*M* = 94.73) than those in the careless (*M* = 39.66, *p* < .01) or random (*M* = 39.23, *p* <.01) conditions. There were no differences in accuracy ratings when comparing the (1) control and careful or (2) careless and random groups, though (see Table 3).

Table 3. Summary of results for the manipulation checks

Condition	Condition *n*	Mean Accuracy Ratings	Conscientiousness Correlation	Directed-Response Items Incorrect (%)	
				Item 1	Item 2
Before Cleaning					
Control	342	94.73_a	.57**$_a$	2%	5%
Careful	335	96.04_a	.58**$_a$	2%	5%
Careless	336	39.66_b	.11*$_b$	41%	40%
Random	330	39.23_b	.21**$_b$	40%	39%
After Cleaning					
Control	284	97.68_a	.62**$_a$	-	-
Careful	287	98.11_a	.68**$_a$	-	-
Careless	274	26.06_b	$-.08_b$	49%	48%
Random	273	26.77_b	$.05_b$	46%	46%

Note. **p* < .05, ***p* < .01. Accuracy ratings were made on a sliding bar ranging from 0 to 100. The conscientiousness correlation refers to the correlation between participant's conscientious scores on the main survey and their conscientiousness scores on the post-survey sliding bar. Within each column, values not connected with the same subscript are significantly different. For the mean accuracy ratings, this was determined with a One-Way ANOVA Tukey HSD post-hoc test. For the conscientiousness correlations, this was determined with a Fisher *r*-to-*z* test.

Next, the correlation between the NEO-IPIP conscientiousness measure and the single item conscientiousness sliding bar was computed for all four conditions. Overall, the correlations were large for both the careful (*r* = .58, *p* < .01) and control (*r* = .57, *p* < .01) conditions. In contrast, the correlations were small for the careless (*r* = .11, *p* < .05) and random (*r* = .21, *p* < .01) conditions. These correlation coefficients were next formally compared using a series of Fisher *r*-to-*z* tests. The results of this analysis indicated that the correlation for the careful condition was significantly different than the correlation for the careless condition (*Z* = 7.12, *p* < .001) and the random condition (*Z* = 5.77, *p* < .001). Likewise, the correlation for control condition was significantly different than the correlation for the careless condition (*Z* = 6.69, *p* < .001) and the random condition (*Z* = 5.60, *p* < .001). There were no significant differences between the (1) control and careful or (2) careless and random conditions, though.

Finally, the number of directed-response items answered incorrectly was computed for each condition. The results indicated that those in the careful and control conditions answered the same number of directed-response items incorrectly (Item 1 = 2%; Item 2 = 5%). Those in the careless and random conditions answered far more of these directed-response items incorrectly (Careless: Item 1 = 41%, Item 2 = 40%; Random: Item 1 = 40%; Item 2 = 39%).

While the results of these analyses provide assurance that respondents generally adhered to the survey instructions, this still might not have been the case for all participants (e.g., someone in the careless condition may have responded carefully and vice-versa). Accordingly, the data were subject to additional cleaning to help ensure that the data in each condition is indicative of the response pattern that it is meant to reflect. This is particularly important since the IRT parameter values for these conditions will be used to inform the parameters for the subsequent simulation study. Accordingly, participants who answered any of the directed-response items incorrectly or provided an accuracy rating ≤ 85, were omitted from the analyses in both the control and careful conditions. Furthermore, participants with an accuracy rating of 100 were omitted from the careless and random conditions. These cleaning criteria values were selected in order to eliminate those who clearly did not adhere to the instructions while also ensuring that the sample size would not be substantially decreased. For example, a considerable number of participants in the careless (*n* = 62) and random conditions (*n* = 57) indicated that they completed the survey very accurately (i.e., selected “100” on the sliding bar) which suggests they did not adhere to the survey instructions. Using the same criteria as the careful condition (i.e., accuracy ratings ≤ 85) to screen participants in the careless and random conditions would have eliminated too many participants and made it impossible to conduct the necessary IRT analyses. Hence, slightly different criteria were used to screen across the careful/control and carless/random conditions. Despite the use of slightly different criteria, this screening method resulted in groups that were roughly equivalent in size (see Table 3).

The results of a One-Way ANOVA again indicated that there was a significant difference in accuracy ratings across the conditions, $F(3, 1105) = 744.37$, $p < .001$. A Tukey HSD post-hoc test comparing accuracy ratings across the four conditions indicated that the pattern of results was exactly the same as that reported prior to data cleaning (see Table 3). Additionally, the conscientiousness correlations were again large for both the careful ($r = .68$, $p < .01$) and control ($r = .62$, $p < .01$) conditions. In contrast, the correlations were small for the careless ($r = -.08$, $p >$ 05) and random ($r = .05$, $p > .05$) conditions. Using a series of Fisher *r*-to-*z* tests to compare the correlation coefficients revealed the same pattern of findings as described above. Specifically, the results of this analysis indicated that the correlation for the careful condition was significantly different than the correlation for the careless condition ($Z = 10.71$, $p < .001$) and the random condition ($Z = 9.17$, $p < .001$). Likewise, the correlation for the control condition was significantly different than the correlation for the careless condition ($Z = 9.46$, $p < .001$) and the random condition ($Z = 7.92$, $p < .001$). There were again no significant differences between the control and careful or careless and random conditions. Lastly, those in the careless and random conditions again answered many directed-response items incorrectly (Careless: Item 1 = 49%, Item 2 = 48%; Random: Item 1 = 46%; Item 2 = 46%), slightly more than the percentage answered incorrectly prior to data cleaning (see Table 3). Overall, the results of this data cleaning procedure provide greater confidence that the survey manipulations functioned as intended and participants properly adhered to the survey instructions.

Item Response Theory Analysis

A series of IRT analyses using the GRM (Samejima, 1969) were conducted using the "mirt" package in R (Chalmers, 2012) and estimated via an expectation maximization algorithm. Prior to deriving the final set of parameters for the subsequent simulation study, it was important to first ensure that the estimated IRT models were appropriate and displayed acceptable fit (Foster, Min, & Zickar, 2017; Nye et al., 2019). Poor functioning items, for instance, may introduce statistical artifacts which could undermine the accuracy of participants' θ estimates. To avoid this issue, a series of steps were taken to ensure that the estimated models were appropriate. This process is described in detail below.

First, the fit of the IRT models was examined with the following fit indices: *M2*, *RMSEA*, and *SRMSR*. Next, these findings were supplemented by conducting a series of CFA models, estimated via maximum likelihood, across the conditions using the "lavaan" package in R (Rosseel, 2012). The following CFA fit indices were examined: χ^2, *RMSEA*, and *SRMSR*. Because the primary goal of these analyses was to generate θ, *a*, and *b* estimates to inform the subsequent simulation study, unidimensional models were estimated. Initially, upon estimating both the IRT and CFA models, the results indicated that the models did not demonstrate acceptable fit. An inspection of the source of model misfit revealed that the reverse-coded items were resulting in model misspecification, which is a well-documented issue when estimating latent variable models (Baumgartner, Weijters, & Pieters, 2018; Spector, Van Katwyk, Brannick, & Chen, 1997; Wang, Chen, & Jin, 2015). Because there are methodological reasons for the removal of these items (Kline, 2016) and given that the goal of these analyses is to derive accurate parameter and θ estimates, all models were estimated without the 29 reverse-coded items included in the IPIP-NEO conscientiousness measure. This helps ensure that methodological artifacts are not distorting the results of the IRT analyses.

Table 4. IRT and CFA fit indices across all four conditions

Fit Indices	Careful	Control	Careless	Random
IRT Fit Indices				
M2	1275.41	1206.94	1124.68	398.69
RMSEA	.10	.10	.10	.03
SRMSR	.09	.09	.09	.06
CFA Fit Indices				
χ^2	1662.00	1628.53	1303.34	576.09
RMSEA	.10	.10	.10	.04
SRMSR	.09	.08	.07	.06

Note. IRT fit indices could not be computed when data was missing. For this reason, the fit indices for the carless (n = 208) and random (n = 197) conditions were computed based on somewhat truncated samples. To be consistent, cases with missing values were also omitted from the CFA analyses.

This fit of these final models is reported in Table 4. As indicated in Table 4, the fit of the model for the careful group demonstrated marginally acceptable fit for both the IRT and CFA analyses. The fit of the control condition was similar to the fit of the careful condition. Model fit, however, was somewhat better for both the careless and random conditions. Moreover, the pattern of model fit results was quite similar across both IRT and CFA methods. An inspection of the source of model misfit for the careful condition indicated that some items had highly correlated error residuals (e.g., "I live order" and "I love order and regularity"). To examine the potential sources of misfit further, a second model was estimated at the facet level. This was accomplished by specifying that each item load on its intended facet and by allowing all six facets of the conscientiousness measure to correlate. A CFA model estimated via maximum likelihood indicated that this six-dimensional model displayed slightly better fit than the unidimensional model (χ^2 = 1119.24, *RMSEA* = .08, *SRMSR* = .08). Nonetheless, because the goal of this study was to derive a set of parameters to inform the subsequent simulation, and because IRT parameters cannot be readily obtained from complex multidimensional models, the results of the unidimensional models are used.

Table 5. Summary of IRT *a* and b_i parameters for all four conditions

	a		b_1		b_2		b_3		b_4	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Careful	**1.69**	[0.64]	**-3.51**	[1.21]	**-2.30**	[0.93]	**-1.29**	[0.71]	**0.81**	[0.44]
Control	**1.93**	[0.74]	**-3.39**	[0.96]	**-2.10**	[0.70]	**-1.00**	[0.53]	**0.64**	[0.34]
Careless	**2.09**	[0.22]	**-1.02**	[0.18]	**-0.32**	[0.13]	**0.36**	[0.07]	**1.30**	[0.18]
Random	**1.09**	[0.23]	**-2.22**	[0.58]	**-0.81**	[0.31]	**0.40**	[0.19]	**2.08**	[0.43]

Note. a = discrimination parameter. b_i = category threshold parameter.

Table 6. Summary of IRT *b*-to-*b* distances for all four conditions

	b_1-b_2		b_2-b_3		b_3-b_4	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*
Careful	**1.22**	[0.42]	**1.01**	[0.34]	**2.10**	[0.57]
Control	**1.30**	[0.49]	**1.10**	[0.41]	**1.70**	[0.49]
Careless	**0.70**	[0.11]	**0.68**	[0.12]	**0.94**	[0.16]
Random	**1.41**	[0.33]	**1.21**	[0.37]	**1.68**	[0.34]

Note. b_i = category threshold parameter.

Upon estimating a series of acceptably fitting unidimensional models, the *a* and b_i parameters as well as the *b*-to-*b* distances were computed for each condition. The *M* and *SD* of the *a* and b_i parameters are reported in Table 5 while the *M* and *SD* of the *b*-to-*b* distances are reported in Table 6 (the item-level parameters for all conditions are reported in Appendix B). Next, following previous research (Lake et al., 2013; Suzuki, Samuel, Pahlen, & Krueger, 2015), the parameters (*a*, b_1, b_2, b_3, and b_4) were formally compared across the four conditions by conducting a series of independent *t*-tests. Cohen's *d* values were also calculated to determine the effect sizes of these comparisons (see Table 7). These results indicate considerable variability in the parameters across all four conditions. While the results indicated differences in the *a* and b_i parameters across the control and careful conditions, these effect sizes were small to moderate in size (Cohen's *d* range 0.11 to 0.57). The differences in parameter values across the careful and careless (Cohen's *d* range 0.84 to 3.27) and careful and random conditions (Cohen's *d* range 1.25 to 3.25) were very large, though. In general, this same pattern of results emerged when comparing the careless and random conditions to the control condition. Finally, these results indicated that the differences in the parameters across the careless and random conditions (Cohen's *d* range 0.28 to 4.44) were in general, very large. This tentatively suggests that instructions to respond carelessly results in different response behavior, and hence different parameter values, compared to the instructions to respond randomly. This possibility will be more rigorously explored in the following simulation study. Overall, the results of these parameter comparisons provide additional assurance that the survey instructions resulted in unique response behaviors and that the parameters generated in this study will be useful in informing the subsequent simulation. These parameters will be discussed further in Study 2 when describing the simulation procedure.

Table 7. Summary of *t*-tests and Cohen's *d* values comparing Study 1 conditions

	a		b_1		b_2		b_3		b_4	
Condition	t	d	t	d	t	d	t	d	t	d
Careful–Control	4.14*	0.35	1.31	0.11	2.90*	0.24	5.52*	0.46	5.16*	0.43
Careful–Careless	9.81*	0.84	33.71*	2.88	34.92*	2.98	38.39*	3.27	17.17*	1.46
Careful–Random	14.62*	1.25	15.96*	1.34	25.51*	2.01	38.05*	3.25	34.52*	2.91
Control–Careless	3.48*	0.29	40.19*	3.43	41.41*	3.54	42.12*	3.53	28.51*	2.43
Control–Random	17.94*	1.53	17.35*	1.48	27.93*	2.38	41.24*	3.46	43.93*	3.71
Careless–Random	50.84*	4.44	32.70*	2.79	24.12*	2.06	3.27*	0.28	27.69*	2.37

Note. **p* < .01. *a* = discrimination parameter. b_i = category threshold parameter. *t* = *t* value from an independent *t* test. *d* = Cohen's *d*.

Study 1 Discussion

The goal of Study 1 was to produce a set of item responses and parameters to inform the subsequent simulation study thereby ensuring a degree of realism that is atypical for CR simulation studies. As noted previously, the nature of CR behavior remains poorly specified. Specifically, it is unclear whether or not CR is analogous to mathematically random responding (e.g., DeSimone et al., 2018; Meade & Craig, 2012). It is also impossible for simulations to rely on existing datasets since CR indices do a poor job at reliably identify careless responders. By using the results of this study to inform the following simulation, however, it is possible to circumvent both of these limitations of previous research. For instance, because response behaviors were shaped experimentally in this study, there is more certainty regarding the nature of the responses (i.e., whether they are careful or careless). The results from the manipulation checks, and subsequent screening and data cleaning that occurred, provide additional confidence that participants' responses reflect the intended response behaviors. Furthermore, the inclusion of separate conditions for careless and random responding will also permit a way to explicitly compare these two types of response behaviors to mathematically random responding (and other forms of CR, described in detail in Study 2).

Interestingly, the results of this study reiterate how CR impacts model fit in unpredictable and seemingly counterintuitive ways (e.g., Beck et al., 2019; Goldammer et al., 2020; Huang et al., 2012). For example, model fit was often better for the careless and random groups in comparison to the careful and control groups for the IRT and CFA analyses. These initial results should be interpreted with caution, though. For instance, it's doubtful that a dataset will ever contain up to 100% CR (see Meade & Craig, 2012). The way that CR affects model fit under this circumstance is not well-specified. Furthermore, it was not possible to include responses with missing values in the computation of the model fit indices. Because there were many missing values in both the careless and random conditions, the computation of model fit indices occurred with a considerably truncated sample, which may have impacted these findings. Despite the fit still being somewhat less than ideal for the careful condition, some of the sources of misfit were identified (e.g., reverse coded items, how dimensionality is modeled, correlated item residuals). Although models could have been further refined to improve fit, this would not be desirable and likely capitalize on change findings. Regardless, the goal of this study was not to assess the absolute fit of the models per se, but to ensure that the models displayed sufficiently acceptable fit so as to produce a set of useful and realistic parameters to inform the subsequent simulation.[2] Nonetheless, these results do underscore the need to know more about the relation between CR and model-data fit.

It is also interesting to note that the results for the control and careful conditions were very similar in many regards. For example, the results from the manipulation checks indicated that these groups were similar in terms of the accuracy of their responses, correlation between the two conscientiousness measures, and the number of directed-response items answered incorrectly. Furthermore, model fit for the two groups was also very similar. While there were some differences in the IRT parameters across these groups, these effects were generally small, especially in comparison to the careless and random conditions. The "statistically significant" differences that were observed are likely being driven by the large size of each comparison group (which is why Cohen's *d* values were computed to supplement these findings). Thus, the instructions to respond carefully did not appear to make much of a difference as participants in

[2] The results of the data generation procedure for the simulation (see Study 2) provide further assurance that the somewhat lower fit observed for the careful condition in Study 1 is not detrimental. Specifically, the IRT parameters from the careful condition in Study 1 that were used to generate careful responses in the simulation resulted in consistently good fitting models (see Appendix C for details).

the control condition tended to also respond carefully, even without explicit instructions to do so. This finding reiterates how interventions designed to reduce CR are often minimally effective (e.g., Francavilla et al., 2019; Ward & Meade, 2018).

Finally, an interesting finding from this study is that the instructions to respond carelessly resulted in different model fit values and parameter estimates compared to the instructions to respond randomly. In general, the effect sizes of these differences were extremely large, far exceeding the Cohen's *d* threshold of .80 that normally constitutes a large effect. These results corroborate the decision to include separate conditions for these response behaviors and tentatively suggest that careless and random responding do not represent equivalent response behaviors. This possibility will be examined in greater detail in the following simulation study.

Chapter 3 - Study 2

The primary goal of Study 2 was to use the parameters and item responses generated in Study 1 to conduct a simulation study examining the (1) relations between CR and the fit of both IRT and CFA latent variable models (2) and compare different types of CR. In doing so, Study 2 demonstrates the utility of using an IRT-based Monte Carlo simulation procedure for studying the effects of CR on the fit of latent variable models. This procedure is described in detail below.

Method

Summary of Conditions

This study manipulated five different independent variables (IVs). Specifically, a mixture of two survey-based and three respondent-based variables were manipulated. These IVs were selected to be realistic and representative of situations that researchers and practitioners commonly encounter (Harwell et al., 1996). Furthermore, the various levels of the IVs were selected due to their realism and ability to ensure the generalizability of the simulation findings across a range of frequently encountered survey contexts (see Highhouse, 2009). These manipulations are described in detail below and summarized in Table 8.

The first survey-based IV manipulated is sample size. Specifically, sample sizes of 200 and 500 will be examined for this simulation study. These particular levels were selected for various reasons. First, because IRT generally requires large samples of up to 500 participants (Embretson & Resise, 2000; Foster et al., 2017), knowledge of how CR impacts model fit in both suboptimal (Sample size = 200) and sufficient (Sample size = 500) conditions is important. Also, sample sizes of 200 and 500 are often the values selected to represent small and moderate/large sample sizes in CR simulations (Woods, 2006). Furthermore, a sample size of 200 is also just slightly higher than the median sample size that is commonly observed in organizational studies (e.g., Shen et al., 2011). Examining sample sizes of 200 and 500 is, therefore, consistent with this study's goal of examining realistic, frequently encountered scenarios.

The second survey-based IV examined is the number of items. Specifically, surveys with 10 and 60 items are examined in this simulation study. Examining the relation between CR and model fit on surveys with varying numbers of items is important given the variability in the number of items contained within psychological measures (see Hinkin, 1998). Furthermore, because CR has been found to increase toward the end of long surveys (e.g., Bowling et al., 2020; Gibson & Bowling, 2019), knowledge of how CR affects model fit for both short (Number of items = 10) and long (Number of items = 60) surveys is important for examining the generalizability of the effects of CR on model-data fit.

The next three IVs that were manipulated represent the respondent-based, CR manipulations. The first of these is entitled, "CR Prevalence" (5%, 15%, 30%), which refers to the percentage of the sample engaging in CR. Typical estimates of the prevalence of CR range from 10%-12% (Meade & Craig, 2012; but see also DeSimone et al., 2018). Hence, a value of "15%" roughly corresponds to the usual prevalence of CR in a sample. The values of "5%" and "30%", therefore, correspond to low and high rates of CR, respectively. Furthermore, these values also roughly align with the values that have been employed in other CR simulation studies (Hong et al., 2020; Niessen et al., 2016).

The second respondent-based IV that was manipulated is "CR Severity" (20%, 50%, 100%), which refers to the percentage of a respondent's responses that are careless. While it is often assumed that respondents will engage in CR across an entire survey, it is more realistic that respondents engage in partial CR (Meade & Craig, 2012). Support for this assertion comes from research showing that CR often increases at the latter portions of a survey, especially for particularly long surveys (Bowling et al., 2020; Galesic & Bosnjak, 2009; Gibson & Bowling, 2019; Yu & Cheng, 2019). Within the context of this IV, a respondent with 100% CR Severity is engaging in CR across their entire response vector. A respondent with 50% CR Severity is only engaging in CR for the final 50% of their response vector. Likewise, a responded with 20% CR Severity is only engaging in CR for the final 20% of their response vector.

The final IV, "CR Type" (Respondent-derived careless responding [CR_C], respondent-derived random responding [RR_C], mathematical random responding [MRR_C], invariant responding [IR_C]), refers to the form of the CR.[3] The first two levels of this variable (i.e., RR_C and CR_C) reflect the response patterns that participants were instructed to exhibit in the "Random" and "Careless" conditions in Study 1, respectively. The next level of this variable, MRR_C refers to mathematically random responding. In the CR literature, it is often assumed that CR follows a uniform distribution (e.g., DeSimone et al., 2018; Huang et al., 2012; Huang et al., 2015; Meade & Craig, 2012). For example, if participants were responding in a mathematically random manner to a survey that employed a 5-point Likert scale, the probability of endorsing each response option is 20%. Hence, the responses to all survey items would be uniformly distributed across the respondent. As noted above, because people find it difficult to behave in mathematically random ways (Nickerson, 2002) and CR indices do a suboptimal job at reliably identifying careless responders, it is unclear if true random responding is analogous to CR. Because participants provided careless and random responses in Study 1, however, it is possible to compare these response patterns to mathematically random response patterns within this simulation study. Finally, the last level of this variable, IR_C, refers to a form of CR where

[3] Here, the "C" subscript is meant to denote "Condition" since there are four "CR Type" conditions. More importantly, this distinguishes CR, which refers to careless responding in a broad sense, from CR_C, which refers to the condition containing the respondent-derived careless responses from Study 1.

participants select the same response option consecutively (e.g., selecting “3” for all items; DeSimone et al., 2018; Meade & Craig, 2012). This form of CR is very distinct from random responding in that there is no variability in the response patterns. To summarize, this simulation study employs a 2 (Sample size: 200 and 500) × 2 (Number of items: 10 and 60) × 3 (CR prevalence: 5%, 15%, 30%) × 3 (CR severity: 20%, 50%, 100%) × 4 (CR type: (Respondent-derived careless responding [CR_{C}], respondent-derived random responding [RR_{C}], mathematical random responding [MRR_{C}], invariant responding [IR_{C}]) design with 144 unique conditions.

Table 8. List of simulation independent variables and levels

Independent Variable	Levels
Sample Size	200 participants
	500 participants
Number of Items	10 items
	60 items
CR Prevalence	5% of participants
	15% of participants
	30% of participants
CR Severity	20% of responses
	50% of responses
	100% of responses
CR Type	Respondent-Derived Careless (CR_{C})
	Respondent-Derived Random (RR_{C})
	Mathematical Random (MRR_{C})
	Invariant Responding (IR_{C})

Note. CR = Careless responding. “Participant-Derived Random” and “Participant-Derived Careless” refer to the type of CR exhibited in the careless and random conditions from Study 1, respectively. There are 144 total conditions. For the “CR Severity” levels of 20% and 50%, the final 20% and 50% of responses will be replaced, respectively. This will be done in order to simulate situations where CR occurs at the end of surveys.

Simulation Procedure

A flowchart summarizing the general simulation procedure that will be used can be seen in Figure 3. The following conditions from Study 1 will be used to derive the parameters and item responses needed for the simulation study: (1) Careful responding, (2) careless responding, and (3) random responding. Because the results of the control condition did not substantially differ from the careful condition within Study 1, the parameters computed for this condition will not be used in the present simulation. As indicated in Figure 3, the simulation procedure contains four primary steps, with steps one, two, and four containing three sub-steps. The first step of this procedure will be to generate a series of *careful* (i.e., uncontaminated) item responses. Careful item responses will be generated for the "Sample size" (200 and 500) and "Number of Items" (10, 60) conditions (2 × 2 = 4 total conditions). These four careful response conditions will serve as the baseline item responses for examining how the introduction of CR impacts model fit under different circumstances. CR will be introduced for each combination of the remaining independent variables: "CR prevalence" (5%, 15%, 30%), "CR severity" (20%, 50%, 100%) and "CR type" (CR_C, RR_C, MRR_C, IR_C). Thus, for each careful condition, 36 different CR combinations will be examined (3 × 3 × 4 = 36 total conditions). Note that multiplying 36 by four (for each of the four baseline careful conditions) yields 144, which is the total number of conditions described above. The procedure summarized in Figure 3 is described further below.

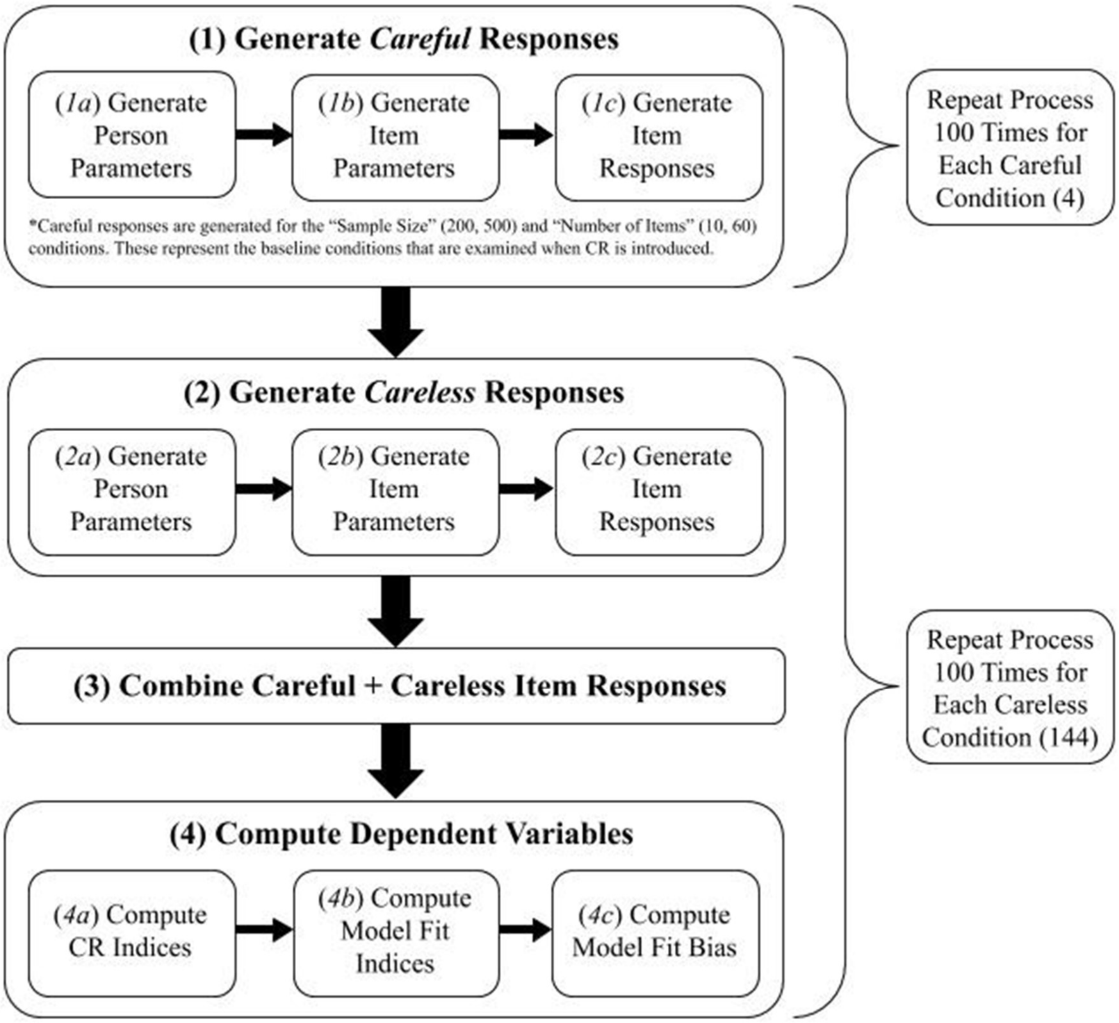

(1) Generate Careful Responses
(1a) Generate Person Parameters
(1b) Generate Item Parameters
(1c) Generate Item Responses
*Careful responses are generated for the "Sample Size" (200, 500) and "Number of Items" (10, 60) conditions. These represent the baseline conditions that are examined when CR is introduced.
Repeat Process 100 Times for Each Careful Condition (4)
(2) Generate Careless Responses
(2a) Generate Person Parameters
(2b) Generate Item Parameters
(2c) Generate Item Responses
(3) Combine Careful + Careless Item Responses
(4) Compute Dependent Variables
(4a) Compute CR Indices
(4b) Compute Model Fit Indices
(4c) Compute Model Fit Bias
Repeat Process 100 Times for Each Careless Condition (144)

Figure 3. Flowchart of the simulation procedure.

Step 1 – Simulating Careful Responses

Step 1a – Generate Careful Person Parameters. As noted above and in Figure 3, the first step of the simulation procedure is to generate a series of careful item responses for the "Sample size" (200 and 500) and "Number of Items" (10, 60) conditions (2 × 2 = 4 total conditions). The first sub-step of Step 1 will be to generate person parameters (θ) within an IRT framework in order to denote each respondent's position on the θ continuum. An examination of the *M* and *SD* of the θ score distributions for the careful condition in Study 1indicated that this group had a *M* of approximately 0 and a *SD* of approximately 1. This is not surprising since the scale of θ is somewhat arbitrary and mean θ values tend to center around 0. Accordingly, in order to place the θ estimates on the standard IRT continuum ranging from -3 to 3, person parameters will be randomly sampled from a *N*(0, 1) normal distribution.

Step 1b – Generate Careful Item Parameters. To generate item parameters, the *a* and b_i values that were computed in the careful condition for Study 1 will be used (see Tables 5 and 6). First, to generate plausible *a* parameters, values will be randomly sampled from a normal distribution of *a* parameters (see Meade, Lautenschlager, & Johnson, 2007). An examination of the distribution of *a* parameters in Study 1 indicated these values had a *M* of 1.69 and *SD* of 0.64. Thus, *a* values will be randomly sampled from a *N*(1.69, 0.64) normal distribution for the simulation. To generate a set of b_i parameters, b_1 values will first be randomly sampled from a normal distribution that will again correspond to the distribution of b_1 values in Study 1. Because the b_1 values for the careful condition had a *M* of -3.51 and *SD* of 1.21, these values will be randomly sampled from a *N*(-3.51, 1.21) normal distribution (see Table 5).

Following previous research (LaHuis, Clark, & O'Brien, 2011; Meade et al., 2007), the randomly sampled b_1 values will be used as the baseline values for the computation of the remaining three b_i values. Typically, previous research using this method will next add constant b_i values to the initial b_1 values. Because this method would yield a set of b_i values with equal *b*-to-*b* distances, which is unrealistic, a slightly different approach will be used in the following study. Instead, *b*-to-*b* distances will be calculated, randomly sampled from a normal distribution of *b*-to-*b* distances, and then added to the initial b_1 value in an iterative manner. Specifically, a b_1-to-b_2 distance value will be randomly sampled from a normal distribution of b_1-to-b_2 distances and then added to the initial b_1 value. This will yield a b_2 value. Next a b_2-to-b_3 distance value will be randomly sampled and then added to the b_2 value to yield a b_3 value. Finally, a b_3-to-b_4 distance value will be randomly sampled and then added to the b_3 value to yield a b_4 value. For the careful condition, this will entail randomly sampling from a b_1-to-b_2 $N(1.22, 0.42)$ normal distribution, a b_2-to-b_3 $N(1.01, 0.34)$ normal distribution, and a b_3-to-b_4 $N(2.10, 0.57)$ normal distribution (see Table 6). A conceptual representation of the b_i and b-to-b distance values used for this process can be seen in Figure 4. This process will be repeated for every item in each condition to generate a realistic set of b_i parameters.

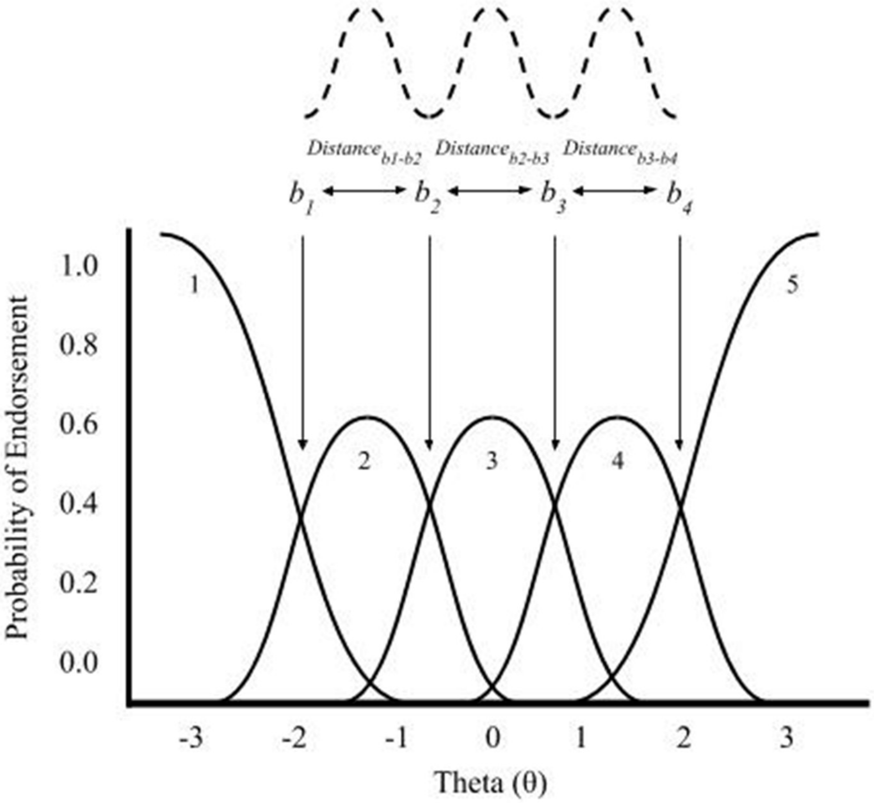


Figure 4. Conceptual representation of b_i and b-to-b distance parameters

Step 1c – Generate Careful Item Responses. To generate realistic item responses that reflect careful responding, an IRT-based response generation procedure will be used with the WinGen software package (Han, 2007). All item responses will have five response categories so as to simulate commonly used 5-point, Likert-type surveys. This also corresponds to the number of response categories used in the Study 1 survey. The θ, *a,* and b_i values generated from Steps 1a and 1b will be used to inform the item responses. While IRT typically provides information about a person's θ level given their response pattern and model parameters, the procedure used to generate item responses for this study will reverse this process and simulate a person's response pattern given the true model parameters and their θ values.

For each careful condition, 20 sets of θ values and 20 sets of *a* and b_i values will be generated. For each set of *a* and b_i values, and the corresponding set of θ values, five sets of item responses (i.e., datasets) will be generated. Each condition, therefore, contains 100 replications (20 condition parameters/θ values × 5 sets of item responses for each set of parameters/θ values = 100 replications). This far exceeds the minimum threshold of 25 replications that has been recommended for IRT simulation studies (Harwell et al., 1996) and is consistent with more recent simulation studies on both CR (e.g., Hong et al., 2020; Niessen et al., 2016) and IRT (e.g., LaHuis et al., 2013; Meade et al., 2007). The goal of generating multiples sets of parameters and item responses within each condition is to derive stable estimates and mitigate error that can occur within the parameter/item generation process. This process (Step 1) will occur four times in total for each of the baseline (i.e., careful) conditions. To be concrete, Steps 1a through Step 1c will occur for the following four baseline conditions: 10 items/200 participants, 10 items/500 participants, 60 items/200 participants, and 60 items/500 participants. These conditions will serve as the baseline groups for understanding how model fit is impacted once CR is introduced. Figure 5 provides a visualization of the general data generation procedure that is used.

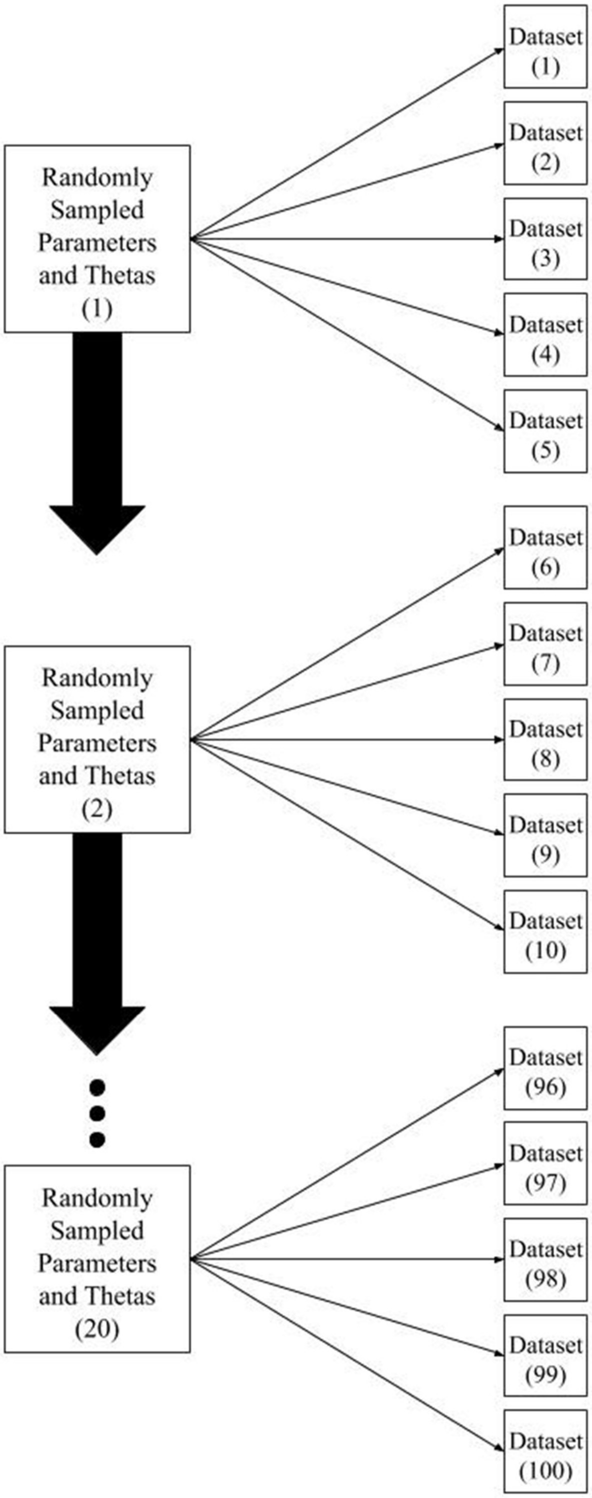


Figure 5. Visual representation of how datasets for each condition are generated

Step 2 – Simulating Careless Responses

Step 2a – Generate Careless Person Parameters. As indicated in Figure 3, the next overall step in the simulation procedure is to generate the *careless* responses. In general, the process of generating careless responses for two of the four “CR Type” conditions (i.e., CR_C and RR_C) will be very similar to the process of generating careful responses described above. First, a set of person parameters will be generated. An examination of the *M* and *SD* of the θ score distributions for these two conditions in Study 1 indicated that these groups both had θ values with a *M* of approximately 0 and a *SD* of approximately 1. Accordingly, for both these conditions, person parameters will be randomly sampled from a *N*(0, 1) normal distribution.

Step 2b – Generate Careless Item Parameters. The process of generating *a* and b_i values for the careless and random conditions will be nearly identical to what is described above for the careful condition. The main difference concerns the values that will be sampled for each condition. For the CR_C condition, *a* values will be sampled from a *N*(2.09, 0.22) normal distribution while for the RR_C condition, *a* values will be sampled from a *N*(1.09, 0.23) normal distribution. Additionally, for the CR_C condition, the b_1 values will be sampled from a *N*(-1.02, 0.18) normal distribution while for the RR_C condition, b_1 values will be sampled from a *N*(-2.22, 0.58) normal distribution (see Table 5). The remaining b_i values will be computed based on the procedure using the *b*-to-*b* distances described in Step 1b. For the CR condition, this will entail randomly sampling from a b_1-to-b_2 *N*(0.70, 0.11) normal distribution, a b_2-to-b_3 *N*(0.68, 0.12) normal distribution, and a b_3-to-b_4 *N*(0.94, 0.16) normal distribution. For the RR condition, this will entail randomly sampling from a b_1-to-b_2 *N*(1.41, 0.33) normal distribution, a b_2-to-b_3 *N*(1.21, 0.37) normal distribution, and a b_3-to-b_4 *N*(1.68, 0.34) normal distribution (see Table 6).

Step 2c – Generate Careless Item Responses. Careless item responses will be generated using the same IRT-based response generation procedure described in Step 1c for the CR_C and RR_C conditions. That is, the θ, *a,* and b_i values generated from Steps 2a and 2b will be used to inform the item responses. For both the CR_C and RR_C conditions, 20 sets of θ values and 20 sets of *a* and b_i values will be generated. For each set of *a* and b_i values, and the corresponding set of θ values, five sets of item responses will be generated. Again, this will yield a set of 100 item responses (i.e., replications) for each condition.

For the remaining other "CR Type" conditions (i.e., MRR_C and IR_C conditions), a different procedure for generating item responses will be used. For the MRR_C condition, responses will be computed so as to align with the assumptions of a uniform distribution. In this case, because responses are being simulated for a 5-point scale, each response option within the MRR_C condition will have a 20% probability of being endorsed. For the IR_C condition, the response option that is simulated will be evenly distributed within each condition. For example, in the condition with 200 items, 10 participants, 5% prevalence and 100% severity, a total of 10 participants' response vectors will be replaced with invariant responses. Of these 10 responses, the first two will contain the first response option selected 10 consecutive times (e.g., "1s" in the Likert scale), the next two response vectors will contain the second response option selected 10 consecutive times (e.g., "2s" in the Likert scale) and so forth. Because there is no reason to expect that participants would generally favor one response option over another when completing an actual survey, this method will hold response options that are selected constant and enable an easier interpretation of the impact of this type of CR (e.g., DeSimone et al., 2018).[4] For each combination of careless conditions that is examined (recall that there are 36 possible combinations for each of the four baseline conditions), this process of generating careless responses will be enacted. Thus, the process of generating careless responses will occur 144 total times (again note that this corresponds to the total number of simulation study conditions described above). Because each condition is replicated 100 times, this will yield a total of 14,400 simulated conditions (100 × 144 = 14,400 total simulated conditions).

Step 3 – Combine Item Responses

Once careful and careless item responses are generated, the next step in the simulation procedure is to combine these responses in the appropriate manner and determine how the introduction of CR affects model-data fit. For each of the 144 conditions, the careless responses will replace a subset of the responses in the careful dataset, thereby yielding a dataset with a mixture of careful and careless responses. The subset of careful item responses that will be replaced will always be the first set of item responses in the careful dataset. For example, in the baseline condition with 10 items and 200 participants, when CR prevalence is 30%, the first 60 careful item responses will be replaced with 60 careless item responses. Likewise, when CR prevalence is 5%, the first 10 careful responses will be replaced with 20 careful responses, and so forth. Holding the responses that are replaced constant will provide a way to explicitly compare the effects of different forms of CR on model fit. An example of this process is described below.

In the following example, the condition that contains 200 participants, 10 items, 5% CR prevalence, 100% CR severity, and the CR type is "RR_C" will be examined. First, a set of 20

[4] Examining other invariant response options (e.g., situations where only the highest response option [e.g., a "5" on a 5-point Likert scale] is selected) would require creating additional conditions. This would, however, increase the total number of conditions to the point that results would be too difficult to interpret. Thus, examining different combinations of invariant responding in greater detail would be better accomplished in future research.

careful θ values will be generated (Step 1a) followed by a set of 20 careful *a* and b_i values (Step 1b). Next, a set of 100 careful item responses will be generated as described in Step 1c. Once the careful item responses are generated for this condition, a set of careless responses are generated using the procedure described in Steps 2a through 2c. Because this condition is examining "RR_C", the parameters generated from the random condition in Study 1 are used to inform the IRT-based response generation procedure for this condition. This will yield a set of 100 careless responses for this condition. The next step is to combine the careful and careless responses to yield a set of 100 item responses (i.e., datasets) containing a mixture of careful and careless responses that correspond to the condition in question. In this case, to combine item responses, 5% of the items in a careful dataset for this condition are replaced with a subset of careless responses from a careless dataset. Since this condition contains 200 participants, and 5% of respondents are engaging in CR, this will entail replacing 10 careful responses with 10 careless responses. Specifically, the responses for the first 10 participants in the careful dataset will be replaced. Because CR severity for this condition is 100%, this means that 100% of the responses of those engaging in CR are careless. In other words, their entire response vector contains careless responses. This process of combining careful and careless responses will occur 100 times for this condition. That is, the first 10 responses from the *first* careless dataset that is generated will replace the first 10 responses in the first *careful* dataset. Next, the first 10 responses from the *second* careless dataset that is generated will replace the first 10 careful responses in the *second* careful dataset, and so on, until this process has been enacted 100 times.

This general procedure will be used for all conditions though will differ slightly depending on the condition being examined. Note that in this example, careful responses were only computed for one baseline condition. As indicated in Figure 3, however, the first step involves computing responses for all four careful baseline conditions, followed by the 144 careless conditions. While the procedure used in this example does not completely match the procedure reported in Figure 3, it does illustrate the general simulation process of generating and combining careful and careless responses that will be used in the simulation study.

Step 4 – Compute Dependent Variables

Step 4a – Estimate CR Indices. Once the item responses are generated for each condition, CR will be examined as a sort of validity check and ensure the simulation method is functioning as intended. For example, the subset of careless responses that are introduced to the careful dataset should be flagged more frequently by the CR indices compared to the remaining careful data. Thus, estimating the amount of CR for the careful and careless responders within each condition will provide a useful way to assess the validity of the method being employed as well as the sensitivity of CR indices to the data generation procedure. Following recommendations from previous research, CR will be computed with the use of multiple CR indices (Meade & Craig, 2012; Niessen et al., 2016; Steedle et al., 2019). Two "traditional" indices and two IRT indices will be used: Mahalanobis distance (*D*), maximum longstring (*MLS*), l_z, and *G.* For each condition, the CR index values that are computed will be averaged across the 100 replications/datasets. These indices are described in detail below.

As noted previously, *D* is a multivariate outlier detection method that represents the distance between a person's response vector and the sample's mean response vector (Mahalanobis, 1936). Put another way, *D* indicates the degree to which a person's response pattern to a set of items deviates from the normative response pattern across all the entire sample. Formally, *D* is defined as:

$$D = \sqrt{(x_i - \bar{x})C_x^{-1}(x_i - \bar{x})^T} \tag{8}$$

Here, *x* represents the set of responses for person i, $\bar{x}$ represents the mean set of responses for the rest of the sample, and C_x^{-1} is the inverse covariance matrix for the set of items. Higher Mahalanobis distance values indicate greater careless responding (i.e., more aberrant response patterns) while lower values indicate more careful responding (i.e., responses that are more similar to the normative response patterns).

The second CR index that will be included is the *MLS* index. The *MLS* index identifies the maximum number of consecutive responses that a person selects for a set of items (Johnson, 2005; Meade & Craig, 2012). For instance, consider the response vectors to a 10-item, 5-point Likert-type scale for three people:

$$\text{Person A} = [1,3,3,1,2,2,3,1,1,2] \tag{9}$$

$$\text{Person B} = [4,4,4,4,4,4,4,4,4,4] \tag{10}$$

$$\text{Person C} = [4,5,5,3,4,4,5,5,3,5] \tag{11}$$

Here, there is some variability for both Person A and Person C. Person B, however, has selected the same response option (“4”) for all 10 items. In general, higher *MLS* values indicate greater CR since it is unlikely that someone would not have any variability in their responses to a set of items (which is especially true when reverse-coded items are included in a scale). Further, because items typically assess different levels of the underlying construct being measured, responses to the items should be somewhat variable. Note that the *MLS* index is best suited to identify invariant responding as opposed to random responding since random responding, by definition, will contain some variability.

The next two CR indices that will be examined are IRT person-fit indices. The first of these is *l_z*. The *l_z* index is a parametric IRT person-fit index that measures the likelihood of a person's response pattern given their theta (θ) level (Drasgow, Levine, & Williams, 1985). That is, the *l_z* index compares a respondent's log-likelihood response pattern to the expected log-likelihood for that θ value. If a person's responses to a set of items are unlikely in comparison to other respondents with a similar θs (i.e., normative response patterns), the person's response pattern will be classified as inconsistent. Generally, this is the basic logic underlying most IRT person-fit indices, though these indices will differ in the way in which response pattern aberrations are assessed (see Karabatsos, 2003). Formally, the *l_z* index is defined as:

$$l_z = \frac{l_0 - E(l_0)}{V(l_0)^{1/2}} \tag{12}$$

Here, *l_0* represents the log-likelihood estimate for a person's responses to a set of polytomous items, *$E(l_0)$* represents the mean of the log-likelihood function, and *$V(l_0)$* represents the variance of the log-likelihood function. Because *l_z* is standardized, normal response patterns will have a value near zero (Drasgow et al., 1985). Lower *l_z* values generally indicate greater CR (i.e., very unusual response patterns).

The final CR index that will be examined is *G*. Originally, *G* was developed for dichotomous items with varying levels of difficulty. *G* represents the number of times a person answers a difficult item correctly, but an easier item incorrectly (Guttman, 1944; see also Curran; 2016; Niessen, Meijer, & Tendeiro, 2016). Because easier items should be answered correctly if difficult items are also answered correctly, the situation where difficult items are correct but easy items are incorrect indicates an inconsistent response pattern. *G* can be applied to polytomous (i.e., Likert) items as well (Emons, 2008). Here, *G* is calculated with the use of item steps (a person’s psychological threshold of response between ordinal response options; see Niessen et al., 2016). Consider a person responding to an item for a 5-point Likert scale where the options range from (1) *strongly disagree* to (5) *strongly agree.* When responding to the items, the person attempts to determine if they should take the item step from (1) *strongly disagree* to (2) *somewhat disagree*. This process continues until a satisfactory option or the last response option (5 *strongly agree*) is reached and is then repeated for all items. The proportion of people who took each item step for all items is then computed and ordered by popularity. Each person’s response pattern is then compared to the item step’s order. In this instance, a Guttman error occurs when a less popular item step was taken (for a hard item) but a more popular item step was not taken (for an easy item). Note that items with a lower probability of being endorsed are considered harder items. Formally, *G* (for polytomous items) is defined as (from Curran, 2016):

$$G = \sum_{h,e} f(X_{nh} > X_{ne} \rightarrow 1; else \rightarrow 0) \tag{13}$$

Here, X_{nh} represents a person’s response to a hard item (i.e., an item with a lower Likert response option value [“3”]) and X_{ne} represents a person’s response to an easy item (i.e., an item with a higher Likert response option value [“5”]). Because *G* represents the number of instances where someone’s response pattern is incongruent with the ordering of item steps, higher values are indicative of CR.

Step 4b – Estimate Model Fit Indices. Model fit represents the primary dependent variable (DV) in this study. To comprehensively understand the relation between CR and model fit, the fit of both IRT and CFA models will be examined. Additionally, multiple fit indices will be estimated for each type of latent variable model. Specifically, for IRT models, the following fit indices will be examined: *M2*, *RMSEA,* and *SRMSR.* For CFA models, the following fit indices will be examined: χ^2, *RMSEA,* and *SRMSR.* These indices were selected given their prominent use within both IRT and CFA frameworks (see Kline, 2016; Maydeu-Olivares & Joe, 2006; Nye et al., 2019; Ropovik, 2015). For each condition, the model fit values that are computed will be averaged across the 100 replications/datasets.

Step 4c – Estimate Bias. Lastly, to better understand the relation between CR and model fit, the degree of bias that is present within each condition will be estimated. Generally, bias assesses the extent to which a sample parameter value deviates from the true population parameter value (Harwell, 2019). Thus, bias estimates the direction of sample-population parameter divergence and the magnitude of this divergence. Formally, bias can be defined as:

$$\frac{\sum_{l=1}^{r}\left(\hat{\xi}_l - \xi\right)}{r} \tag{14}$$

Here, ξ represents the true population value for a parameter, $\hat{\xi}$ represents the sample estimate of the parameter, and *r* represent the number of replications. While bias is generally used for examining model parameters, this method will be slightly adapted in the present study for purposes of examining model-data fit. Thus, for each condition, the model fit value that is derived after CR is introduced within the dataset can be compared to the model fit for the baseline model (i.e., the “careful” datasets). Recall that for this study there are four baseline datasets, comprising the “Sample Size” and “Number of Items” conditions, that are uncontaminated with CR. Estimating model bias will provide a way to quantify the direction and magnitude of the change in model fit that occurs due to CR. For each condition, bias values that are computed will be averaged across the 100 replications/datasets.

Results

Validity Checks – Careless Responding Indices

First, to verify the validity of the simulation procedure, four unique CR indices were computed for all 144 conditions. All validity check results were derived from the combined (i.e., contaminated; see Step 3 in Figure 3) datasets/responses. The l_z and *G* indices were computed with the R package, "PerFit" (Tendeiro, Meijer, & Niessen, 2016) while the *D* and *MLS* indices were computed with the R package, "careless" (Yentes & Wilhelm, 2018). Specifically, the four CR indices were computed for each participant across the 100 datasets (i.e., replications) for a given condition. Next, the CR values were averaged for the subset of careless responses. CR values were also averaged for the subset of careful responses. This general process was repeated for all 144 conditions and was done in order to facilitate comparisons between the careless and careful responses within the combined datasets (Appendix D provides a more detailed example of this process). Overall, it was expected that the CR indices would indicate greater amounts of CR for the subset of careless responses compared to the subset of careful responses.

Specifically, it was expected that three of the CR indices (i.e., l_z, *G*, and *D*) would indicate higher amounts of CR for the careless responses compared to the careful responses for the following CR Type conditions: careless, respondent-derived random, and mathematical random. For the *MLS* index, it was expected that this index would only indicate greater CR for invariant responding. This is because random responding and invariant responding are distinct response patterns and the *MLS* is designed specifically to detect long strings of identical response options, not random response patterns. In contrast, the other three indices (i.e., l_z, *G*, and *D*) are better suited to detect aberrant (i.e., random) responses. For this reason, *MLS* values are expected to be lower for all CR Types (i.e., careless, respondent-derived random, mathematical random) except for invariant responding, in which case *MLS* values will actually be higher. Likewise, l_z, *G*, and *D* are expected to better detect CR within the three non-invariant "CR Type" conditions.

Summary of Validity Checks. The results of this analysis showed that the subset of careless responses was more likely to be flagged as careless compared to the subset of careful responses in most cases (see Appendix E for the CR values across all 144 conditions). To summarize the findings from all 144 conditions in a succinct manner, Table 9 contains the average CR index values broken down by each baseline condition. As a reminder, higher *G*, *D*, and *MLS* values, but lower l_z values, indicate greater amounts of CR. As can be seen, the subset of careless responses was consistently identified as having more CR compared to the careful responses. To rigorously assess these comparisons, a series of *t*-tests comparing the CR index values between the careless and careful responses were conducted. As Table 9 indicates, 14 out of the 16 possible comparisons were significant. While there are a few exceptions to this trend for some of the specific conditions (see Appendix D), these findings show that, overall, this simulation procedure is functioning as intended and emulating the desired response patterns.

Table 9. Summary of validity checks averaged across baseline conditions

	IRT Person-Fit Indices				Traditional CR Indices			
	l_z		*G*		*D*		*MLS*	
Condition	Careless Subset	Careful Subset	Careless Subset	Careful Subset	Careless Subset	Careful Subset	Careless Subset	Careful Subset
200 Participants; 10 Items	**-1.18***	0.27	**46.70***	17.24	**17.83***	8.87	**3.56**	3.07
200 Participants; 60 Items	**-3.41***	0.18	**1894.44***	710.99	**79.13***	56.39	**12.22***	5.27
500 Participants; 10 Items	**-1.35***	0.23	**48.62***	16.94	**18.44***	8.85	**3.55**	2.90
500 Participants; 60 Items	**-3.32***	0.38	**1943.68***	754.03	**88.88***	55.84	**12.20***	5.22

Note. **t*-test indicates significant difference between careful/careless CR index pair ($p < .05$). Bolded CR index values for each CR index pair indicate higher amounts of CR. l_z = standardized loglikelihood. *G* = Guttman error. *D* = Mahalanobis distance. *MLS* = Maximum longstring. Higher *G, D,* and *MLS,* and lower l_z values, indicate greater CR. Bolded l_z, *G, D,* and *MLS* values are indicative of higher CR. Values represent mean values averaged across both conditions and 100 replications (see Appendix D for detailed results).

Careless Responding and Model Fit

After establishing the validity of the simulation procedure, model fit was computed within each condition. As a reminder, model fit was computed with the combined datasets (Step 3 of Figure 3) and the results within each condition were averaged across 100 replications. Furthermore, as noted above, this entailed computing three IRT fit indices (*M2*, *RMSEA,* and *SRMSR*) and three CFA fit indices (χ^2, *RMSEA,* and *SRMSR*). Unidimensional models were estimated for all IRT and CFA models (so as to align with the models from Study 1) such that each indicator/item loaded onto a single latent construct. The specific way this occurred (e.g., whether there were 10 or 60 items/indicators) slightly varied depending on the condition. A summary of the model fit indices for all 144 conditions is provided in Appendix F. Note that the values reported in Tables 34 through 37 in Appendix F represent the simulation DVs.

Careless Responding and IRT Model Fit. While the above information is useful for summarizing model fit across all conditions, it is difficult to determine the specific ways in which CR affects model fit based solely on this descriptive information. Accordingly, to more rigorously understand how the five simulation IVs (i.e., Sample size, number of items, CR prevalence, CR severity, CR Type) affect model fit, a series of factorial ANOVAs were conducted. All five conditions (and their associated levels), as well as each possible two-way interaction[5] (10 total), were entered as IVs with each of the three IRT model fit indices serving as the DVs. For each condition, all 100 replications were included in the analyses. Thus, each ANOVA was conducted on a dataset containing 14,400 total rows (which is analogous to “participants” within typical ANOVA frameworks), with each condition having 100 replications

[5] To ensure model convergence and ease the interpretation of the results, interactions containing more than two terms were not included in any of the estimated models.

(i.e., “participants”). Finally, because the large number of data points would likely make every result significant, η^2 (which is a measure of effect size) was also calculated to supplement these findings and assess the magnitude of the IVs’ effects.

As seen in Table 10, the overall models for the *M2* ($F = 1430.20, p < .001$)*, RMSEA* ($F = 353.29, p < .001$) and *SRMSR* ($F = 786.60, p < .001$) IRT model fit indices were significant. There are several notable aspects of these results. First, “number of items” has the largest impact on the *M2* fit index ($\eta^2 = .44$). This is not surprising, however, since *M2* is a function of item number. Second, the *RMSEA* fit index is most impacted by CR severity ($\eta^2 = .19$) followed by CR prevalence ($\eta^2 = .11$). Third, *SRMSR* is equally impacted by CR prevalence and CR severity (η^2s = .20). Fourth, CR type had only small/moderate effects across the three IRT fit indices (i.e., η^2s ≤ .05)[6] and a much smaller impact compared to CR prevalence and CR severity. Fifth, the effects of CR across the three fit indices are quite variable. This suggests that the manner in which CR affects fit is a partial function of which model fit index is employed. Lastly, many of the interactions had meaningful effect sizes (i.e., η^2s ≥ .01) thus further underscoring the complex manner in which CR affects fit. To better understand these interactions, a visualization of each meaningful interaction where $\eta^2 \geq .01$ is presented in Appendix G. Notably, both the “CR Prevalence × CR Severity” and “CR Severity × CR Type” interactions had small/moderate effect sizes across the three fit indices. For the CR Prevalence × CR Severity interaction, model fit deteriorates rapidly as CR prevalence increases and when CR severity is 20%/50% but remains mostly unchanged when prevalence increases and CR severity is 100% (Figure 8, interaction F; Figure 9, interaction B; Figure 10, interaction, A). For the CR Severity × CR Type interaction, model fit was often worse when CR was 50% (compared to either 20% or 100%) across CR

[6] For η^2, .01-.05 is considered a small effect, .06-13 a moderate effect, and > .14 a large effect (see Cohen, 1988).

types. This interaction also indicated that model fit is consistently worse for invariant responding across differing levels of CR severity compared to the other CR types (Figure 8, interaction H; Figure 9, interaction D; Figure 10, interaction C).

Table 10. ANOVA results summarizing IVs effects on IRT model fit indices

	M2 (IRT)		*RMSEA* (IRT)		*SRMSR* (IRT)	
Condition	Model *F*	η^2	Model *F*	η^2	Model *F*	η^2
Sample Size	1653.00*	.02	7.24*	.00	1385.90*	.03
Number of Items	31398.20*	.44	224.31*	.01	1264.90*	.03
CR Prevalence	1614.80*	.05	1560.49*	.11	4647.70*	.20
CR Severity	2260.00*	.06	2707.03*	.19	4481.10*	.20
CR Type	348.60*	.02	145.84*	.02	689.20*	.05
Sample Size × Number of Items	1616.90*	.02	18.36*	.00	141.20*	.00
Sample Size × CR Prevalence	247.40*	.01	6.67*	.00	35.10*	.00
Number of Items × CR Prevalence	1576.00*	.04	7.49*	.00	28.20*	.00
Sample Size × CR Severity	348.10*	.01	7.83*	.00	88.70*	.00
Number of Items × CR Severity	2217.90*	.06	436.60*	.03	31.60*	.00
CR Prevalence × CR Severity	297.10*	.02	280.85*	.04	868.10*	.08
Sample Size × CR Type	74.10*	.00	5.57*	.00	21.10*	.00
Number of Items × CR Type	339.40*	.01	17.73*	.00	12.10*	.00
CR Prevalence × CR Type	87.60*	.00	31.09*	.01	115.10*	.02
CR Severity × CR Type	335.10*	.03	435.34*	.09	619.40*	.08
Overall Model	1430.20*	-	353.29*	-	786.60*	-

Note. $^*p < .001$.

Next, to better understand the main effects reported in Table 10, a series of Tukey post-hoc tests were conducted. Cohen's *d* values were also computed for these comparisons in order to assess the magnitude of these effects. Much like the computation of η^2 above, the calculation of Cohen's *d* for each comparison can counteract the fact that most comparisons are likely to be significant due to the large sample size. For the "Sample Size" IV, the results indicated that *M2* was much smaller in the condition with 200 participants compared to the condition with 500 participants (*d* = 0.73). Furthermore, the *SRMSR* index was higher for the condition with 200 participants compared to the one with 500 participants (*d* = 0.62). The *RMSEA* index was not notably affected by the number of participants, though (*d* = 0.04; see Table 11).

Table 11. Tukey post-hoc tests examining the effects of sample size on IRT model fit

Comparison				
Sample Size	Sample Size	Mean Difference	*t* value	Cohen's *d*
M2				
200 Participants → (*M* = 1502.00)	500 Participants (*M* = 2385.00)	-883.00	-40.70**	0.68
RMSEA				
200 Participants → (*M* = .05)	500 Participants (*M* = .05)	.002	2.69*	0.04
SRMSR				
200 Participants → (*M* = .08)	500 Participants (*M* = .07)	.01	37.20**	0.62

Note. **p* < .01, ***p* < .001.

For the "Number of Items" IV, the Tukey post-hoc test results indicated that model fit is consistently better for the 10-item condition compared to the 60-item condition for the *M2* index (*d* = 2.95), *RMSEA* index (*d* = 0.25) and *SRMSR* index (*d* = 0.59). Because *M2* is a function of the number of items (in addition to sample size), this particular finding is not too surprising. Additionally, while fit is better when there are fewer items, this effect is larger for the *SRMSR* index (*d* = 0.59) compared to the *RMSEA* index (*d* = 0.25; see Table 12).

Table 12. Tukey post-hoc tests examining the effects of item number on IRT model fit

Comparison				
Number of Items	Number of Items	Mean Difference	*t* value	Cohen's *d*
M2				
10 Items (*M* = 18.60)	→ 60 Items (*M* = 3867.60)	-3849.00	-177.00*	2.95
RMSEA				
10 Items (*M* = .04)	→ 60 Items (*M* = .05)	-.01	-15.00*	0.25
SRMSR				
10 Items (*M* = .07)	→ 60 Items (*M* = .08)	-.01	-35.60*	0.59

Note. **p* < .001.

For the “CR Prevalence” IV, all Tukey post-hoc comparisons were significant for each IRT model fit index examined. More importantly, every Cohen’s *d* value was ≥ 0.53, which demonstrates that every comparison displayed at least a moderate effect size. The largest effect was for the 5%-30% comparison for the *SRMSR* index such that model fit was better when CR prevalence was 5% compared to 30% (*d* = 1.95). The smallest effect was for the 15%-30% comparison for the *M2* index (*d* = 0.53). The pattern of these results is very consistent across model fit indices. These results clearly show that IRT model fit tends to decrease as the prevalence of CR increases (see Table 13).

Table 13. Tukey post-hoc tests examining the effects of CR prevalence on IRT model fit

Comparison				
CR Prevalence	CR Prevalence	Mean Difference	*t* value	Cohen’s *d*
M2				
5% Prevalence (*M* = 1166.00)	→ 15% Prevalence (*M* = 1988.00)	-823.00	-30.90*	0.63
	→ 30% Prevalence (*M* = 2675.00)	-1510.00	-56.81*	1.16
15% Prevalence (*M* = 1988.00)	→ 30% Prevalence (*M* = 2675.00)	-687.00	-25.80*	0.53
RMSEA				
5% Prevalence (*M* = .03)	→ 15% Prevalence (*M* = .05)	-.02	-33.85*	0.69
	→ 30% Prevalence (*M* = .07)	-.04	-55.40*	1.13
15% Prevalence (*M* = .05)	→ 30% Prevalence (*M* = .07)	-.02	-21.60*	0.44
SRMSR				
5% Prevalence (*M* = .05)	→ 15% Prevalence (*M* = .08)	-.02	-57.70*	1.16
	→ 30% Prevalence (*M* = .09)	-.04	-95.80*	1.95
15% Prevalence (*M* = .08)	→ 30% Prevalence (*M* = .09)	-.01	-38.90*	0.79

Note. *$p < .001$.

For the “CR Severity” IV, the results of the Tukey post-hoc tests indicated that all the comparisons were significant, with the majority of comparisons indicating moderate/large effects. The largest effect was for the 20%-50% comparison for the *SRMSR* index (*d* = 1.84) such that model fit was better when CR severity was 20% compared to when it was 50%. The smallest effect was for the 20%-50% comparison for the *RMSEA* index (*d* = 0.08) such that model fit was better when CR severity was 20% compared to when it was 50%. The results of this analysis revealed some other interesting findings. First, the effects of CR severity on fit are not entirely consistent across model fit indices. For example, fit is worse for *RMSEA* when CR severity is 20% compared to when CR severity is 100%. In contrast, fit is better for the *SRMSR* index when CR severity is 20% compared to when CR severity is 100%. Second, model fit is consistently worse when CR severity is 50% than when it is 100% across all three IRT fit indices. This somewhat counterintuitive finding suggests that partial CR results in worse IRT model fit compared to complete CR. Table 14 summarizes these findings.

Table 14. Tukey post-hoc tests examining the effects of CR severity on IRT model fit

Comparison				
CR Severity	CR Severity	Mean Difference	*t* value	Cohen’s *d*
M2				
20% Severity (*M* = 2236.00)	→ 50% Severity (*M* = 2654.00)	-418.00	-15.70*	0.32
	→ 100% Severity (*M* = 939.00)	1297.00	48.85*	0.99
50% Severity (*M* = 2654.00)	→ 100% Severity (*M* = 939.00)	1715.00	64.50*	1.32
RMSEA				
20% Severity (*M* = .06)	→ 50% Severity (*M* = .06)	.002	3.92*	0.08
	→ 100% Severity (*M* = .02)	.04	65.59*	1.34
50% Severity (*M* = .06)	→ 100% Severity (*M* = .02)	.04	61.67*	1.26
SRMSR				
20% Severity (*M* = .06)	→ 50% Severity (*M* = .09)	-.03	-90.00*	1.84
	→ 100% Severity (*M* = .07)	-.01	-19.70*	0.40

CR Type	CR Type	Mean Difference	*t* value	Cohen's *d*
50% Severity (*M* = .09)	→ 100% Severity (*M* = .07)	.02	70.30*	1.44

Note. **p* < .001.

Lastly, for the "CR Type" IV, the Tukey post-hoc test results revealed many significant differences between CR types across the three IRT model fit indices. The largest effect was for the mathematical random responding-invariant responding comparison for the *SRMSR* fit index such that that fit was worse for invariant responding compared to true random responding (*d* = 1.01). The smallest (significant) effect was for the CR-mathematically responding comparison for the *RMSEA* index such that fit was worse for the invariant conditions (*d* = 0.10). These results also show that invariant responding is consistently associated with the worse model fit across fit indices. Also, the results indicate that person-based random responding and true mathematical randomly responding often have different effects on model fit, though these effects tend to be quite small. This provides some tentative support that these two types of CR may not in fact be equivalent (see also below results for CFA model fit).

Table 15. Tukey post-hoc tests examining the effects of CR type on IRT model fit

Comparison				
CR Type	CR Type	Mean Difference	*t* value	Cohen's *d*
M2				
Careless (*M* = 1775.00)	→ Person Random (*M* = 1835.00)	-60.40	-1.97	0.05
	→ True Random (*M* = 1625.00)	149.30	4.86*	0.11
	→ Invariant (*M* = 2537.00)	-762.10	-24.81*	0.58
Person Random (*M* = 1835.00)	→ True Random (*M* = 1625.00)	209.70	6.83*	0.16
	→ Invariant (*M* = 2537.00)	-701.70	-22.84*	0.54
True Random (*M* = 1625.00)	→ Invariant (*M* = 2537.00)	-911.40	-29.67*	0.70
RMSEA				
Careless (*M* = .05)	→ Person Random (*M* = .05)	-.001	-2.30	0.05
	→ True Random (*M* = .05)	.003	4.28*	0.10

	→ Invariant (M = .06)	-.01	-15.53*	0.37
Person Random (M = .05)	→ True Random (M = .05)	.005	6.58*	0.16
	→ Invariant (M = .06)	-.01	-13.23*	0.31
True Random (M = .05)	→ Invariant (M = .06)	-.01	-19.80*	0.47
SRMSR				
Careless (M = .08)	→ Person Random (M = .07)	.01	14.03*	0.33
	→ True Random (M = .07)	.01	23.54*	0.56
	→ Invariant (M = .09)	-.01	-19.17*	0.45
Person Random (M = .07)	→ True Random (M = .07)	.004	9.51*	0.22
	→ Invariant (M = .09)	-.02	-33.20*	0.78
True Random (M = .07)	→ Invariant (M = .09)	-.02	-42.71*	1.01

Note. *p < .001.

Careless Responding and CFA Model Fit. To understand how the IVs affect the model fit of CFA models, another series of factorial ANOVAs were conducted. Aside from examining different DVs, these ANOVS were set up in the same manner as those described above. As seen in Table 16, the overall models for the χ^2 (F = 7257.90, p < .001), *RMSEA* (F = 543.62, p < .001) and *SRMSR* (F = 935.60, p < .001) CFA model fit indices were significant. There are several notable aspects of these findings. First, "number of items" has the largest impact on the χ^2 index (η^2 = .77). This is not surprising since this fit index is a function of the number of items. Second, CR prevalence (η^2 = .13), CR severity (η^2 = .13), and CR type (η^2 = .13) have an equal impact on *RMSEA.* Third, CR severity (η^2 = .21) and CR prevalence (η^2 = .18) had the largest impact on the *SRMSR* index, thereby again showing that the IVs impact different model fit indices in unique ways. Fourth, there were quite a few interactions that displayed meaningful effect sizes (i.e., η^2s > .01). A visualization of every interaction where $\eta^2 \geq .01$ is presented in Appendix H. Notably, the "CR Prevalence × CR Type" interaction consistently displayed small/moderate effects across the three CFA indices. These interactions consistently show that while increases in CR prevalence are associated with decreases in model fit across all CR types, this increase is more prominent for invariant responding compared to the other CR types (Figure 11, interaction F; Figure 12, interaction F; Figure 13, interaction B).

There are also some notable ways in which these results are similar to/different from the results of the IRT model fit indices. First, CR type has much larger effect on *RMSEA* for CFA model fit (η^2 = .13) than IRT model fit (η^2 = .02). Second, the number of items affects χ^2 (η^2 = .77) much more than it affects *M2* (η^2 = .44), though both effects are very large. Third, the *SRMSR* fit index is similarly impacted by CR prevalence (CFA η^2 = .18; IRT η^2 = .20) and CR severity (CFA η^2 = .21; IRT η^2 = .20) across CFA and IRT models. Fourth, while there was a lot of overlap in the main effects and interactions that emerged across CFA and IRT models, there were still differences present. This affirms that these IVs differentially impact CFA and IRT models. Put another way, the effects of CR appear to be a function of the latent variable model employed.

Table 16. ANOVA results summarizing IVs effects on CFA model fit indices

	χ^2 (CFA)		*RMSEA* (CFA)		*SRMSR* (CFA)	
Condition	Model *F*	η^2	Model *F*	η^2	Model *F*	η^2
Sample Size	7876.20*	.03	218.64*	.01	973.30*	.02
Number of Items	233759.70*	.77	2883.13*	.08	3687.30*	.07
CR Prevalence	3356.00*	.02	2252.06*	.13	4719.40*	.18
CR Severity	2168.10*	.01	2384.07*	.13	5408.30*	.21
CR Type	3003.50*	.03	1605.87*	.13	1047.10*	.06
Sample Size × Number of Items	5835.20*	.02	235.20*	.01	201.90*	.00
Sample Size × CR Prevalence	514.40*	.00	0.13	.00	17.10*	.00
Number of Items × CR Prevalence	2547.00*	.02	335.52*	.02	46.30*	.00
Sample Size × CR Severity	241.30*	.00	29.65*	.00	84.50*	.00
Number of Items × CR Severity	1571.80*	.01	478.68*	.03	33.50*	.00
CR Prevalence × CR Severity	133.70*	.00	88.20*	.01	897.50*	.07
Sample Size × CR Type	522.20*	.01	10.72*	.00	29.80*	.00
Number of Items × CR Type	2320.00*	.02	199.62*	.02	24.30*	.00
CR Prevalence × CR Type	566.60*	.01	218.51*	.04	187.00*	.02
CR Severity × CR Type	84.70*	.00	67.37*	.01	662.70*	.08
Overall Model	7257.90*	-	543.62*	-	935.60*	-

Note. *p < .001.

Similar to above, a series of Tukey post-hoc tests were conducted for each IV and the three CFA model fit indices. For the “Sample Size” variable, the results unsurprisingly indicated that χ^2 was much lower when there were 200 participants compared to when there were 500 participants (*d* = 1.48). The results also indicated the *RMSEA* and *SRMSR* fit indices were higher when there were 200 participants compared to when there were 500 participants (Table 17).

Table 17. Tukey post-hoc tests examining the effects of sample size on CFA model fit

Comparison				
Sample Size	Sample Size	Mean Difference	*t* value	Cohen’s *d*
χ^2				
200 Participants → (*M* = 1797.00)	500 Participants (*M* = 2537.00)	-740.00	-88.70*	1.48
RMSEA				
200 Participants → (*M* = .08)	500 Participants (*M* = .07)	.01	14.80*	0.25
SRMSR				
200 Participants → (*M* = .08)	500 Participants (*M* = .07)	.01	31.20*	0.52

Note. *$p < .001$.

For the “Number of Items” IV, the results indicated that χ^2 was far higher when there were 60 items compared to when there were 10 (*d* = 8.06). Also, *RMSEA* was higher when there were 10 items compared to when there were 60 (*d* = 0.90). In contrast, *SRMSR* values were lower when there 10 items compared to when there 60 (*d* = 1.01; see Table 18).

Table 18. Tukey post-hoc tests examining the effects of item number on CFA model fit

Comparison				
Number of Items	Number of Items	Mean Difference	*t* value	Cohen’s *d*
χ^2				
10 Items (*M* = 152.00)	→ 60 Items (*M* = 4182.00)	-4030.00	-56.00*	8.06
RMSEA				
10 Items (*M* = .08)	→ 60 Items (*M* = .06)	.02	53.70*	0.90
SRMSR				
10 Items (*M* = .06)	→ 60 Items (*M* = .08)	-.02	-60.70*	1.01

Note. *$p < .001$.

For the “CR prevalence” IV, the results of the Tukey post-hoc test indicated that every comparison was significant and displayed moderate/large effects. The largest effect was for the 5%-30% comparison for the *SRMSR* index such that fit was better when prevalence was 5% compared to when it was 30% (*d* = 1.97). The smallest effect was for the 15%-30% comparison for the *RMSEA* index such that fit was better when prevalence was 15% compared to when it was 30% (*d* = 0.46). Like the above results, these results demonstrate that increases in CR prevalence are associated with decreases in model fit. This is consistent across all three CFA fit indices (see Table 19).

Table 19. Tukey post-hoc tests examining the effects of CR prevalence on CFA model fit

Comparison				
CR Prevalence	CR Prevalence	Mean Difference	*t* value	Cohen’s *d*
χ^2				
5% Prevalence (M = 1718.00)	→ 15% Prevalence (M = 2236.00)	-518.00	-50.70*	1.04
	→ 30% Prevalence (M = 2546.00)	-828.00	-81.10*	1.66
15% Prevalence (M = 2236.00)	→ 30% Prevalence (M = 2546.00)	-310.00	-30.30*	0.62
RMSEA				
5% Prevalence (M = .06)	→ 15% Prevalence (M = .08)	-.02	-43.50*	0.89
	→ 30% Prevalence (M = .09)	-.03	-66.00*	1.35
15% Prevalence (M = .08)	→ 30% Prevalence (M = .09)	-.01	-22.50*	0.46
SRMSR				
5% Prevalence (M = .05)	→ 15% Prevalence (M = .07)	-.02	-58.30*	1.19
	→ 30% Prevalence (M = .08)	-.03	-96.50*	1.97
15% Prevalence (M = .07)	→ 30% Prevalence (M = .08)	-.01	-38.20*	0.78

Note. *p < .001.

For the “CR severity” IV, the results of the Tukey post-hoc tests revealed that nearly every comparison displayed moderate/large effects. The largest effect was for the 20%-50%

comparison for the *SRMSR* fit index such that fit was worse when severity was 50% compared to when it was 20% (*d* = 1.99). The smallest effect was for the 50%-100% comparison for the *RMSEA* fit index such that fit was worse when severity was 50% compared to when it was 100% (*d* = 0.29). Like the above results for the IRT model fit indices, these results again show that model fit is worse when CR severity is 50% compared to when it is 100%. Furthermore, this is consistent across all three CFA model fit indices. The results also show that when CR severity is 20%, model fit is consistently better than when CR severity is either 50% or 100% (Table 20).

Table 20. Tukey post-hoc tests examining the effects of CR severity on CFA model fit

Comparison				
CR Severity	CR Severity	Mean Difference	*t* value	Cohen's *d*
χ^2				
20% Severity (M = 1790.00)	→ 50% Severity (M = 2434.00)	-644.00	-63.10*	1.29
	→ 100% Severity (M = 2278.00)	-488.00	-47.80*	0.98
50% Severity (M = 2434.00)	→ 100% Severity (M = 2278.00)	156.00	15.30*	0.31
RMSEA				
20% Severity (M = .06)	→ 50% Severity (M = .09)	-.03	-65.80*	1.34
	→ 100% Severity (M = .08)	-.02	-51.10*	1.04
50% Severity (M = .09)	→ 100% Severity (M = .08)	.01	14.70*	0.29
SRMSR				
20% Severity (M = .05)	→ 50% Severity (M = .09)	-.04	-97.40*	1.99
	→ 100% Severity (M = .06)	-.01	-17.20*	0.35
50% Severity (M = .09)	→ 100% Severity (M = .06)	.03	80.30*	1.64

Note. *p < .001.

Lastly, for the "CR type" IV, the Tukey post-hoc test results again indicated that all comparisons were significant. Compared to the above results of the IRT model fit indices, the effect sizes of these comparisons tended to be slightly larger. This suggests that CR type might have a greater impact on the fit of CFA models than the fit of IRT models. The largest effect observed was for the mathematically random responding-invariant responding comparison for the *RMSEA* fit index such that fit was worse for invariant responding compared to mathematical random responding (*d* = 2.08). In fact, the largest effects were typically observed for comparisons involving invariant responding which again shows that this form of CR tends to have the largest (i.e., worse) impact on model fit. The smallest effect was for the person random-mathematically randomly responding comparison for the *RMSEA* index such that fit was higher for person-based random responding compared to mathematically random responding condition (*d* = 0.32). These results also show that there are consistent differences between person-based random responding and true mathematical randomness across the three CFA fit indices, though these effects are small/moderate. Specifically, model fit is consistently worse for person-based random responding compared to mathematically random responding. This, coupled with the above findings regarding the differential effect of these two CR types, provides additional evidence that these careless response behaviors might not necessarily be equivalent (Table 21).

Table 21. Tukey post-hoc tests examining the effects of CR type on CFA model fit

Comparison				
CR Type	CR Type	Mean Difference	*t* value	Cohen's *d*
χ^2				
Careless (*M* = 2124.00)	→ Person Random (*M* = 1948.00)	177.00	15.00*	0.35
	→ True Random (*M* = 1777.00)	347.00	29.40*	0.69
	→ Invariant (*M* = 2818.00)	-694.00	-58.90*	1.39
Person Random (*M* = 1948.00)	→ True Random (*M* = 1777.00)	170.00	14.50*	0.34
	→ Invariant (*M* = 2818.00)	-871.00	-73.90*	1.74
True Random (*M* = 1777.00)	→ Invariant (*M* = 2818.00)	-1041.00	-88.30*	2.08
RMSEA				

Careless (*M* = .08)	→ Person Random (*M* = .07)	.01	15.00*	0.35
	→ True Random (*M* = .06)	.02	28.90*	0.68
	→ Invariant (*M* = .10)	-.02	-36.90*	0.87
Person Random (*M* = .07)	→ True Random (*M* = .06)	.01	13.90*	0.32
	→ Invariant (*M* = .10)	-.03	-51.90*	1.22
True Random (*M* = .06)	→ Invariant (*M* = .10)	-.04	-65.80*	1.55
SRMSR				
Careless (*M* = .07)	→ Person Random (*M* = .07)	.004	11.60*	0.28
	→ True Random (*M* = .06)	.01	29.50*	0.70
	→ Invariant (*M* = .08)	-.01	-25.10*	0.59
Person Random (*M* = .07)	→ True Random (*M* = .06)	.01	17.80*	0.42
	→ Invariant (*M* = .08)	-.02	-36.80*	0.86
True Random (*M* = .06)	→ Invariant (*M* = .08)	-.02	-54.60*	1.29

Note. *p < .001.

Variability of Model Fit Indices. While the above ANOVAs account for the within-level variability of the model fit indices by including all 100 replications within each condition, the amount of within-condition variability is not explicitly apparent (but see Tables 35 through 38 in Appendix F for the average variability of model fit for all conditions). Table 22 displays the average variability of the model fit indices broken down by the sample size/item number conditions. These results show there is high variability (*SD*) for the *M2* and χ^2 fit indices, especially when there are 60 items. These results also show there is high variability for the *RMSEA* and *SRMSR* indices, though this variability is higher when there are 10 items compared to when there are 60 items. Collectively, the results from Table 22 and Appendix F indicate that even within an identical circumstance, there can be considerable variability in how CR affects model fit. The implications of this finding are expanded upon in the discussion section below.

Table 22. Average variability of model fit indices across baseline conditions

	IRT Model Fit *SD*			CFA Model Fit *SD*		
Condition	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
200 Participants; 10 Items	10.43	.04	.02	38.99	.03	.02
200 Participants; 60 Items	542.64	.01	.01	311.87	.01	.01
500 Participants; 10 Items	22.76	.04	.02	103.22	.03	.02
500 Participants; 60 Items	1028.23	.01	.01	550.70	.01	.01

Note. Values represent mean values averaged across both conditions and 100 replications (see Appendix F for *SD* values for all fit indices across all 144 conditions).

Careless Responding and Model Fit Bias

Model fit bias was next computed for each of the 144 conditions. Values were again averaged across the 100 replications for each condition. Model fit bias represents the model fit of the contaminated (Step 3 in Figure 3) data subtracted from the model fit of the baseline/careful data (Step 1 in Figure 3) and thus provides useful information concerning the magnitude of the effect that CR has on model fit relative to the uncontaminated fit (i.e., the fit of entirely careful responses). In other words, model fit bias connotates how much fit changes when CR is introduced into to the dataset (which occurs in Step 3 in Figure 3). Because model fit and bias are closely related, a detailed account of model fit bias will not be provided here as this would be highly redundant with the previous section. Instead, a basic summary is provided. A comprehensive summary of model fit bias for all 144 conditions is provided in Appendix I.

To illustrate how the various IVs impact model fit – and also to illustrate the relation between model fit and model fit bias – a visual representation of the five conditions with the highest bias and the five conditions with the lowest bias is presented in Figure 6 below and in Appendix J. Results are broken down by sample size and number of items conditions for all six model fit indices.[7] Only the results for the condition with 200 participants and 10 items are reported here due to the high degree of similarity between the sample size/number of items conditions. The remaining visualizations are provided in Appendix J. These visualizations illustrate several other important findings. First, bias tends to be largest (i.e., fit deteriorates the most) for invariant responding, when CR prevalence is 30% and when CR severity is either 20% or 50%. Interestingly, model fit rarely deteriorates (i.e., bias is unchanged) when severity is 100%, thus further demonstrating that partial CR is worse than full CR. Second, there are some combinations of conditions where the fit of the contaminated data is equal to/better than the fit of the careful/baseline data (e.g., *M2* for 200 participants/10 items [upper left portion of Figure 6]). This tends to occur when CR prevalence is low and, surprisingly, when CR severity is 100% across different types of CR (see above example and Appendix J for details). Thus, CR is sometimes unrelated to fit and/or associated with improvements in model fit. This does, however, again appear to be a function of the model fit index and latent variable model (i.e., CFA vs. IRT) that is employed.

[7] Model fit bias is not aggregated across all conditions since *M2* and χ^2 are heavily influenced by the number of items. For instance, as seen in Appendix I, the average *M2* values for the condition with 10 items is about 10.00 while the average χ^2 value for the condition with 10 items is around 100. When there are 60 items, however, these values increase by thousands. Thus, examining bias by aggregating across all conditions (rather than across the sample size/number of items conditions) is not appropriate.

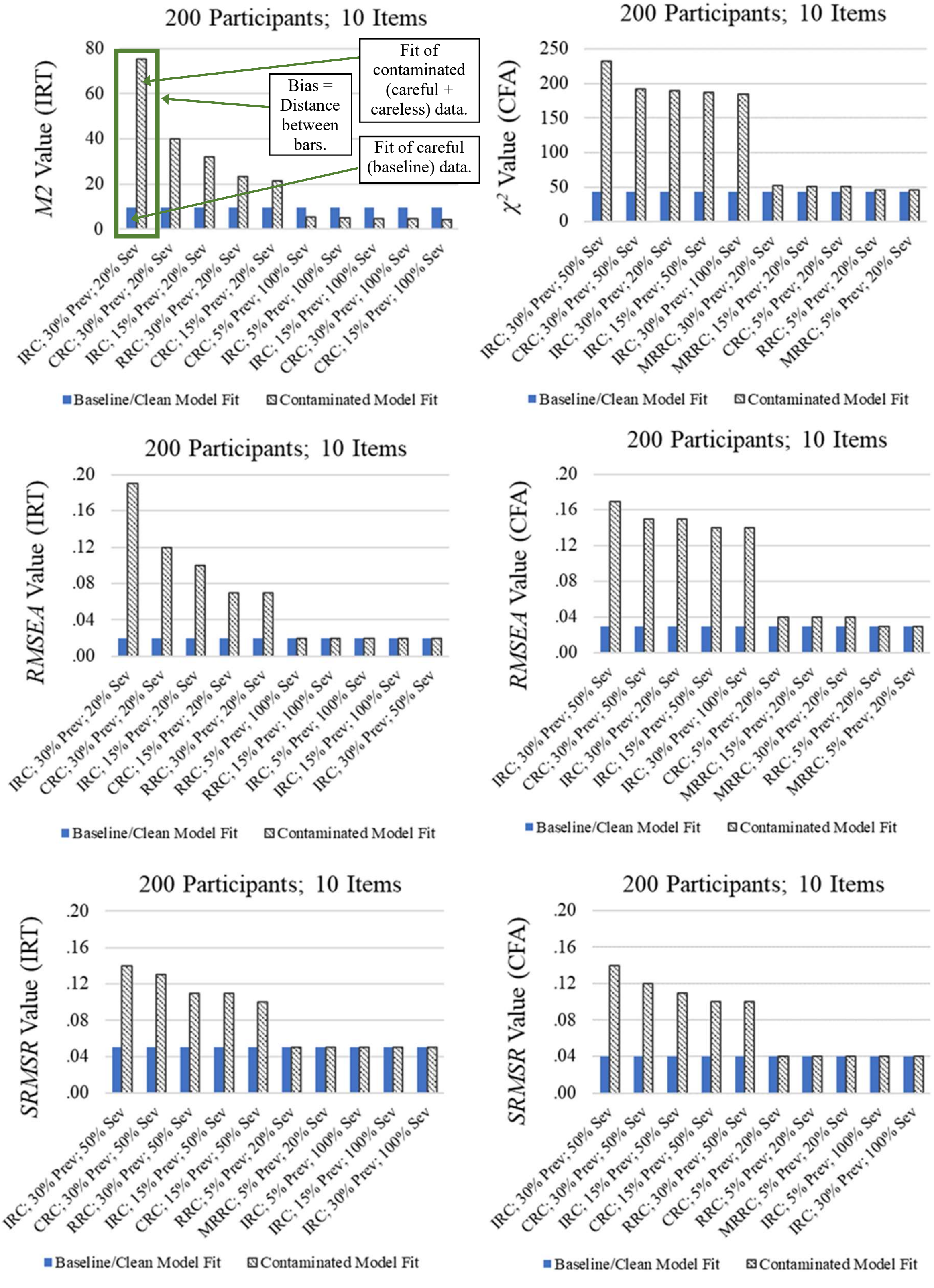


Figure 6. Visual representation of model fit/bias (200 participants/10 items)

Summary of Research Questions

The primary goal of this study was to answer the following questions: (1) "What are the consequences of CR on the fit of both CFA and IRT latent variable models across a range of different conditions?" and (2) "To what extend are CR and mathematically random responding analogous response behaviors?". Regarding the first research question, the results indicate that CR tends to worsen model fit under most circumstances. Specifically, across the 144 conditions examined, *M2* worsened in 79% of cases, IRT *RMSEA* worsened in 84% of cases, IRT *SRMSR* worsened in 95% of cases, χ^2 worsened in 100% of cases, CFA *RMSEA* worsened in 97% of cases, and CFA *SRMSR* worsened in 93% of cases. As the results also indicate, these effects are very complex and a function of different IVs, combinations of IVs, model fit indices, and latent variable models. For example, CFA *RMSEA* deteriorated rapidly in the condition with 200 items, 10 items, 30% prevalence, 20% severity, and careless responding. When mathematically random responding was examined, and all other variables were the same, the CFA *RMSEA* model fit index was, however, hardly impacted. This study revealed many such interesting findings.

Regarding the second research question, the results tentatively suggest that mathematical randomness and person-based randomness are not necessarily analogous response behaviors, despite this being a common assumption. For example, these two types of CR consistently had differential effects on model fit across different latent variable models and fit indices. If these two different types of CR represent different response patterns, it is difficult to determine why they have differential effects on fit. In fact, all four types of CR examined in this simulation frequently had unique effects on model fit. This suggests that there are many ways in which CR can manifest, with each manifestation having its own unique set of consequences. The implications of these findings are further discussed below.

Chapter 4 - General Discussion

The primary goal of this study was to examine the consequences of CR on the fit of both IRT and CFA latent variable models across various conditions using a highly realistic simulation paradigm. This study also sought to further elucidate the nature of CR and determine the extent to which CR and random responding are similar response behaviors. To accomplish this, a multiphase simulation approach was employed. Specifically, participants' response behaviors were first experimentally manipulated in Study 1. The results from this study were then used in Study 2 as a basis for the primary simulation study. Overall, the results from this study provide many important insights regarding the nature and consequences of CR.

First, this study indicates that while CR is generally associated with decreases in model fit, the effects of CR on model fit are highly nuanced and complex, contingent on many factors, and quite variable. This general finding is important and can potentially help explain some of the inconsistent findings reported in the literature. For example, previous studies have found that CR is related to poorer model fit (e.g., Arias et al., 2020; Woods, 2006), improved model fit (e.g., Goldammer et al., 2020), and/or unrelated to model fit (e.g., Beck et al., 2019; Liu et al., 2019). As the results of this study indicate, the effects of CR on model fit are contingent on many factors (i.e., sample size, number of items, type of CR, severity of CR, prevalence of CR). Thus, in one circumstance, CR might have negligible effects on model fit while in another circumstance, the effects can be quite substantial. In addition, the results of this study show that there is not only considerable variability *between* circumstances (i.e., conditions), but that there is some variability *within* a given circumstance (see Appendix F). Thus, even within an identical situation (i.e., when there is the same number of items, participants, CR severity, etc.), the effects of CR on model fit are likely to be somewhat variable from one dataset to the next. Given this high amount of between-level and within-level variability regarding the CR-model fit linkage, it is not surprising that the effects of CR on model fit that have been reported within the literature are also highly variable. Unfortunately, this finding indicates that the way CR affects model fit might not be easily predictable in non-simulation contexts where the nature of responses (i.e., whether responses are careful/valid or careless/invalid) cannot be known with certainty and where replications cannot be used to specify aggregate effects. Consequently, it is crucial to proactively address CR throughout all phases of data collection in order to help ensure that any effects of CR on model fit will be minimal (see also Aguinis, & Vandenberg, 2014; DeSimone et al., 2015; Meade & Craig, 2012).

This study revealed many other notable findings regarding the CR-model fit linkage. First, invariant responding was frequently associated with the largest deteriorations in model fit compared to other types of CR. This was true regardless of the condition, latent variable model, or model fit index that was examined. Second, while model fit consistently worsened as CR prevalence increased, model fit did not consistently worsen as CR severity increased. Somewhat counterintuitively, model fit was often worse when severity was 50% compared to when it was either 20% or 100%. Thus, partial CR appears to have worse consequences on model fit compared to complete CR. Furthermore, even when just 5% of the sample and only 20% of participants' responses are careless, model fit still tends to notably deteriorate in many circumstances. This suggests that even minimal amounts of CR are capable of reducing model fit (e.g., Arias et al., 2020). Third, while CR tended to almost always worsen fit, the values of the model fit indices often remained below the thresholds for what constitutes poor fit for the *RMSEA* and *SRMSR* indices (see Hu & Bentler, 1999; but see also Nye & Drasgow, 2011). While this is partially due to the low initial baseline model fit values (i.e., the fit of the careful/uncontaminated models was exceptionally good), this does suggest that seemingly good model fit values can obscure the impact of CR. For example, this study shows that acceptable fit values can often be attained even when up to 30% of a sample engages in CR. For other situations where the baseline/uncontaminated fit values are closer to the thresholds that researchers commonly use (e.g., *RMSEA* is .07 rather than .02), such models are much more likely to be rejected since CR is more likely to cause the model fit indices to exceed such thresholds. Fourth, in addition to showing how sample size, number of items, CR prevalence, CR severity, and CR type impact fit in unique and interactive ways, this study also indicates the manner in which CR affects fit is also a function of the latent variable model and fit indices that are employed. Furthermore, model fit indices and latent variable models do not always align with each other. Thus, CR may substantially impact one fit index, but fail to impact a different fit index within an identical situation. To fully understand how CR affects model fit, one must, therefore, also be mindful of the latent variable model and fit indices that are being used. In fact, the present study indicated a few cases when model fit information was actually contradictory across fit indices and latent variable models. This further underscores the complex manner in which CR is capable of affecting model fit. This also affirms that when evaluating model fit, multiple indices must be employed as CR does not impact fit indices in a consistent manner.

In addition to elucidating how CR and model fit are related, this study also offers new, important insights concerning the nature of CR behavior, which is a surprisingly understudied topic. For example, the results of Study 1 indicated that when participants were instructed to respond carelessly, their response patterns were different from those that emerged when participants were instructed to respond randomly. This difference in random and careless responding was further corroborated in Study 2, as these two CR types also impacted model fit in unique ways. This finding has important implications for CR research. For example, there has been some question as to whether random responding is indeed analogous to CR (e.g., DeSimone et al., 2018; Meade and Craig, 2012). The results of this study suggest that CR and random responding, while both constituting careless behavior, are somewhat unique manifestations of survey carelessness. The results of the validity check items for Study 2 seem to suggest that careless behavior contains more invariant responses than random responding (as evidenced by the higher longstring values for the careless condition compared to either the person-based random or true random conditions). It will be important for future simulations to keep this distinction in mind when studying CR and recognize that random responding is not necessarily equivalent to CR, but merely a particular subset of the broader CR construct.

In addition to some of the differences that were observed between CR and random responding, Study 2 also indicated there were differences between random responding and true mathematical randomness, though these effects tended to be small/moderate in size. It is well-documented that people cannot consistently recognize and engage in truly random behavior (e.g., Nickerson, 2002). The results of this study are consistent with this view and provide some preliminary evidence that people's inability to enact true random behavior extends to survey responses. Future research will need to further examine the differences between participants' random behavior and true randomness in other contexts, however, as this study is only able to make such comparisons concerning model fit information. The findings from this study are nonetheless important given that the nature of CR is not fully specified. For example, the results from this study offer some novel insights into the nature and consequences of unique types of careless behaviors that can be used to help inform subsequent studies on CR. The findings from this study also offer many important theoretical and practical contributions.

Theoretical Contributions

The results of this study have some important implications for latent variable modeling and validity theory (e.g., Bollen, 2002; Borboom, 2008). Perhaps most importantly, the findings from this study reaffirm that model-data fit and response validity are unique issues. Validity refers to when a posited psychological attribute serves as a causal determinant for a set of observed indicators (i.e., survey responses; Borsboom et al., 2004) while model-data fit refers to the extent by which the observed values in a dataset correspond to those predicted by the statistical model (Kline, 2016). Although this study shows that CR often worsens fit, these empirical effects are too variable and not consistent enough for model fit to serve as a reliable proxy for CR/response validity. Even if CR was reliably associated with deteriorations in model fit, using fit to make inferences about CR/response validity would still not necessarily be appropriate from a conceptual standpoint.

This conceptual (and empirical) orthogonality between response validity and fit – particularity when interpreted within the context of the results of this study – can give rise to some undesirable theoretical scenarios within a latent variable model paradigm. For instance, as this study shows, good model fit can be attained enough when up to 30% of a sample is engaging in CR. It is debatable that such a latent variable model, even with good fit, can still be considered valid (i.e., it truly represents the data) since many responses lack validity. The failure to account for CR would, however, result in this very conclusion (i.e., that good fit implies a good model). In reality, good model fit is a necessary, but insufficient, feature of a valid latent variable model. Latent variable models must also have valid (i.e., not careless) responses underlying the model. Evaluating the validity of models solely based on fit is, therefore, not appropriate.

The results of this study also have some important implications for theory development and testing more generally. For instance, it would seem as though addressing CR is necessary to develop accurate and useful theories. If, for example, CR causes the model fit indices for a latent variable model to exceed the threshold for what constitutes good fit (i.e., the fit of the *RMSEA* index changes from .07 to .09 due to the presence of CR), one might conclude that the model is unsuitable (i.e., a false negative) when this poor fit is actually being driven by invalid responses underlying the model. When investigating poor fit, it is often common to utilize modification indices and/or other psychometric information to determine the sources of model misfit (Kline, 2016). As the results of this study show, poor fit may also be attributable to CR. Unfortunately, the results typically provided by a latent variable model (e.g., modification indices, factor loadings, fit indices, etc.) are not necessarily useful for directly assessing the validity of responses that comprise a latent variable model. Consequently, to avoid these Type II errors, model (mis)fit and CR must be separately accounted for and addressed.

As this study also shows, CR can sometimes improve model fit. Thus, the fit of a seemingly "good" model may be inflated by the presence of CR. In such a situation, if CR were to not be addressed, one would falsely conclude that the model is good (i.e., a false positive), when it in fact not. Thus, accounting for CR can also be useful for avoiding these Type I errors. While it is common to investigate the sources of misfit for poor-fitting models, the fit of good-fitting models is hardly ever scrutinized. Somewhat counterintuitively, the results of this study suggest that one should also investigate the reasons for good fit, as good fit may be driven by CR in certain circumstances. The fact that CR can alter models fit indices, coupled with the fact that CR can also alter other psychometric information as well as the correlations between substantive variables (e.g., DeSimone et al., 2018; Huang et al., 2015), affirms that CR needs to be proactively addressed for accurate theories to be developed and tested. In fact, the failure to address CR may also be related to replicability issues within psychology (see Curran, 2016), particularly with regard to the replicability of latent variable models. That is, depending on the extent to which CR exists, low levels of CR may in one circumstance cause one to accept a latent variable model but in another circumstance with higher CR, cause one to reject the same latent variable model, all other things being equal. The failure to adequately address CR is, therefore, likely to results in inaccurate findings and produce false theoretical conclusions.

To extend these theoretical implications even further, the results of this study suggest that the use of arbitrary model fit cutoffs (e.g., *RMSEA* and *SRMSR* < .08) does not actually prevent studies that are laden with CR from being published. As noted earlier, "good" model fit can often be attained despite high levels of CR prevalence and severity. Thus, it is very conceivable that CR underlies many of the "good-fitting" models reported within the literature. The publication of low-quality data that is laden with CR is not ideal for theory development and/or science, and as noted above, could be a contributing factor to replication issues within psychology. To combat this problem, addressing CR should become the standard practice for any situation involving survey data. Unfortunately, such efforts are rarely reported in great detail within the literature. Accordingly, it is unclear the extent to which researchers ensure survey data-quality prior to reporting research results. Recently, Flake and Fried (2020) have introduced the term, "questionable measurement practices", which refers to, "Decisions researchers make that raise doubts about the validity of the measures used in a study, and ultimately the validity of the final conclusion." (Flake & Fried, 2020, p. 458). The failure to account for CR seems to be consistent with this definition since the decision to not address CR can both conceptually and empirically undermine the validity of survey data and subsequent conclusions that follow. Addressing CR is, therefore, necessary for ensuring high-quality measurement practices and something that should become a transparent practice throughout psychology for theories to be properly developed.

Practical Contributions

The results of this study also have various practical implications for practitioners who are interested in attaining higher quality survey data. For example, this study highlights situations where CR is likely to be of most concern as well as the situations where the effects of CR are negligible. Such knowledge can be useful for helping practitioners identify the situations where strategies for reducing and addressing CR are needed. This can be particularly useful in high-stakes, organizational contexts where survey findings have the potential to meaningfully impact organizational practices and policies. For instance, the results of this study suggest that partial CR is more detrimental to model fit than full CR. Because CR becomes more likely as the length of a survey increases (e.g., Bowling et al., 2020; Gibson & Bowling, 2019), practitioners could consider using shorter surveys when CR is cause for concern, as CR will be less likely to occur in such instances. Likewise, if longer surveys are needed, practitioners could consider placing instructions and messages throughout the survey that encourage careful responding (e.g., Huang et al., 2012). This may help decrease the likelihood that a person's responses become more careless toward the latter portion of the survey (but see Gibson & Bowling, 2019). Moreover, the results of this study suggest that invariant responding is associated with the largest deteriorations in model fit. It would, therefore, be beneficial for these responses to be removed and for practitioners to determine when such responses are likely to occur so that their occurrence can be minimized. Fortunately, such responses are easy to identify. This recommendation to actively address CR is consistent with a large body of research on CR (e.g., DeSimone et al., 2015; Hong et al., 2020; Meade & Craig, 2012). This study builds on this research, however, by adding another compelling reason for why CR should be addressed (i.e., CR reduces model fit). Indeed, being able to address CR and ensure higher quality results is also likely to translate into more efficacious and justifiable data-driven decision making within organizations. Put another way, one's applied research results will be more defensible if poor-quality data can be ruled out as a confounding explanation for one's results.

The results of this study may also be used by practitioners in a broader sense. To illustrate, if practitioners are interested in gaining advanced insights into how CR may affect their data prior to data collection, the results of this study can be used as a source of guidance for what to expect in their own datasets. For example, if a practitioner is interested in how long surveys with small samples and high levels of CR affect their data, they could look to the conditions with 60 items, 200 participants, and 30% prevalence as a source of guidance about what to expect in this circumstance. Such information could be useful for planning interventions and determining when additional steps will be needed for assuring data quality.

As another example, a practitioner might be interested in addressing CR that is collected using a crowdsourced marketplace, such as MTurk or Qualtrics. Given concerns about CR in these circumstances (see Brühlmann et al., 2020), particularly the use of nonhuman algorithms that are used to generate responses (e.g., Dupuis et al., 2019; Dupuis et al., 2020), a practitioner could look at the results for the conditions with mathematically generated responses to better understand how computer-generated random responses might influence their data. In fact, using the results of this study in this way to plan research studies and gain advanced insights about how CR might impact model fit can be especially valuable given that this is a highly controlled and comprehensive simulation study based on realistic data and a total of 14,400 replications (which helps increase certainty in the results).

The results of this study largely support the view that model fit indices should not be the primary method for making inferences about CR. While CR often has a negative effect on model fit, this effect is still quite variable and contingent on too many factors for model fit to serve as a dependable proxy for CR. For example, *M2* and χ^2 are heavily influenced by the number of items and sample size, but hardly affected by CR prevalence, CR severity, or CR type. Thus, these indices are minimally sensitive to CR, and should not be used by practitioners to make inferences about CR. While the other fit indices examined in this study may function as better proxies for CR (e.g., *RMSEA*), it is still probably better for practitioners to utilize methods that are specifically designed to assess CR (e.g., Mahalanobis distance, longstring, etc.; see Curran, 2016). For instance, this study indicated that there is a lot of within-condition variance in model fit indices and that certain fit indices are positively related to CR. For these reasons, model fit is unlikely to reliably detect CR. Additionally, good model fit can be attained even when large amounts of CR are present (e.g., one could observe an *RMSEA* value of .07 even when 30% of the sample is engaging in CR). Consequently, model fit might not exceed the fit thresholds that constitute poor fit, even when high amounts of CR are present. Relying on model fit to make inferences about CR in such situations will likely result in retaining a lot of low-quality data and making incorrect conclusions about the quality of one's data. Thus, practitioners should account for misfit and CR, but do so separately. In fact, addressing CR should occur regardless of whether fit is poor or good. This is especially important for ensuring that good model fit is not actually being driven by CR, which as this study indicates, is sometimes a possibility.

A final practical contribution of this study concerns the IRT-based simulation method that was used. While various simulation studies of CR have been conducted (e.g., DeSimone et al., 2018; Hong et al., 2020; Meade & Craig, 2012) this is the first CR-related study to use real, experimentally-shaped response behaviors as a basis for a simulation. This study is, therefore, one of the most realistic, and thus generalizable, CR simulations that has been conducted (see Harwell et al., 1996). Researchers interested in conducting more realistic simulation studies on CR can adopt the method that was used in this study for their own research projects. Since the way CR manifests is still not fully specified, using this IRT-based data generation method can help researchers examine more complex response behaviors rather than just uniform distributions (i.e., true random responding) or other distributions with known properties (e.g., normal distributions). This is especially useful since this study clearly shows that there are many unique manifestations of CR, not all of which adhere to the assumptions of known distributions.

Limitations and Future Directions

Despite the many strengths and contributions of this study, some limitations and directions for future research must be noted. First, the generalizability of this simulation is somewhat limited by the parameters that were derived in Study 1 and used as a basis for the simulation in Study 2. Specifically, given the measure of conscientiousness that was used, the parameters used to inform the simulation were based on negatively skewed data. This is not too surprising, however, since personality and attitudinal measures frequently contain negatively skewed responses (e.g., Danner, Aichholzer, & Rammstedt, 2015). While the results of this study will likely generalize to other constructs that are also negatively skewed (e.g., job performance, job satisfaction, etc.), future research is needed to determine if the current findings generalize to situations with other parameter values and response distributions. It could be the case that the results of this study are partially attributable to the distribution of parameters underlying the IRT simulation procedure. Examining how different response distributions and model parameters affect model fit represents a fruitful avenue for future research.

Second, while various careless behaviors were experimentally shaped in Study 1, and in a manner consistent with how other response sets are manipulated (e.g., faking; Zickar et al., 2004), it is unclear if participants' experimentally shaped careless behavior reflects naturalistically occurring careless behavior that is not experimentally manipulated. That is, the way people engage in CR without explicit instructions might differ from how they engage in CR with explicit instructions. Determining if this is the case, however, can be very difficult since it can never be known with complete certainty if participants' responses are careful or careless. This is, unfortunately, an inherent limitation of all studies of CR.

Third, while the primary goal of this study was to comprehensively examine how CR affects model fit, there are still other IVs, and combinations of IVs, that could have been examined. For example, it is conceivable that different participants engage in different forms of CR (e.g., random responding versus invariant responding) within a given dataset (e.g., between-level effects). It may also be possible for a single respondent to engage in different forms of CR across their entire response vector (e.g., within-level effects). Although more complex configurations of the "CR Type" variable could have been examined in the present study, this would have drastically increased the number of total conditions and made the results very difficult to interpret. Accordingly, future research should examine more complex between-level and within-level combinations of CR types to more fully understand how CR affects model fit. In a similar manner, it would be useful for future research to examine other model fit indices (e.g., Comparative Fit Index [CFI], Akaike Information Criterion [AIC], etc.) and further assess the generalizability of findings from the current study. This would be especially useful since this study shows that the effects of CR on model fit are not entirely consistent across model fit indices. It would also be useful to examine more complex latent variable models (e.g., multidimensional and/or hierarchical models) or structural models that contain multiple constructs as well as specified paths between latent constructs.

Fourth, while not a limitation per se, there was considerable variability in the model fit index values reported within some of the conditions. This could potentially be due to the IRT simulation method that was used. For instance, each condition was replicated by randomly sampling 20 sets of parameters/thetas and generating five replications for each set of sampled parameters/thetas. Because parameters and thetas were randomly sampled from the same underlying distribution, however, this cannot fully account for this large variability. Consequently, it would be beneficial to conduct even more detailed simulation studies that examine how CR affects model fit. Such studies could build off the results reported here and offer a more nuanced account on how CR and model fit are related.

Lastly, although this study provides compelling evidence that CR causes deteriorations in model fit, as well as a comprehensive account of when this is likely to occur, this study cannot fully explain *why* this occurs. Although some explanations have been provided (e.g., CR increases measurement error; see Arias et al., 2020), it would be beneficial for future research to further elucidate the mediating mechanisms that explain the CR-model fit linkage. As the results of this study suggest, however, this will be challenging since the manner in which CR affects model fit is highly variable. Thus, the reason why CR affects model fit in a one circumstance might be different from why CR affects model fit in another circumstance.

Conclusion

Data is an integral aspect of modern society and surveys remain one of the most popular, useful, and efficient ways to gather data within academic and organizational settings. To fully leverage the data that are gathered with surveys, it is essential to ensure that the data are of high-quality. Unfortunately, research shows that many participants do not provide valid data and respond carelessly when completing surveys. The failure to address survey data that are laden with CR can lead to false theoretical conclusions and misinformed workplace decision making. Indeed, as this study shows, the presence of CR can be detrimental to data-quality and reduce model fit under many different circumstances. This is true across different latent variable models, model fit indices, and a variety of other circumstances. It is, therefore, essential to proactively address CR in all phases of the research life cycle. The failure to do so all but guarantees the inaccuracy, invalidity, and uselessness of one's survey results.

References

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, *53*, 63-70. https://doi.org/10.1016/j.jom.2017.06.001

Aguinis, H., & Vandenberg, R. J. (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior, 1*, 569-595. https://doi.org/10.1146/annurev-orgpsych-031413-091231

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, *52*, 2589-2505. https://doi.org/10.3758/s13428-020-01401-8

Arthur Jr., W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*, 435-442. https://doi.org/10.1037/0021-9010.93.2.435

Arthur Jr., W., Hagen, E. & George Jr., F. (2020). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior, 8,* 105-137. https://doi.org/10.1146/annurev-orgpsych-012420-055324

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education*, *53*, 182-200. https://doi.org/10.1007/s11162-011-9251-2

Baumgartner, H., Weijters, B., & Pieters, R. (2018). Misresponse to survey questions: A conceptual framework and empirical test of the effects of reversals, negations, and polar opposite core concepts. *Journal of Marketing Research*, *55*, 869-883. https://doi.org/10.1509/jmr.15.0117

Beck, M. F., Albano, A. D., & Smith, W. M. (2019). Person-Fit as an index of inattentive responding: A comparison of methods using polytomous survey data. *Applied Psychological Measurement*, *43*, 374-387. https://doi.org/10.1177/0146621618798666

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*, 478-494. https://doi.org/10.1037/0021-9010.74.3.478

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*, 317-335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605-634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203-219. https://doi.org/10.1037/0033-295x.110.2.203

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071. https://doi.org/10.1037/0033-295x.111.4.1061

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 25-53.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*, 218-229. https://doi.org/10.1037/pspp0000085

Bowling, N. A., & Huang, J. L. (2018). Your attention please! Toward a better understanding of research participant carelessness. *Applied Psychology*, *67*, 227-230. https://doi.org/10.1111/apps.12143

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C K. (2020). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods.* https://doi.org/10.1177/1094428120947794

Bradford, L. (2018, October, 11). Why all employees need data skills in 2019 (and beyond). Retrieved from: https://www.forbes.com/sites/laurencebradford/2018/10/11/ why-all-employees-need-data-skills-in-2019-and-beyond/#58815bbf510f

Brower, C. K. (2020). *What are you looking at? Using eye-tracking to provide insight into careless responding and its measurement*. (Unpublished doctoral dissertation). Wright State University.

Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, 2, 100022. https://doi.org/10.1016/j.metip.2020.100022

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3-5. https://doi.org/10.1037/e527772014-223

Bulut, O., & Sünbül, Ö. (2017). Monte Carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology*, *8*, 266-287.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1-29. https://doi.org/10.18637/jss.v048.i06

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*, 347-361. https://doi.org/10.1007/s10869-016-9458-5

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

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*, 596-612. https://doi.org/10.1177/0013164410366686

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19. https://doi.org/10.1016/j.jesp.2015.07.006

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119-130. https://doi.org/10.1016/j.jrp.2015.05.004

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, *90*, 70-76.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*, 171-181. https://doi.org/10.1002/job.1962

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology*, *67*, 309-338. https://doi.org/10.1111/apps.12117

DeSimone, J. A., & Harms, P. D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology*, *33*, 559-577. https://doi.org/10.1007/s10869-017-9514-9

DeSimone, J. A., Davison, H. K., Schoen, J. L., & Bing, M. N. (2020). Insufficient effort responding as a partial function of implicit aggression. *Organizational Research Methods*, *23*, 154-180. https://doi.org/10.1177/1094428118799486

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*, 105-113. https://doi.org/10.1177/001316446802800110

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. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, *33*, 105-121. https://doi.org/10.1007/s10869-016-9479-0

Dupuis, M., Meier, E., & Cuneo, F. (2019). Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavior Research Methods*, *51*, 2228-2237. https://doi.org/10.3758/s13428-018-1103-y

Dupuis, M., Meier, E., Gholam-Rezaee, M., Gmel, G., Strippoli, M. P. F., & Renaud, O. (2020). Detecting computer-generated random responding in online questionnaires: An extension of Dupuis, Meier & Cuneo (2019) on dichotomous data. *Personality and Individual Differences*, *157*, 109812. https://doi.org/10.1016/j.paid.2020.109812

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*, 155-174. https://doi.org/10.1037/1082-989x.5.2.155

Edwards, J. R. (2019). Response invalidity in empirical research: Causes, detection, and remedies. *Journal of Operations Management*, *65*, 62-76. https://doi.org/10.1016/j.jom.2018.12.002

Embretson S.E. & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Emons, W. H. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224-247. https://doi.org/10.1177/0146621607302479

Evans, J. R., & Mathur, A. (2018). The value of online surveys: A look back and a look ahead. *Internet Research*, 28, 854-887. https://doi.org/10.1108/intr-03-2018-0089

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*, 36-49. https://doi.org/10.1111/emip.12111

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*, 456-465. https://doi.org/10.31234/osf.io/hs7wm

Francavilla, N. M., Meade, A. W., & Young, A. L. (2019). Social interaction and internet-based surveys: Examining the effects of virtual and in-person proctors on careless response. *Applied Psychology*, *68*, 223-249. https://doi.org/10.1111/apps.12159

Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods*, *20*, 465-486. https://doi.org/10.1177/1094428116689708

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*, 349-360. https://doi.org/10.1093/poq/nfp031

Gibson, A. (2019). *Stop what you're doing, right now! Effects of interactive messages on careless responding* (Unpublished doctoral dissertation). Wright State University.

Gibson, A. M., & Bowling, N. A. (2019). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*, 36, 410-420. https://doi.org/10.1027/1015-5759/a000526

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*, 84-96. https://doi.org/10.1016/j.jrp.2005.08.007

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*, 101384. https://doi.org/10.1016/j.leaqua.2020.101384

Grau, I., Ebbeler, C., & Banse, R. (2019). Cultural differences in careless responding. *Journal of Cross-Cultural Psychology*, *50*, 336-357. https://doi.org/10.1177/0022022119827379

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, *33*, 313-317. https://doi.org/10.1027/1015-5759/a000450

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

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*, 457-459. https://doi.org/10.1177/0146621607299271

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101-125. https://doi.org/10.1177/014662169602000201

Harwell, M. (2019). A strategy for using bias and RMSEA as outcomes in Monte Carlo studies in statistics. *Journal of Modern Applied Statistical Methods*.

Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods, 12*, 554-566. https://doi.org/10.1177/1094428107300396

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*, 104-121. https://doi.org/10.1177/109442819800100106

Holden, R. R., Marjanovic, Z., & Troister, T. (2019). Indiscriminate responding can increase effect sizes for clinical phenomena in nonclinical populations: A cautionary note. *Journal of Psychoeducational Assessment*, *37*, 464-472. https://doi.org/10.1177/0734282918758809

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*, *80*, 312-345. https://doi.org/10.1177/0013164419865316

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 1-55. https://doi.org/10.1080/10705519909540118

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. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2014). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*, 299-311. https://doi.org/10.1007/s10869-014-9357-6

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. https://doi.org/10.1037/a0038510

Huang, J. L., & DeSimone, J. A. (2020). Insufficient effort responding as a potential confound between survey measures and objective tests. *Journal of Business and Psychology*, 1-22. https://doi.org/10.1007/s10869-020-09707-2

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*, 103-129. https://doi.org/10.1016/j.jrp.2004.09.009

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430-454. https://doi.org/10.1016/0010-0285(72)90016-3

Kam, C. C. S. (2019). Careless responding threatens factorial analytic results and construct validity of personality measure. *Frontiers in Psychology*, *10*, 1258. https://doi.org/10.3389/fpsyg.2019.01258

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*, 512-541. https://doi.org/10.1177/1094428115571894

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277-298. https://doi.org/10.1207/s15324818ame1604_2

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., ... & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, *42*, 139-15. https://doi.org/10.1080/00461520701416231

Kline, R. B. (2016). Principles and practice of structural equation modeling (4th ed.). New York, NY: Guildford Press.

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology*, *93*, 140-154. https://doi.org/10.1037/0021-9010.93.1.140

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236. https://doi.org/10.1002/acp.2350050305

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

Kung, F. Y., Kwok, N., & Brown, D. J. (2018). Are attention check questions a threat to scale validity? *Applied Psychology*, *67*, 264-283. https://doi.org/10.1111/apps.12108

LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An examination of item response theory item fit indices for the graded response model. *Organizational Research Methods*, *14*, 10-23. https://doi.org/10.1177/1094428109350930

Lake, C. J., Withrow, S., Zickar, M. J., Wood, N. L., Dalal, D. K., & Bochinski, J. (2013). Understanding the relation between attitude involvement and response latitude using item response theory. *Educational and Psychological Measurement*, *73*, 690-712. https://doi.org/10.1177/0013164413482920

Liu, T., Lan, T., & Xin, T. (2019). Detecting random responses in a personality scale using IRT-based person-fit indices. *European Journal of Psychological Assessment*, *35*, 126-136. https://doi.org/10.1027/1015-5759/a000369

Liu, T., Sun, Y., Li, Z., & Xin, T. (2019). The impact of aberrant response on reliability and validity. *Measurement: Interdisciplinary Research and Perspectives*, *17*, 133-142. https://doi.org/10.1080/15366367.2019.1584848

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

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences*, *84*, 79-83. https://doi.org/10.1016/j.paid.2014.08.021

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Marr, B. (2019, October, 28). The 10+ most important job skills every company will be looking for in 2020. Retrieved from: https://www.forbes.com/sites/bernardmarr/2019/10/28/the-10-most-important-job-skills-every-company-will-be-looking-for-in-2020/#530d21d67b67

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713-732. https://doi.org/10.1007/s11336-005-1295-9

McGonagle, A. K., Huang, J. L., & Walsh, B. M. (2016). Insufficient effort survey responding: An under-appreciated problem in work and organisational health psychology research. *Applied Psychology*, *65*, 287-321. https://doi.org/10.1111/apps.12058

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. https://doi.org/10.1016/j.chb.2018.03.007

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement*, *31*, 430-455. https://doi.org/10.1177/0146621606297316

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437-455. https://doi.org/10.1037/a0028085

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135. https://doi.org/10.1177/01466210122031957

Mosimann, J. E., Wiseman, C. V., & Edelman, R. E. (1995). Data fabrication: Can people generate random digits? *Accountability in Research*, *4*, 31-55. https://doi.org/10.1080/08989629508573866

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, *109*, 330-357. https://doi.org/10.1037//0033-295X.109.2.330

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology, 45,* 239 –250. https://doi.org/10.1002/1097-4679(198903)45:2<239::aid-jclp2270450210>3.0.co;2-1

Nichols, A. L., & Edlund, J. E. (2020). Why don't we care more about carelessness? Understanding the causes and consequences of careless participants. *International Journal of Social Research Methodology*, 1-14. https://doi.org/10.1080/13645579.2020.1719618

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1-11. https://doi.org/10.1016/j.jrp.2016.04.010

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods*, *14*, 548-570. https://doi.org/10.1177/1094428110368562

Nye, C. D., Joo, S. H., Zhang, B., & Stark, S. (2019). Advancing and evaluating IRT model data fit indices in organizational research. *Organizational Research Methods*, *23*, 457-486. https://doi.org/10.1177/1094428119833158

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. https://doi.org/10.1016/j.jesp.2009.03.009

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology*, *1*, 220. https://doi.org/10.3389/fpsyg.2010.00220

Patton, J. M., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, *44*, 309-341. https://doi.org/10.3102/1076998618825116

Rabin, M., & Vayanos, D. (2010). The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, *77*, 730-778. https://doi.org/10.1111/j.1467-937x.2009.00582.x

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology, 6*, 1715. https://doi.org/10.3389/fpsyg.2015.01715

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. *Journal of Statistical Software*, *48*, 1-36.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometric Monograph No. 17. Richmond, VA: Psychometric Society.

Schneider, S., May, M., & Stone, A. A. (2018). Careless responding in internet-based quality of life assessments. *Quality of Life Research*, *27*, 1077-1088. https://doi.org/10.1007/s11136-017-1767-2

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology*, *96*, 1055-1064. https://doi.org/10.1037/a0023322

Shulruf, B., Hattie, J., & Dixon, R. (2008). Factors affecting responses to Likert type questionnaires: Introduction of the ImpExp, a new comprehensive model. *Social Psychology of Education*, *11*, 59-78. https://doi.org/10.1007/s11218-007-9035-x

Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management*, *23*, 659-677. https://doi.org/10.1177/014920639702300503

Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social–emotional learning competencies. *Educational Measurement: Issues and Practice*, *38*, 101-111. https://doi.org/10.1111/emip.12256

Suzuki, T., Samuel, D. B., Pahlen, S., & Krueger, R. F. (2015). DSM-5 alternative personality disorder model traits as maladaptive extreme variants of the five-factor model: An item-response theory analysis. *Journal of Abnormal Psychology*, *124*, 343-354. https://doi.org/10.1037/abn0000035

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. https://doi.org/10.18637/jss.v074.i05

Tonidandel, S., King, E., & Cortina, J. (Eds.). (2015). *Big data at work: The data science revolution and organizational psychology*. New York, NY: Routledge. https://doi.org/10.4324/9781315780504

Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education, 26,* 169-181. https://doi.org/10.1108/qae-06-2017-0034

Tourangeau, R., Rips, L.J. and Rasinski, K. (2000). *The psychology of survey response*. Cambridge, England: Cambridge University Press.

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*, 195-217. https://doi.org/10.1093/ijpor/eds021

Voss, N. M., & Vangsness, L. (2020). Is procrastination related to low-quality data? *Educational Measurement: Issues and Practice, 39,* 95-104. https://doi.org/10.1111/emip.12355

Wang, W. C., Chen, H. F., & Jin, K. Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, *75*, 157-178. https://doi.org/10.1177/0013164414528209

Ward, M. K., & Pond III, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554-568. https://doi.org/10.1016/j.chb.2015.01.070

Ward, M. K., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior*, *76*, 417-430. https://doi.org/10.1016/j.chb.2017.06.032

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, *67*, 231-263. https://doi.org/10.1111/apps.12118

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*, 409-422. https://doi.org/10.1007/s11747-007-0077-6

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*, 454-464. https://doi.org/10.1177/1948550617703168

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*, 189-194. https://doi.org/10.1007/s10862-005-9004-7

Yentes, R.D., & Wilhelm, F. (2018) careless: Procedures for computing indices of careless responding. R packages version 1.1.0 url: https://github.com/ryentes/careless

Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods*. https://doi.org/10.1037/met0000212

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, *7*, 168-190. https://doi.org/10.1177/1094428104263674

Appendix A - List of Study 1 Survey Items

List of NEO IPIP conscientiousness items (Goldberg et al., 2006) included in Study 1. All responses were provided one a 5-point, Likert scale ranging from (1) *Very Inaccurate* to (5) *Very Accurate.*

1. I waste my time.
2. I am not highly motivated to succeed.
3. I do the opposite of what is asked.
4. I turn plans into actions.
5. I find it difficult to get down to work.
6. I often make last-minute plans.
7. I like order.
8. I make rash decisions.
9. I get to work at once.
10. I go straight for the goal.
11. I am always prepared.
12. I demand quality.
13. I like to tidy up.
14. I set high standards for myself and others.
15. I don't understand things.
16. I leave my belongings around.
17. I choose my words with care.
18. I need a push to get started.
19. I do just enough work to get by.

20. I rush into things.

21. I complete tasks successfully.

22. I am not bothered by messy people.

23. I like to act on a whim.

24. I get others to do my duties.

25. I listen to my conscience.

26. I handle tasks smoothly.

27. I break rules.

28. I do crazy things.

29. I work hard.

30. I carry out my plans.

31. I postpone decisions.

32. I misrepresent the facts.

33. I jump into things without thinking.

34. I tell the truth.

35. I love order and regularity.

36. I plunge into tasks with all my heart.

37. I know how to get things done.

38. I want everything to be "just right."

39. I am sure of my ground.

40. I don't see the consequences of things.

41. I keep my promises.

42. I often forget to put things back in their proper place.

43. I try to follow the rules.

44. I stick to my chosen path.

45. I put little time and effort into my work.

46. I leave a mess in my room.

47. I avoid mistakes.

48. I excel in what I do.

49. I have difficulty starting tasks.

50. I pay my bills on time.

51. I act without thinking.

52. I am not bothered by disorder.

53. I do things according to a plan.

54. I misjudge situations.

55. I come up with good solutions.

56. I have little to contribute.

57. I get chores done right away.

58. I break my promises.

59. I do more than what's expected of me.

60. I start tasks right away.

Appendix B - Item-Level IRT Parameters from Study 1 Conditions

Table 23. Item-level IRT parameters from Study 1 careful and control conditions

Item	Careful Condition					Control Condition				
	a	b_1	b_2	b_3	b_4	a	b_1	b_2	b_3	b_4
1	1.76	-3.72	-2.77	-2.19	0.47	2.90	-3.31	-2.39	-1.34	0.29
2	2.02	-3.19	-1.92	-1.22	1.01	3.06	-2.64	-1.67	-0.90	0.55
3	2.47	-2.39	-1.72	-1.13	0.60	2.51	-2.69	-1.91	-1.01	0.41
4	1.70	-3.16	-1.75	-0.75	1.20	2.39	-3.26	-1.65	-0.79	0.69
5	1.89	-3.59	-2.10	-0.90	1.05	3.09	-2.86	-1.65	-0.64	0.46
6	1.53	-3.64	-2.91	-1.74	1.03	2.04	-3.59	-2.20	-1.17	0.53
7	1.05	-4.31	-3.03	-1.87	0.25	1.00	-4.48	-3.19	-1.61	0.68
8	1.23	-3.04	-1.69	-0.91	0.80	1.36	-3.29	-1.90	-1.01	0.66
9	1.30	-3.37	-2.41	-1.29	0.35	1.13	-4.27	-2.78	-1.42	0.63
10	1.11	-4.36	-2.66	-1.42	0.91	1.04	-4.13	-2.63	-1.31	0.63
11	2.36	-3.02	-1.87	-0.96	0.77	2.00	-2.98	-2.08	-1.12	0.81
12	0.82	-6.45	-3.95	-2.25	0.96	1.10	-5.00	-3.12	-1.78	0.91
13	1.05	-5.89	-3.81	-2.51	0.23	1.28	-3.38	-2.08	0.13	NA
14	1.26	-4.21	-3.46	-2.45	0.22	1.51	-3.30	-2.16	-0.06	NA
15	1.13	-4.57	-3.00	-2.06	0.58	1.10	-5.65	-3.76	-1.97	0.52
16	0.65	-6.19	-4.60	-3.09	-0.52	0.99	-4.13	-3.63	-2.39	-0.38
17	2.36	-3.05	-1.71	-1.02	0.96	2.83	-2.58	-1.54	-0.77	0.67
18	2.43	-2.48	-1.26	-0.59	0.98	2.86	-2.55	-1.39	-0.58	0.48
19	1.42	-3.83	-2.16	-1.01	0.91	1.55	-4.37	-2.17	-1.06	0.61
20	1.47	-3.14	-2.17	-1.19	0.74	1.99	-2.95	-1.89	-0.97	0.51
21	2.67	-2.59	-1.64	-1.00	0.25	2.77	-2.75	-1.91	-1.01	0.05
22	1.88	-2.31	-1.33	-0.42	0.98	1.73	-2.76	-1.56	-0.57	0.89
23	1.83	-2.43	-1.66	-0.70	1.03	1.92	-2.72	-1.65	-0.68	1.09
24	1.93	-2.04	-1.07	-0.37	1.20	2.10	-2.42	-1.14	-0.51	0.66
25	1.83	-3.01	-1.58	-0.93	1.33	1.90	-3.06	-1.48	-0.87	1.13
26	3.37	-2.58	-1.86	-1.18	0.53	3.42	-2.30	-1.69	-0.91	0.26
27	1.97	-1.87	-0.99	-0.19	1.11	1.81	-2.51	-1.09	-0.51	0.79
28	2.60	-1.90	-1.17	-0.43	0.92	2.55	-2.24	-1.17	-0.52	0.57
29	0.89	-4.87	-3.48	-1.71	1.28	0.87	-6.06	-2.96	-1.54	1.22
30	1.77	-2.95	-2.08	-0.77	1.33	1.86	-3.32	-2.02	-0.76	0.89
31	0.79	-4.70	-3.34	-1.65	1.69	1.16	-3.69	-2.52	-1.25	1.22
Mean	**1.69**	**-3.51**	**-2.30**	**-1.29**	**0.81**	**1.93**	**-3.39**	**-2.10**	**-1.00**	**0.64**
SD	**0.64**	**1.21**	**0.93**	**0.71**	**0.44**	**0.74**	**0.96**	**0.70**	**0.53**	**0.34**

Table 24. Item-level IRT parameters from Study 1 careless and random conditions

Item	Careless Condition					Random Condition				
	a	b_1	b_2	b_3	b_4	*a*	b_1	b_2	b_3	b_4
1	2.40	-0.67	-0.09	0.28	1.35	1.46	-0.93	-0.18	0.26	1.52
2	2.03	-1.14	-0.37	0.34	1.36	0.87	-2.62	-1.00	0.62	2.65
3	2.15	-1.20	-0.44	0.35	1.27	0.95	-2.41	-1.07	0.50	2.17
4	1.71	-1.25	-0.50	0.26	1.33	1.04	-2.02	-0.75	0.29	1.81
5	1.92	-1.16	-0.45	0.32	1.22	1.25	-2.37	-0.81	0.39	2.12
6	1.92	-1.11	-0.39	0.20	1.28	0.98	-1.94	-0.82	0.60	2.21
7	2.24	-1.01	-0.30	0.33	1.19	0.92	-2.65	-0.89	0.21	2.03
8	2.06	-1.08	-0.31	0.40	1.55	1.10	-2.48	-0.97	0.41	2.06
9	2.09	-1.08	-0.38	0.36	1.16	1.02	-2.69	-0.89	-0.03	1.44
10	2.08	-1.09	-0.44	0.28	1.11	1.06	-2.63	-0.95	0.37	2.02
11	2.13	-0.95	-0.27	0.42	1.39	1.12	-2.41	-0.86	0.59	2.09
12	2.14	-1.02	-0.34	0.32	1.43	0.92	-2.66	-0.91	0.47	2.78
13	2.56	-0.91	-0.25	0.32	1.11	1.23	-2.36	-0.94	0.02	2.13
14	2.55	-0.40	-0.06	0.33	1.02	1.63	-0.90	-0.25	0.35	1.50
15	2.15	-0.88	-0.40	0.37	1.16	1.45	-1.65	-0.67	0.41	1.82
16	2.05	-1.03	-0.40	0.23	0.84	1.18	-1.96	-0.79	0.05	1.34
17	1.86	-1.08	-0.30	0.43	1.66	0.88	-3.01	-1.05	0.57	2.73
18	2.10	-1.07	-0.48	0.45	1.30	1.16	-1.71	-0.62	0.29	1.83
19	2.10	-0.98	-0.23	0.42	1.43	0.97	-2.50	-0.72	0.52	2.16
20	2.06	-1.06	-0.28	0.43	1.34	1.27	-2.14	-0.73	0.14	1.71
21	2.43	-0.91	-0.24	0.40	1.09	1.50	-1.55	-0.56	0.30	1.32
22	1.79	-1.11	-0.25	0.47	1.49	0.96	-2.53	-0.90	0.47	2.34
23	2.15	-0.99	-0.34	0.41	1.29	0.69	-3.36	-1.65	0.48	2.93
24	1.68	-1.23	-0.38	0.39	1.53	0.85	-2.56	-0.75	0.64	2.56
25	2.21	-0.65	0.09	0.50	1.51	1.34	-1.11	-0.10	0.57	2.23
26	2.46	-0.95	-0.34	0.35	1.12	1.31	-1.60	-0.51	0.39	1.58
27	2.00	-1.24	-0.36	0.43	1.31	0.91	-1.99	-0.56	0.70	2.25
28	1.91	-0.96	-0.38	0.33	1.26	0.78	-2.81	-1.43	0.35	2.52
29	1.81	-1.05	-0.32	0.38	1.62	1.06	-2.40	-0.84	0.55	2.21
30	1.93	-1.24	-0.50	0.27	1.33	0.99	-2.63	-0.99	0.35	2.02
31	2.00	-1.02	-0.25	0.41	1.23	0.94	-2.29	-1.09	0.56	2.39
Mean	**2.09**	**-1.02**	**-0.32**	**0.36**	**1.30**	**1.09**	**-2.22**	**-0.81**	**0.40**	**2.08**
SD	**0.22**	**0.18**	**0.13**	**0.07**	**0.18**	**0.23**	**0.58**	**0.31**	**0.19**	**0.43**

Table 25. Item-level *b*-to-*b* distance values from Study 1 careful and control conditions

Item	Careful Condition			Control Condition		
	b_1-b_2	b_2-b_3	b_3-b_4	b_1-b_2	b_2-b_3	b_3-b_4
1	0.95	0.58	2.67	0.93	1.05	1.62
2	1.28	0.70	2.23	0.97	0.77	1.45
3	0.66	0.60	1.72	0.78	0.90	1.41
4	1.41	1.00	1.94	1.61	0.86	1.48
5	1.49	1.20	1.95	1.21	1.01	1.10
6	0.73	1.16	2.78	1.39	1.03	1.70
7	1.28	1.16	2.12	1.29	1.59	2.28
8	1.35	0.78	1.72	1.40	0.89	1.67
9	0.96	1.11	1.64	1.49	1.36	2.04
10	1.70	1.24	2.33	1.50	1.32	1.94
11	1.14	0.91	1.73	0.90	0.96	1.93
12	2.50	1.70	3.21	1.88	1.34	2.69
13	2.08	1.30	2.74	1.30	2.21	NA
14	0.75	1.01	2.67	1.14	2.11	NA
15	1.57	0.94	2.64	1.89	1.79	2.49
16	1.59	1.51	2.57	0.50	1.24	2.01
17	1.34	0.69	1.98	1.04	0.77	1.43
18	1.22	0.67	1.57	1.16	0.81	1.06
19	1.67	1.15	1.92	2.20	1.11	1.67
20	0.97	0.98	1.93	1.06	0.92	1.49
21	0.96	0.63	1.26	0.83	0.90	1.06
22	0.98	0.91	1.40	1.20	0.99	1.46
23	0.77	0.96	1.73	1.07	0.97	1.77
24	0.98	0.70	1.57	1.28	0.63	1.17
25	1.43	0.66	2.25	1.58	0.60	2.00
26	0.72	0.68	1.71	0.62	0.78	1.17
27	0.88	0.80	1.30	1.43	0.58	1.30
28	0.73	0.74	1.35	1.06	0.66	1.09
29	1.39	1.78	2.99	3.11	1.42	2.76
30	0.86	1.32	2.09	1.30	1.25	1.65
31	1.37	1.69	3.34	1.18	1.27	2.47
Mean	**1.22**	**1.01**	**2.10**	**1.30**	**1.10**	**1.70**
SD	**0.42**	**0.34**	**0.57**	**0.49**	**0.41**	**0.49**

Table 26. Item-level *b*-to-*b* distance values from Study 1 careless and random conditions

Item	Careless Condition			Random Condition		
	b_1-b_2	b_2-b_3	b_3-b_4	b_1-b_2	b_2-b_3	b_3-b_4
1	0.58	0.38	1.07	0.75	0.44	1.27
2	0.77	0.71	1.02	1.62	1.63	2.02
3	0.77	0.78	0.92	1.34	1.57	1.67
4	0.75	0.76	1.07	1.28	1.04	1.53
5	0.71	0.77	0.90	1.56	1.20	1.73
6	0.72	0.59	1.08	1.12	1.42	1.61
7	0.71	0.63	0.86	1.77	1.10	1.82
8	0.77	0.71	1.15	1.51	1.38	1.66
9	0.71	0.74	0.79	1.79	0.86	1.47
10	0.65	0.72	0.83	1.67	1.33	1.65
11	0.68	0.69	0.97	1.55	1.46	1.50
12	0.68	0.65	1.12	1.75	1.38	2.32
13	0.66	0.57	0.79	1.42	0.96	2.12
14	0.34	0.39	0.69	0.65	0.61	1.15
15	0.49	0.76	0.79	0.98	1.08	1.41
16	0.63	0.63	0.61	1.16	0.84	1.30
17	0.78	0.73	1.23	1.96	1.62	2.16
18	0.59	0.93	0.86	1.09	0.91	1.54
19	0.76	0.65	1.01	1.78	1.24	1.64
20	0.79	0.71	0.91	1.42	0.86	1.58
21	0.67	0.64	0.69	0.99	0.85	1.02
22	0.86	0.72	1.02	1.63	1.37	1.87
23	0.65	0.75	0.88	1.71	2.13	2.46
24	0.86	0.76	1.14	1.81	1.38	1.92
25	0.74	0.42	1.01	1.01	0.68	1.65
26	0.61	0.69	0.77	1.09	0.89	1.19
27	0.89	0.79	0.88	1.43	1.26	1.56
28	0.58	0.71	0.94	1.38	1.78	2.17
29	0.73	0.69	1.24	1.56	1.39	1.66
30	0.74	0.77	1.06	1.64	1.34	1.67
31	0.77	0.66	0.83	1.20	1.65	1.83
Mean	**0.70**	**0.68**	**0.94**	**1.41**	**1.21**	**1.68**
SD	**0.11**	**0.12**	**0.16**	**0.33**	**0.37**	**0.34**

Appendix C - Model Fit Indices for Baseline-Careful Conditions

Table 27. Model fit values for all careful/baseline replications (200 participants/10 items)

	IRT Fit			CFA Fit		
Replication	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
1	4.87	.00	.04	30.40	.00	.03
2	12.87	.08	.05	53.23	.05	.04
3	4.30	.00	.04	33.47	.00	.03
4	4.72	.00	.05	53.24	.05	.04
5	4.88	.00	.04	38.14	.02	.04
6	7.83	.02	.05	38.38	.02	.04
7	10.35	.04	.05	38.00	.02	.04
8	7.92	.00	.06	43.79	.04	.05
9	12.21	.04	.05	39.41	.03	.04
10	13.55	.05	.05	30.56	.00	.04
11	3.54	.00	.05	53.48	.05	.05
12	3.63	.00	.05	39.36	.03	.04
13	6.57	.00	.04	28.92	.00	.04
14	6.01	.00	.04	33.99	.00	.04
15	7.93	.00	.05	35.86	.01	.04
16	5.41	.00	.04	33.71	.00	.04
17	13.39	.07	.05	60.25	.00	.05
18	8.02	.00	.04	37.56	.02	.04
19	6.51	.02	.05	48.29	.04	.04
20	12.75	.05	.05	46.14	.04	.04
21	4.32	.00	.04	36.05	.01	.04
22	9.08	.03	.05	52.49	.05	.05
23	6.93	.00	.05	46.01	.04	.04
24	27.18	.10	.05	40.97	.03	.04
25	5.66	.00	.04	41.77	.03	.04
26	4.68	.00	.04	28.72	.00	.03
27	23.13	.06	.05	51.67	.05	.05
28	19.63	.05	.06	67.58	.07	.05
29	18.16	.05	.06	48.37	.04	.05
30	21.23	.06	.06	55.15	.05	.05
31	9.89	.00	.06	61.72	.06	.05
32	12.71	.04	.05	55.33	.05	.04
33	13.06	.04	.05	63.00	.06	.05
34	20.80	.07	.06	57.15	.06	.05
35	19.75	.06	.05	52.79	.05	.04
36	16.60	.04	.05	62.63	.06	.05
37	6.94	.00	.03	22.49	.00	.03
38	4.19	.00	.05	60.47	.06	.05
39	9.64	.03	.04	38.59	.02	.04
40	5.37	.00	.04	46.28	.04	.04

41	4.13	.00	.04	37.97	.02	.03
42	15.90	.04	.05	45.96	.04	.04
43	16.44	.04	.05	52.38	.05	.05
44	9.99	.00	.04	30.16	.00	.04
45	18.92	.05	.05	53.17	.05	.04
46	9.47	.00	.04	30.71	.00	.04
47	10.13	.04	.04	45.65	.04	.04
48	5.22	.00	.04	23.97	.00	.03
49	15.69	.07	.05	44.08	.04	.04
50	9.31	.00	.05	40.05	.03	.04
51	6.90	.00	.05	40.91	.03	.04
52	6.70	.00	.05	52.52	.05	.04
53	9.72	.02	.06	52.89	.05	.05
54	6.58	.00	.05	40.74	.03	.04
55	4.49	.00	.05	38.63	.02	.04
56	7.33	.00	.06	29.81	.00	.04
57	8.76	.00	.04	31.04	.00	.03
58	22.02	.05	.05	50.23	.05	.04
59	20.22	.06	.05	51.45	.05	.04
60	4.63	.00	.05	32.63	.00	.03
61	10.22	.00	.04	33.90	.00	.03
62	4.96	.00	.05	50.12	.05	.04
63	4.87	.00	.04	32.14	.00	.03
64	4.39	.00	.04	57.87	.06	.04
65	14.69	.06	.04	41.86	.03	.04
66	5.72	.00	.04	39.40	.03	.04
67	7.81	.00	.04	40.37	.03	.04
68	8.34	.03	.04	32.32	.00	.04
69	5.45	.00	.04	32.78	.00	.04
70	7.09	.03	.05	27.03	.00	.04
71	4.05	.00	.04	34.27	.00	.04
72	4.19	.00	.05	31.07	.00	.04
73	4.72	.00	.05	48.92	.05	.04
74	3.62	.00	.05	49.75	.05	.04
75	7.62	.04	.05	56.87	.06	.05
76	4.69	.00	.05	35.39	.01	.04
77	18.57	.05	.05	42.57	.03	.05
78	7.34	.00	.05	38.36	.02	.04
79	16.65	.06	.06	58.27	.06	.05
80	11.95	.02	.05	58.53	.06	.05
81	8.97	.00	.05	35.41	.01	.04
82	10.61	.00	.05	47.57	.04	.04
83	7.90	.00	.05	50.22	.05	.04
84	9.00	.03	.04	39.16	.02	.03
85	12.59	.04	.05	45.26	.04	.04
86	4.52	.00	.04	24.43	.00	.03

87	1.36	.00	.04	17.25	.00	.03
88	12.34	.04	.05	39.17	.02	.04
89	7.95	.03	.05	62.18	.06	.05
90	10.17	.05	.05	42.91	.03	.04
91	12.82	.05	.05	41.01	.03	.04
92	6.67	.00	.04	34.59	.00	.04
93	5.68	.00	.04	36.36	.01	.04
94	4.08	.00	.03	16.81	.00	.03
95	15.77	.03	.05	51.42	.05	.05
96	10.34	.01	.05	38.30	.02	.04
97	4.12	.00	.05	30.34	.00	.04
98	5.50	.00	.06	68.81	.07	.05
99	1.92	.00	.04	28.44	.00	.04
100	13.01	.07	.06	53.68	.05	.05
Mean	**9.51**	**.02**	**.05**	**42.75**	**.03**	**.04**
SD	**5.41**	**.03**	**.01**	**11.21**	**.02**	**.01**

Note.

Table 28. Model fit values for all careful/baseline replications (200 participants/ 60 items)

	IRT Fit			CFA Fit		
Replication	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
1	1547.74	.00	.05	2106.53	.04	.05
2	1638.54	.02	.05	2064.72	.04	.05
3	1526.47	.00	.05	2123.87	.04	.05
4	1627.63	.02	.06	2187.13	.04	.05
5	1687.59	.02	.06	2353.12	.05	.05
6	1697.81	.02	.06	2304.44	.04	.05
7	1506.33	.00	.05	2265.41	.04	.05
8	1547.08	.00	.06	2203.11	.04	.05
9	1596.24	.01	.05	2189.10	.04	.05
10	1577.45	.01	.05	2097.98	.04	.05
11	1581.09	.01	.05	2151.60	.04	.05
12	1545.67	.00	.05	2131.03	.04	.05
13	1562.92	.00	.05	2079.01	.04	.05
14	1622.60	.02	.05	2257.77	.04	.05
15	1608.00	.01	.05	2112.43	.04	.05
16	1581.72	.01	.06	2109.51	.04	.05
17	1595.21	.01	.06	2157.58	.04	.05
18	1578.00	.01	.06	2101.28	.04	.05
19	1724.56	.02	.06	2188.92	.04	.05
20	1588.88	.01	.05	2137.82	.04	.05
21	1583.91	.01	.06	2081.35	.04	.05
22	1546.31	.00	.06	2093.47	.04	.05
23	1572.66	.01	.05	2139.07	.04	.05
24	1635.39	.02	.05	2107.95	.04	.05
25	1643.61	.02	.06	2235.50	.04	.05
26	1673.77	.02	.05	2193.63	.04	.05
27	1636.60	.02	.05	2202.08	.04	.05
28	1597.14	.01	.05	2147.82	.04	.05
29	1660.54	.02	.05	2189.16	.04	.05
30	1643.37	.02	.05	2150.18	.04	.05
31	1597.73	.01	.06	2219.10	.04	.05
32	1523.37	.00	.05	2093.14	.04	.05
33	1648.85	.02	.06	2321.93	.05	.05
34	1659.23	.02	.06	2389.74	.05	.05
35	1589.88	.01	.05	2086.92	.04	.05
36	1512.49	.00	.06	2181.19	.04	.05
37	1635.76	.02	.05	2053.92	.04	.05
38	1468.09	.00	.05	1916.06	.03	.05
39	1611.80	.01	.05	2094.02	.04	.05
40	1508.24	.00	.05	2028.82	.03	.05
41	1540.47	.00	.05	2002.80	.03	.05
42	1536.31	.00	.05	2238.74	.04	.05
43	1637.77	.02	.05	2300.98	.04	.05

44	1555.80	.00	.05	2246.55	.04	.05
45	1569.17	.01	.05	2187.30	.04	.05
46	1647.27	.02	.05	2093.61	.04	.05
47	1597.76	.01	.05	2268.89	.04	.05
48	1615.34	.01	.05	1996.99	.03	.05
49	1597.58	.01	.05	2046.30	.04	.05
50	1716.67	.02	.05	2176.36	.04	.05
51	1598.44	.01	.05	1998.03	.03	.05
52	1549.37	.00	.05	1948.55	.03	.05
53	1620.39	.01	.05	2242.31	.04	.05
54	1522.08	.00	.05	2156.53	.04	.05
55	1665.04	.02	.05	2228.76	.04	.05
56	1570.16	.00	.05	2225.86	.04	.05
57	1506.66	.00	.05	2036.28	.03	.05
58	1599.41	.01	.05	2275.07	.04	.05
59	1504.62	.00	.05	2058.40	.04	.05
60	1570.81	.01	.05	2137.81	.04	.05
61	1612.46	.01	.05	2350.54	.05	.05
62	1563.68	.00	.05	2136.16	.04	.05
63	1569.13	.01	.05	2104.40	.04	.05
64	1625.45	.01	.05	2257.84	.04	.05
65	1573.35	.01	.05	2187.57	.04	.05
66	1594.11	.01	.05	2222.07	.04	.05
67	1618.08	.01	.06	2229.12	.04	.05
68	1601.40	.01	.05	2145.70	.04	.05
69	1601.96	.01	.05	2248.52	.04	.05
70	1532.18	.00	.05	1990.09	.03	.05
71	1640.38	.02	.05	2138.12	.04	.05
72	1599.00	.01	.05	2058.41	.04	.05
73	1584.86	.01	.05	2020.80	.03	.05
74	1644.38	.02	.05	2112.16	.04	.05
75	1491.95	.00	.05	2169.96	.04	.05
76	1546.82	.00	.05	2319.36	.05	.05
77	1611.66	.01	.05	2352.66	.05	.05
78	1641.63	.02	.05	2345.95	.05	.05
79	1559.29	.00	.05	2083.98	.04	.05
80	1743.19	.02	.05	2496.87	.05	.05
81	1527.69	.00	.05	2162.65	.04	.05
82	1586.08	.01	.05	2274.56	.04	.05
83	1616.02	.01	.05	2143.87	.04	.05
84	1616.16	.01	.05	2318.27	.05	.05
85	1630.46	.01	.05	2173.20	.04	.05
86	1687.56	.02	.06	2225.07	.04	.05
87	1553.73	.00	.05	2015.93	.03	.05
88	1750.88	.02	.05	2220.34	.04	.05
89	1680.86	.02	.05	2242.61	.04	.05

90	1558.33	.00	.05	2042.96	.03	.05
91	1560.62	.00	.05	2041.14	.03	.05
92	1614.61	.01	.05	2270.15	.04	.05
93	1530.79	.00	.05	2199.68	.04	.05
94	1556.79	.00	.05	2305.75	.04	.05
95	1640.35	.02	.05	2307.02	.05	.05
96	1573.75	.01	.05	2218.00	.04	.05
97	1588.22	.01	.05	2263.73	.04	.05
98	1685.37	.02	.05	2212.15	.04	.05
99	1697.69	.02	.06	2226.61	.04	.05
100	1609.71	.01	.05	2105.65	.04	.05
Mean	**1598.32**	**.01**	**.05**	**2170.53**	**.04**	**.05**
SD	**56.39**	**.01**	**.00**	**104.57**	**.00**	**.00**

Table 29. Model fit values for all careful/baseline replications (500 participants/10 items).

	IRT Fit			CFA Fit		
Replication	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
1	5.32	.00	.03	36.47	.01	.02
2	3.78	.00	.03	40.25	.02	.03
3	6.24	.01	.03	54.89	.03	.03
4	7.18	.00	.03	63.27	.04	.03
5	6.82	.00	.03	54.90	.03	.03
6	9.97	.00	.03	54.45	.03	.03
7	4.45	.00	.03	45.74	.03	.03
8	20.63	.05	.03	75.36	.05	.03
9	8.29	.00	.03	66.84	.04	.03
10	0.46	.00	.03	71.02	.05	.03
11	9.14	.00	.04	36.82	.01	.02
12	6.39	.00	.03	50.01	.03	.03
13	11.52	.04	.02	36.47	.01	.02
14	1.58	.00	.03	81.43	.05	.04
15	1.22	.00	.03	40.73	.02	.03
16	6.17	.00	.03	29.74	.00	.02
17	12.81	.04	.03	36.92	.01	.03
18	7.35	.00	.03	40.55	.02	.03
19	5.10	.00	.03	60.16	.04	.03
20	4.53	.00	.03	30.07	.00	.02
21	9.98	.01	.03	54.38	.03	.03
22	6.33	.00	.03	46.90	.03	.03
23	3.28	.00	.03	28.97	.00	.02
24	7.06	.00	.03	51.74	.03	.02
25	5.10	.00	.03	38.20	.01	.03
26	7.08	.02	.03	50.69	.03	.03
27	13.79	.03	.03	30.79	.00	.02
28	3.74	.00	.03	30.38	.00	.02
29	7.83	.02	.03	37.44	.01	.03
30	2.18	.00	.03	46.13	.03	.03
31	14.00	.01	.03	46.37	.03	.03
32	11.77	.00	.03	39.03	.02	.02
33	10.16	.00	.03	45.93	.03	.03
34	18.84	.03	.03	44.41	.02	.03
35	6.45	.00	.03	38.93	.02	.02
36	6.18	.00	.03	47.44	.03	.03
37	5.77	.02	.02	33.42	.00	.02
38	1.77	.00	.03	47.53	.03	.03
39	2.91	.00	.03	42.53	.02	.02
40	5.40	.00	.02	30.68	.00	.02
41	4.56	.00	.03	33.93	.00	.02
42	1.79	.00	.03	27.18	.00	.02
43	5.46	.00	.03	37.74	.01	.03

44	11.23	.04	.03	53.89	.03	.03
45	10.78	.04	.03	46.59	.03	.03
46	3.73	.00	.02	32.53	.00	.02
47	7.43	.02	.04	43.00	.02	.03
48	3.11	.00	.03	35.13	.00	.02
49	5.20	.00	.03	37.79	.01	.03
50	2.57	.00	.03	24.06	.00	.02
51	8.29	.00	.03	43.98	.02	.03
52	5.48	.00	.03	40.85	.02	.03
53	11.74	.00	.03	70.80	.05	.03
54	21.64	.05	.04	78.56	.05	.04
55	9.48	.00	.03	49.69	.03	.03
56	4.27	.00	.02	33.09	.00	.02
57	10.86	.01	.04	61.78	.04	.03
58	16.68	.03	.03	82.10	.05	.04
59	8.37	.01	.04	74.19	.05	.04
60	11.33	.03	.03	63.63	.04	.03
61	12.74	.04	.03	40.30	.02	.03
62	18.79	.05	.03	64.07	.04	.03
63	1.69	.00	.03	37.51	.01	.03
64	4.28	.00	.03	37.02	.01	.03
65	3.42	.00	.03	42.67	.02	.03
66	4.56	.00	.03	48.15	.03	.03
67	7.61	.00	.04	99.35	.06	.04
68	10.42	.02	.03	35.76	.01	.03
69	4.69	.00	.02	22.64	.00	.02
70	8.95	.00	.03	39.55	.02	.03
71	3.12	.00	.03	39.02	.02	.03
72	3.01	.00	.03	56.74	.04	.03
73	14.08	.03	.03	39.53	.02	.03
74	5.71	.00	.03	39.43	.02	.03
75	6.04	.00	.02	22.76	.00	.02
76	6.82	.03	.03	53.48	.03	.03
77	4.24	.00	.03	44.33	.02	.03
78	3.66	.00	.03	47.52	.03	.02
79	4.56	.00	.04	81.02	.05	.04
80	9.58	.03	.03	40.74	.02	.02
81	10.31	.05	.03	55.16	.03	.03
82	3.43	.00	.04	45.68	.03	.03
83	1.18	.00	.03	50.03	.03	.03
84	12.15	.05	.03	47.10	.03	.03
85	6.00	.00	.03	41.24	.02	.03
86	7.60	.01	.02	29.76	.00	.02
87	4.61	.00	.03	38.76	.02	.03
88	15.49	.04	.03	49.13	.03	.03
89	17.10	.05	.03	60.84	.04	.03

90	6.93	.00	.03	32.13	.00	.02
91	23.65	.07	.04	55.82	.03	.03
92	5.13	.00	.03	47.76	.03	.03
93	12.61	.04	.03	52.52	.03	.03
94	6.09	.01	.03	57.02	.04	.03
95	3.17	.00	.02	21.30	.00	.02
96	2.70	.00	.03	47.82	.03	.03
97	13.79	.05	.04	55.54	.03	.03
98	12.86	.03	.03	53.58	.03	.03
99	6.01	.00	.03	45.76	.03	.03
100	7.97	.03	.03	42.12	.02	.03
Mean	**7.69**	**.01**	**.03**	**46.69**	**.02**	**.03**
SD	**4.80**	**.02**	**.00**	**14.38**	**.02**	**.00**

Table 30. Model fit values for all careful/baseline replications (500 participants/60 items)

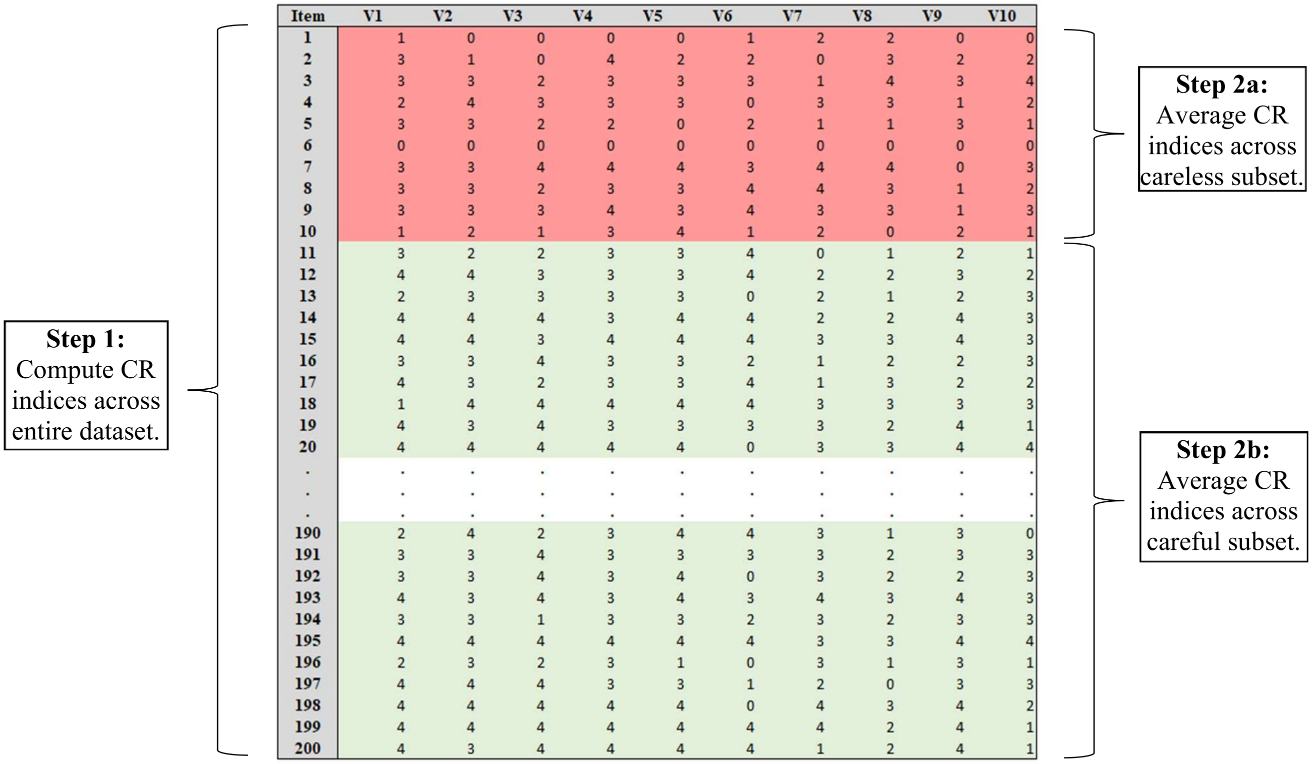
	IRT Fit			CFA Fit		
Replication	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
1	1645.05	.01	.03	2435.34	.03	.04
2	1574.81	.01	.03	2043.44	.02	.03
3	1522.65	.00	.03	2044.31	.02	.03
4	1637.03	.01	.04	2063.78	.02	.04
5	1602.17	.01	.04	1981.08	.02	.03
6	1693.64	.01	.04	2137.87	.02	.04
7	1600.15	.01	.04	2111.45	.02	.04
8	1580.38	.01	.04	1952.28	.02	.03
9	1554.51	.00	.04	2076.49	.02	.03
10	1562.78	.00	.03	2095.84	.02	.03
11	1568.92	.00	.03	2014.37	.02	.03
12	1708.28	.01	.04	2214.14	.03	.03
13	1428.10	.00	.03	2374.86	.03	.04
14	1515.10	.00	.04	2091.23	.02	.03
15	1526.36	.00	.03	2181.17	.03	.03
16	1593.85	.01	.03	2147.13	.02	.03
17	1562.56	.01	.04	2185.36	.03	.03
18	1528.88	.00	.03	2288.31	.03	.03
19	1592.09	.01	.03	2155.04	.03	.03
20	1554.62	.00	.03	2094.05	.02	.03
21	1553.30	.00	.03	2048.03	.02	.03
22	1680.64	.01	.03	2188.09	.03	.03
23	1480.28	.00	.03	1979.17	.02	.03
24	1562.95	.00	.03	2244.63	.03	.03
25	1594.29	.01	.03	2070.04	.02	.03
26	1702.76	.01	.03	2253.38	.03	.03
27	1660.50	.01	.03	2081.44	.02	.03
28	1534.71	.00	.03	2062.37	.02	.03
29	1624.13	.01	.03	2140.30	.02	.03
30	1653.39	.01	.03	2232.37	.03	.03
31	1507.80	.00	.03	1976.01	.02	.03
32	1571.25	.00	.03	2015.95	.02	.03
33	1593.55	.01	.03	2109.08	.02	.03
34	1575.25	.01	.03	2002.51	.02	.03
35	1530.87	.00	.03	2222.27	.03	.03
36	1636.18	.01	.03	2164.37	.03	.03
37	1566.73	.01	.03	2084.63	.02	.03
38	1533.89	.00	.03	1973.04	.02	.03
39	1577.25	.01	.03	1965.32	.02	.03
40	1679.13	.01	.03	2114.96	.02	.03
41	1580.45	.01	.03	2053.46	.02	.03
42	1614.86	.01	.04	2078.26	.02	.04
43	1550.67	.00	.03	1939.28	.02	.03

44	1657.02	.01	.04	1984.63	.02	.03
45	1534.05	.00	.03	1960.79	.02	.03
46	1604.15	.01	.03	2370.83	.03	.04
47	1543.93	.00	.04	2047.70	.02	.04
48	1629.31	.01	.03	2215.09	.03	.03
49	1614.08	.01	.03	2123.78	.02	.03
50	1598.39	.01	.03	2043.34	.02	.03
51	1611.88	.01	.03	2194.14	.03	.03
52	1576.77	.01	.03	2104.73	.02	.03
53	1558.76	.00	.03	1944.90	.02	.03
54	1586.26	.01	.03	1991.65	.02	.03
55	1560.18	.00	.03	1954.96	.02	.03
56	1522.43	.00	.03	1935.79	.02	.03
57	1615.46	.01	.03	2038.20	.02	.03
58	1538.78	.00	.03	1995.47	.02	.03
59	1573.91	.01	.03	2104.91	.02	.03
60	1619.24	.01	.03	2101.69	.02	.03
61	1539.35	.00	.03	2054.45	.02	.03
62	1582.27	.01	.03	2057.52	.02	.03
63	1603.33	.01	.03	2003.00	.02	.03
64	1503.36	.00	.03	2007.84	.02	.03
65	1636.95	.01	.03	2220.41	.03	.03
66	1590.01	.01	.03	2050.13	.02	.03
67	1605.73	.01	.03	2112.02	.02	.03
68	1541.26	.00	.03	2151.41	.03	.03
69	1584.38	.01	.03	2123.38	.02	.03
70	1502.13	.00	.03	2007.58	.02	.03
71	1545.08	.00	.03	2130.52	.02	.03
72	1537.71	.00	.03	2118.06	.02	.03
73	1523.15	.00	.03	2048.65	.02	.03
74	1589.69	.01	.03	2137.44	.02	.03
75	1600.08	.01	.03	2039.31	.02	.03
76	1599.04	.01	.03	2022.05	.02	.03
77	1560.93	.00	.03	1997.87	.02	.03
78	1560.90	.00	.03	2030.37	.02	.03
79	1641.50	.01	.03	2082.19	.02	.03
80	1554.02	.00	.03	2041.30	.02	.03
81	1625.75	.01	.03	2038.44	.02	.03
82	1636.12	.01	.03	2111.28	.02	.03
83	1590.76	.01	.03	2121.67	.02	.03
84	1523.50	.00	.03	2039.00	.02	.03
85	1614.34	.01	.03	2075.67	.02	.03
86	1527.54	.00	.03	2062.20	.02	.03
87	1536.19	.00	.03	2072.72	.02	.03
88	1596.52	.01	.03	2182.40	.03	.03
89	1578.01	.01	.03	2148.33	.03	.03

90	1603.56	.01	.03	2042.19	.02	.03
91	1617.59	.01	.03	2120.32	.02	.03
92	1638.28	.01	.03	2194.59	.03	.03
93	1563.42	.00	.03	2096.70	.02	.03
94	1700.06	.01	.03	2171.10	.03	.03
95	1645.29	.01	.03	2214.21	.03	.03
96	1592.62	.01	.03	2146.94	.02	.03
97	1524.35	.00	.03	2091.01	.02	.03
98	1475.90	.00	.03	2003.93	.02	.03
99	1518.67	.00	.03	2079.94	.02	.03
100	1588.00	.01	.03	2144.80	.02	.03
Mean	**1581.59**	**.01**	**.03**	**2093.92**	**.02**	**.03**
SD	**51.83**	**.00**	**.00**	**95.00**	**.00**	**.00**

Appendix D - Example of Validity Check Computation

This is an example and visual representation of how the validity checks were computed. This example is based on the following condition: 200 participants, 10 items, 5% prevalence, 100% severity, “Carless” CR Type. Step 1 entailed computing the four CR indices (*l_z*, *G*, *D*, and *MLS)* across the entire combined dataset (see Step 3 in Figure 3). This yielded four unique CR index values for each participant (i.e., row). Step 2a involved averaging the CR index values for the subset of careless responses (highlighted in red below; 10 items in this case) and Step 2b involved averaging the CR index values for the subset of careful responses (highlighted in green below; 190 items in this case). This produced an average CR index (for each of the four indices) for the subset of careless responses and an average CR index value (for each of the four indices) for the subset of careful responses. Step 3 involved repeating this process 100 times (due to there being 100 replications for each condition). The 100 CR index values for the careful responses were then averaged and the 100 CR index values for the careless responses were also averaged. This entire process was done for all 144 conditions (Appendix E displays all these values).



Step 1: Compute CR indices across entire dataset.

Step 2a: Average CR indices across careless subset.

Step 2b: Average CR indices across careful subset.

Item	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	1	0	0	0	0	1	2	2	0	0
2	3	1	0	4	2	2	0	3	2	2
3	3	3	2	3	3	3	1	4	3	4
4	2	4	3	3	3	0	3	3	1	2
5	3	3	2	2	0	2	1	1	3	1
6	0	0	0	0	0	0	0	0	0	0
7	3	3	4	4	4	3	4	4	0	3
8	3	3	2	3	3	4	4	3	1	2
9	3	3	3	4	3	4	3	3	1	3
10	1	2	1	3	4	1	2	0	2	1
11	3	2	2	3	3	4	0	1	2	1
12	4	4	3	3	3	4	2	2	3	2
13	2	3	3	3	3	0	2	1	2	3
14	4	4	4	3	4	4	2	2	4	3
15	4	4	3	4	4	4	3	3	4	3
16	3	3	4	3	3	2	1	2	2	3
17	4	3	2	3	3	4	1	3	2	2
18	1	4	4	4	4	4	3	3	3	3
19	4	3	4	3	3	3	3	2	4	1
20	4	4	4	4	4	0	3	3	4	4
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
190	2	4	2	3	4	4	3	1	3	0
191	3	3	4	3	3	3	3	2	3	3
192	3	3	4	3	4	0	3	2	2	3
193	4	3	4	3	4	3	4	3	4	3
194	3	3	1	3	3	2	3	2	3	3
195	4	4	4	4	4	4	3	3	4	4
196	2	3	2	3	1	0	3	1	3	1
197	4	4	4	3	3	1	2	0	3	3
198	4	4	4	4	4	0	4	3	4	2
199	4	4	4	4	4	4	4	2	4	1
200	4	3	4	4	4	4	1	2	4	1

Figure 7. Visual example of how CR validity checks are computed

Appendix E - Validity Checks for all Study 2 Conditions

Table 31. Summary of validity check items (200 participants/10 items)

			IRT Person-Fit Indices				Traditional CR Indices			
Condition			l_z		*G*		*D*		*MLS*	
CR Type	Prevalence	Severity	CR$_S$	C$_S$	CR$_S$	C$_S$	CR$_S$	C$_S$	CR$_S$	C$_S$
CR$_C$	5%	20%	**-1.04**	0.14	**37.09**	16.39	**17.96**	9.53	2.86	**3.07**
CR$_C$	5%	50%	**-2.13**	0.32	**57.62**	16.50	**25.02**	9.16	2.87	**3.07**
CR$_C$	5%	100%	**-2.01**	0.12	**52.13**	16.41	**29.61**	8.92	3.19	**3.07**
CR$_C$	15%	20%	**-0.70**	0.24	**33.95**	16.81	**14.01**	9.23	2.89	**3.07**
CR$_C$	15%	50%	**-1.12**	0.44	**51.99**	17.43	**17.39**	8.64	2.91	**3.07**
CR$_C$	15%	100%	**-1.17**	0.36	**48.63**	16.96	**20.70**	8.05	2.91	**3.07**
CR$_C$	30%	20%	**-0.39**	0.32	**31.18**	17.73	**11.93**	9.10	2.88	**3.07**
CR$_C$	30%	50%	**-0.52**	0.48	**45.13**	18.89	**13.77**	8.31	2.94	**3.07**
CR$_C$	30%	100%	**-0.72**	0.58	**46.68**	18.29	**16.17**	7.28	**3.13**	3.07
RR$_C$	5%	20%	**-0.99**	0.13	**34.39**	16.39	**17.39**	9.56	2.79	**3.07**
RR$_C$	5%	50%	**-2.19**	0.30	**55.72**	16.50	**24.95**	9.16	2.69	**3.07**
RR$_C$	5%	100%	**-2.84**	0.31	**71.78**	16.48	**32.41**	8.77	2.46	**3.07**
RR$_C$	15%	20%	**-0.68**	0.18	**31.71**	16.74	**13.81**	9.27	2.88	**3.07**
RR$_C$	15%	50%	**-1.30**	0.44	**48.33**	17.22	**17.62**	8.60	2.68	**3.07**
RR$_C$	15%	100%	**-1.54**	0.49	**62.27**	17.28	**22.17**	7.79	2.47	**3.07**
RR$_C$	30%	20%	**-0.36**	0.27	**29.43**	17.43	**11.90**	9.12	2.86	**3.07**
RR$_C$	30%	50%	**-0.65**	0.56	**42.57**	18.67	**13.97**	8.23	2.68	**3.07**
RR$_C$	30%	100%	**-0.75**	0.60	**57.45**	19.17	**16.89**	6.97	2.44	**3.07**
MRR$_C$	5%	20%	**-1.10**	0.15	**38.50**	16.39	**18.87**	9.48	2.78	**3.07**
MRR$_C$	5%	50%	**-2.71**	0.29	**68.68**	16.50	**27.46**	9.03	2.52	**3.07**
MRR$_C$	5%	100%	**-3.92**	0.28	**100.90**	16.44	**41.37**	8.30	2.19	**3.07**
MRR$_C$	15%	20%	**-0.77**	0.19	**35.61**	16.83	**14.65**	9.12	2.87	**3.07**
MRR$_C$	15%	50%	**-1.75**	0.51	**61.17**	17.30	**20.48**	8.09	2.60	**3.07**
MRR$_C$	15%	100%	**-2.68**	0.65	**96.98**	16.93	**28.52**	6.67	2.16	**3.07**
MRR$_C$	30%	20%	**-0.42**	0.29	**32.37**	17.64	**12.28**	8.95	2.84	**3.07**
MRR$_C$	30%	50%	**-1.06**	0.64	**56.36**	18.92	**15.48**	7.58	2.55	**3.07**
MRR$_C$	30%	100%	**-1.69**	0.89	**93.48**	18.15	**20.20**	5.56	2.18	**3.07**
IR$_C$	5%	20%	**-1.12**	0.19	**37.47**	16.39	**17.12**	9.57	2.96	**3.07**
IR$_C$	5%	50%	**-1.57**	0.31	**49.42**	16.47	**18.77**	9.49	**5.31**	3.07
IR$_C$	5%	100%	**-0.61**	-0.30	13.64	**16.37**	**14.51**	9.71	**10.00**	3.07
IR$_C$	15%	20%	**-0.77**	0.20	**33.39**	16.80	**12.75**	9.46	3.06	**3.07**
IR$_C$	15%	50%	**-0.33**	0.28	**41.95**	17.26	**11.51**	9.67	**5.37**	3.07
IR$_C$	15%	100%	-0.21	**-0.45**	9.56	**16.79**	6.86	**10.50**	**10.00**	3.07
IR$_C$	30%	20%	**-0.40**	0.30	**30.78**	17.64	**10.77**	9.60	3.03	**3.07**
IR$_C$	30%	50%	**0.14**	0.22	**37.10**	18.82	8.73	**10.47**	**5.32**	3.07
IR$_C$	30%	100%	-0.35	**-1.05**	5.82	**17.79**	3.91	**12.54**	**10.00**	3.07

Note. CR$_S$ = Careless subset. C$_S$ = Careful subset. CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. l_z = standardized loglikelihood. *G* = Guttman error. *D* = Mahalanobis distance. *MLS* = Maximum longstring. Bolded CR index values for each CR index pair indicate higher amounts of CR. Values for each condition represent mean values averaged across 100 replications.

Table 32. Summary of validity check items (200 participants/60 items)

			IRT Person-Fit Indices				Traditional CR Indices			
Condition			l_z		*G*		*D*		*MLS*	
CR Type	Prevalence	Severity	CR$_S$	C$_S$	CR$_S$	C$_S$	CR$_S$	C$_S$	CR$_S$	C$_S$
CR$_C$	5%	20%	**-3.28**	-0.45	**1569.13**	673.20	**82.35**	58.51	**5.55**	5.28
CR$_C$	5%	50%	**-5.76**	-0.20	**2239.54**	675.91	**97.16**	57.73	**5.93**	5.28
CR$_C$	5%	100%	**-5.49**	0.01	**2286.83**	673.63	**110.83**	57.01	**6.36**	5.28
CR$_C$	15%	20%	**-2.26**	-0.02	**1356.41**	690.93	**71.17**	57.68	**5.50**	5.27
CR$_C$	15%	50%	**-3.53**	0.50	**1966.09**	706.36	**81.80**	55.80	**5.91**	5.27
CR$_C$	15%	100%	**-3.81**	0.72	**2136.88**	699.21	**96.34**	53.23	**6.11**	5.27
CR$_C$	30%	20%	**-1.68**	-0.09	**1241.43**	734.10	**65.82**	57.08	**5.50**	5.27
CR$_C$	30%	50%	**-1.78**	0.53	**1756.25**	789.17	**72.84**	45.07	**6.09**	5.27
CR$_C$	30%	100%	**-2.40**	1.06	**1905.24**	767.35	**83.77**	49.38	**6.02**	5.27
RR$_C$	5%	20%	**-3.02**	-0.33	**1414.07**	672.14	**84.17**	58.41	5.07	**5.28**
RR$_C$	5%	50%	**-5.89**	-0.04	**2222.93**	675.42	**103.16**	57.41	4.67	**5.28**
RR$_C$	5%	100%	**-7.50**	0.34	**2861.36**	675.55	**120.50**	56.50	3.98	**5.28**
RR$_C$	15%	20%	**-2.32**	-0.20	**1318.50**	690.72	**72.90**	57.37	5.11	**5.27**
RR$_C$	15%	50%	**-4.07**	0.41	**1956.78**	704.94	**86.74**	54.93	4.80	**5.27**
RR$_C$	15%	100%	**-4.40**	0.83	**2667.78**	711.29	**105.37**	51.64	2.88	**5.27**
RR$_C$	30%	20%	**-1.53**	-0.07	**1175.95**	728.87	**66.57**	56.76	5.11	**5.27**
RR$_C$	30%	50%	**-2.27**	0.73	**1744.74**	781.56	**76.04**	52.70	4.74	**5.27**
RR$_C$	30%	100%	**-2.33**	1.01	**2373.74**	804.15	**90.32**	46.58	3.96	**5.27**
MRR$_C$	5%	20%	**-3.28**	-0.39	**1575.02**	672.72	**92.44**	57.98	5.01	**5.28**
MRR$_C$	5%	50%	**-6.57**	-0.18	**2686.55**	675.85	**119.02**	56.58	4.49	**5.28**
MRR$_C$	5%	100%	**-10.36**	0.43	**4094.53**	674.14	**143.38**	55.30	3.26	**5.28**
MRR$_C$	15%	20%	**-2.38**	-0.19	**1434.58**	690.50	**78.27**	56.42	5.05	**5.27**
MRR$_C$	15%	50%	**-4.73**	0.52	**2452.55**	705.92	**99.92**	52.60	4.45	**5.27**
MRR$_C$	15%	100%	**-7.09**	1.12	**3955.74**	697.79	**125.75**	48.04	3.28	**5.27**
MRR$_C$	30%	20%	**-1.60**	0.03	**1311.59**	730.56	**69.59**	55.46	5.03	**5.27**
MRR$_C$	30%	50%	**-3.08**	0.92	**2231.80**	776.60	**84.21**	49.19	4.52	**5.27**
MRR$_C$	30%	100%	**-4.60**	1.59	**3783.39**	759.22	**106.16**	39.79	3.30	**5.27**
IR$_C$	5%	20%	**-3.39**	-0.50	**1484.28**	672.30	**67.86**	59.27	**12.41**	5.28
IR$_C$	5%	50%	**-4.94**	-0.08	**1901.36**	674.62	57.87	**59.80**	**30.37**	5.28
IR$_C$	5%	100%	**-1.89**	-0.20	**703.96**	671.49	25.00	**51.53**	**60.00**	5.28
IR$_C$	15%	20%	**-2.44**	-0.21	**1326.47**	689.96	57.64	**60.06**	**12.41**	5.27
IR$_C$	15%	50%	**-1.76**	0.10	**1631.65**	703.83	44.21	**62.43**	**30.37**	5.27
IR$_C$	15%	100%	-0.28	**-0.42**	518.03	**691.81**	9.88	**68.49**	**60.00**	5.27
IR$_C$	30%	20%	**-1.65**	0.04	**1193.72**	729.12	54.12	**62.09**	**12.41**	5.27
IR$_C$	30%	50%	**-0.21**	0.02	**1396.80**	777.66	39.69	**68.28**	**30.36**	5.27
IR$_C$	30%	100%	0.97	**-0.82**	324.33	**746.94**	5.70	**83.11**	**60.00**	5.27

Note. CR$_S$ = Careless subset. C$_S$ = Careful subset. CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. l_z = standardized loglikelihood. *G* = Guttman error. *D* = Mahalanobis distance. *MLS* = Maximum longstring. Bolded CR index values for each CR index pair indicate higher amounts of CR. Values for each condition represent mean values averaged across 100 replications.

Table 33. Summary of validity check items (500 participants/10 items)

			IRT Person-Fit Indices				Traditional CR Indices			
Condition			*l_z*		*G*		*D*		*MLS*	
CR Type	Prevalence	Severity	CR$_S$	C$_S$	CR$_S$	C$_S$	CR$_S$	C$_S$	CR$_S$	C$_S$
CR$_C$	5%	20%	**-1.21**	0.04	**38.87**	16.05	**18.72**	9.52	2.77	**2.90**
CR$_C$	5%	50%	**-2.15**	0.28	**56.53**	16.12	**25.30**	9.17	**2.95**	2.90
CR$_C$	5%	100%	**-2.28**	0.18	**61.96**	16.06	**32.69**	8.78	**3.17**	2.90
CR$_C$	15%	20%	**-0.67**	0.29	**34.13**	16.43	**14.25**	9.23	2.82	2.90
CR$_C$	15%	50%	**-1.29**	0.40	**49.36**	16.83	**17.54**	8.65	**2.95**	2.90
CR$_C$	15%	100%	**-1.42**	0.43	**55.49**	16.64	**21.83**	7.89	**3.20**	2.90
CR$_C$	30%	20%	**-0.40**	0.32	**30.42**	17.49	**11.98**	9.12	2.81	**2.91**
CR$_C$	30%	50%	**-0.52**	0.50	**44.22**	18.72	**13.71**	8.38	2.91	2.91
CR$_C$	30%	100%	**-0.81**	0.62	**52.28**	18.46	**16.52**	7.18	**3.15**	2.91
RR$_C$	5%	20%	**-1.04**	0.08	**34.63**	16.03	**17.39**	9.59	2.78	**2.90**
RR$_C$	5%	50%	**-2.31**	0.12	**54.66**	16.09	**24.96**	9.19	2.68	**2.90**
RR$_C$	5%	100%	**-3.21**	0.36	**81.30**	16.13	**34.14**	8.71	2.44	**2.90**
RR$_C$	15%	20%	**-0.65**	0.30	**33.10**	16.41	**14.17**	9.24	2.81	**2.90**
RR$_C$	15%	50%	**-1.49**	0.33	**50.10**	16.75	**17.92**	8.58	2.70	**2.90**
RR$_C$	15%	100%	**-1.67**	0.51	**72.53**	16.94	**23.41**	7.61	2.48	**2.90**
RR$_C$	30%	20%	**-0.34**	0.31	**28.23**	17.27	**11.83**	9.19	2.78	**2.91**
RR$_C$	30%	50%	**-0.75**	0.52	**42.22**	18.55	**13.75**	8.37	2.70	**2.91**
RR$_C$	30%	100%	**-0.76**	0.61	**61.39**	19.27	**17.13**	6.92	2.46	**2.91**
MRR$_C$	5%	20%	**-1.21**	0.01	**38.05**	16.03	**18.96**	9.51	2.76	**2.90**
MRR$_C$	5%	50%	**-2.59**	0.20	**67.03**	16.11	**29.74**	8.94	2.59	**2.90**
MRR$_C$	5%	100%	**-4.27**	0.40	**110.12**	16.06	**46.06**	8.08	2.18	**2.90**
MRR$_C$	15%	20%	**-0.74**	0.27	**35.19**	16.39	**14.80**	9.13	2.79	**2.90**
MRR$_C$	15%	50%	**-1.70**	0.49	**60.64**	16.80	**20.43**	8.14	2.61	**2.90**
MRR$_C$	15%	100%	**-2.71**	0.65	**103.91**	16.58	**29.64**	6.51	2.17	**2.90**
MRR$_C$	30%	20%	**-0.39**	0.31	**31.67**	17.36	**12.26**	9.00	2.77	**2.91**
MRR$_C$	30%	50%	**-1.09**	0.59	**54.55**	18.51	**15.29**	7.71	2.58	**2.91**
MRR$_C$	30%	100%	**-1.64**	0.84	**99.38**	18.19	**20.50**	5.47	2.17	**2.91**
IR$_C$	5%	20%	**-1.24**	0.05	**37.81**	16.03	**17.37**	9.59	**2.92**	2.90
IR$_C$	5%	50%	**-1.69**	0.29	**47.45**	16.10	**19.28**	9.49	**5.34**	2.90
IR$_C$	5%	100%	**-2.26**	-0.98	**20.03**	16.03	**15.84**	9.67	**10.00**	2.90
IR$_C$	15%	20%	**-0.81**	0.17	**33.04**	16.40	**13.12**	9.43	**2.96**	2.90
IR$_C$	15%	50%	**-0.39**	0.30	**40.45**	16.78	**11.94**	9.63	**5.33**	2.90
IR$_C$	15%	100%	**-1.56**	-1.03	15.10	**16.54**	7.26	**10.46**	**10.00**	2.90
IR$_C$	30%	20%	**-0.43**	0.29	**29.70**	17.33	**10.98**	9.55	**2.94**	2.91
IR$_C$	30%	50%	0.20	**0.18**	**34.72**	18.44	8.91	**10.44**	**5.23**	2.91
IR$_C$	30%	100%	**-0.94**	-0.80	9.91	**18.03**	4.04	**12.52**	**10.00**	2.91

Note. CR$_S$ = Careless subset. C$_S$ = Careful subset. CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. *l_z* = standardized loglikelihood. *G* = Guttman error. *D* = Mahalanobis distance. *MLS* = Maximum longstring. Bolded CR index values for each CR index pair indicate higher amounts of CR. Values for each condition represent mean values averaged across 100 replications.

Table 34. Summary of validity check items (500 participants/60 items)

			IRT Person-Fit Indices				Traditional CR Indices			
Condition			l_z		*G*		*D*		*MLS*	
CR Type	Prevalence	Severity	CR$_S$	C$_S$	CR$_S$	C$_S$	CR$_S$	C$_S$	**CR$_S$**	C$_S$
CR$_C$	5%	20%	**-3.13**	-0.03	**1570.64**	717.36	**88.95**	58.35	**5.58**	5.22
CR$_C$	5%	50%	**-5.91**	0.16	**2295.98**	719.58	**111.61**	57.16	**5.78**	5.22
CR$_C$	5%	100%	**-5.70**	0.25	**2349.56**	718.11	**137.72**	55.78	**5.76**	5.22
CR$_C$	15%	20%	**-2.20**	0.15	**1438.98**	736.04	**74.49**	57.30	**5.40**	5.22
CR$_C$	15%	50%	**-3.50**	0.57	**2089.96**	750.61	**89.38**	54.67	**5.99**	5.22
CR$_C$	15%	100%	**-3.73**	0.74	**2120.07**	740.90	**109.16**	51.18	**6.16**	5.22
CR$_C$	30%	20%	**-1.36**	0.36	**1294.21**	780.26	**67.42**	56.65	**5.44**	5.22
CR$_C$	30%	50%	**-1.50**	0.71	**1830.60**	824.21	**76.75**	52.65	**5.84**	5.22
CR$_C$	30%	100%	**-2.41**	1.08	**1901.23**	805.94	**89.90**	47.01	**6.06**	5.22
RR$_C$	5%	20%	**-3.10**	-0.08	**1507.26**	717.13	**92.22**	58.18	5.07	**5.22**
RR$_C$	5%	50%	**-6.12**	0.20	**2327.73**	719.89	**120.71**	56.68	4.71	**5.22**
RR$_C$	5%	100%	**-7.57**	0.46	**2925.62**	720.05	**153.41**	54.96	3.99	**5.22**
RR$_C$	15%	20%	**-1.97**	0.28	**1349.73**	733.14	**76.55**	56.94	5.00	**5.22**
RR$_C$	15%	50%	**-3.93**	0.59	**2043.48**	748.30	**95.26**	53.46	4.68	**5.22**
RR$_C$	15%	100%	**-4.51**	0.84	**2694.06**	753.35	**121.04**	49.09	3.96	**5.22**
RR$_C$	30%	20%	**-1.31**	0.28	**1230.20**	773.49	**68.62**	56.13	5.00	**5.22**
RR$_C$	30%	50%	**-2.06**	0.68	**1786.22**	824.47	**79.80**	51.34	4.67	**5.22**
RR$_C$	30%	100%	**-2.25**	1.00	**2322.81**	843.55	**96.96**	43.99	3.95	**5.22**
MRR$_C$	5%	20%	**-3.30**	0.02	**1660.70**	717.44	**104.35**	57.54	4.98	**5.22**
MRR$_C$	5%	50%	**-6.98**	0.19	**2775.41**	719.63	**149.54**	55.16	4.49	**5.22**
MRR$_C$	5%	100%	**-10.56**	0.62	**4153.25**	717.97	**204.84**	52.25	3.27	**5.22**
MRR$_C$	15%	20%	**-2.23**	0.27	**1512.08**	734.69	**83.57**	55.70	4.94	**5.22**
MRR$_C$	15%	50%	**-4.67**	0.70	**2540.95**	747.85	**114.34**	50.27	4.44	**5.22**
MRR$_C$	15%	100%	**-7.25**	1.16	**4005.33**	739.25	**157.26**	42.69	3.27	**5.22**
MRR$_C$	30%	20%	**-1.34**	0.50	**1376.68**	776.23	**72.27**	54.57	4.94	**5.22**
MRR$_C$	30%	50%	**-2.90**	1.22	**2304.82**	817.23	**90.65**	46.69	4.45	**5.22**
MRR$_C$	30%	100%	**-4.78**	1.61	**3815.60**	796.47	**119.01**	34.54	3.27	**5.22**
IR$_C$	5%	20%	**-3.19**	-0.06	**1540.03**	717.29	**69.13**	59.39	**12.41**	5.22
IR$_C$	5%	50%	**-4.50**	0.18	**1926.71**	719.22	**57.47**	60.01	**30.34**	5.22
IR$_C$	5%	100%	**-1.17**	-0.05	679.51	**716.85**	**23.42**	61.80	**60.00**	5.22
IR$_C$	15%	20%	**-2.29**	0.15	**1394.73**	734.54	57.26	**60.34**	**12.39**	5.22
IR$_C$	15%	50%	**-1.11**	0.23	**1695.37**	746.68	42.11	**63.02**	**30.33**	5.22
IR$_C$	15%	100%	**-0.72**	-0.64	502.37	**736.17**	9.39	**68.79**	**60.00**	5.22
IR$_C$	30%	20%	**-1.37**	0.42	**1246.91**	775.55	53.27	**62.71**	**12.40**	5.22
IR$_C$	30%	50%	0.73	**-0.07**	**1448.22**	814.99	36.99	**69.69**	**30.25**	5.22
IR$_C$	30%	100%	0.34	**-0.90**	315.44	**790.81**	4.88	**83.45**	**60.00**	5.22

Note. CR$_S$ = Careless subset. C$_S$ = Careful subset. CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. l_z = standardized loglikelihood. *G* = Guttman error. *D* = Mahalanobis distance. *MLS* = Maximum longstring. Bolded CR index values for each CR index pair indicate higher amounts of CR. Values for each condition represent mean values averaged across 100 replications.

Appendix F - Model Fit Values Across All Conditions

Table 35. Model fit value across all simulation conditions (200 participants/10 items)

			IRT Model Fit						CFA Model Fit					
Type	Prevalence	Severity	*M2*		*RMSEA*		*SRMSR*		χ^2		*RMSEA*		*SRMSR*	
CR$_C$	5%	20%	**10.64**	[6.35]	**.03**	[.03]	**.05**	[.01]	**50.14**	[14.89]	**.04**	[.03]	**.04**	[.01]
CR$_C$	5%	50%	**10.63**	[6.47]	**.04**	[.04]	**.07**	[.02]	**85.90**	[40.50]	**.08**	[.03]	**.07**	[.02]
CR$_C$	5%	100%	**5.31**	[3.47]	**.02**	[.03]	**.06**	[.03]	**69.30**	[23.52]	**.06**	[.03]	**.05**	[.01]
CR$_C$	15%	20%	**21.37**	[26.25]	**.07**	[.05]	**.06**	[.02]	**70.02**	[34.41]	**.06**	[.03]	**.06**	[.01]
CR$_C$	15%	50%	**15.57**	[16.76]	**.06**	[.06]	**.10**	[.03]	**158.60**	[70.09]	**.13**	[.04]	**.10**	[.03]
CR$_C$	15%	100%	**4.42**	[3.12]	**.02**	[.03]	**.07**	[.01]	**91.28**	[34.43]	**.09**	[.03]	**.05**	[.01]
CR$_C$	30%	20%	**39.96**	[26.78]	**.12**	[.06]	**.08**	[.01]	**94.51**	[36.28]	**.09**	[.03]	**.07**	[.02]
CR$_C$	30%	50%	**13.76**	[30.66]	**.04**	[.06]	**.13**	[.03]	**192.04**	[64.33]	**.15**	[.03]	**.12**	[.02]
CR$_C$	30%	100%	**4.62**	[2.99]	**.02**	[.03]	**.07**	[.01]	**94.28**	[33.35]	**.09**	[.02]	**.05**	[.01]
RR$_C$	5%	20%	**10.09**	[5.63]	**.03**	[.03]	**.05**	[.01]	**45.79**	[14.34]	**.03**	[.03]	**.04**	[.01]
RR$_C$	5%	50%	**11.06**	[8.35]	**.04**	[.04]	**.06**	[.01]	**72.39**	[28.69]	**.07**	[.03]	**.05**	[.01]
RR$_C$	5%	100%	**6.60**	[4.66]	**.02**	[.03]	**.06**	[.01]	**73.37**	[32.00]	**.07**	[.03]	**.05**	[.01]
RR$_C$	15%	20%	**15.36**	[10.03]	**.05**	[.04]	**.06**	[.01]	**57.04**	[21.81]	**.05**	[.03]	**.05**	[.01]
RR$_C$	15%	50%	**14.24**	[10.45]	**.06**	[.05]	**.09**	[.03]	**116.89**	[69.01]	**.10**	[.04]	**.08**	[.03]
RR$_C$	15%	100%	**5.58**	[3.48]	**.02**	[.03]	**.07**	[.01]	**94.98**	[31.22]	**.09**	[.03]	**.06**	[.01]
RR$_C$	30%	20%	**23.38**	[19.04]	**.07**	[.06]	**.06**	[.02]	**63.02**	[25.43]	**.06**	[.03]	**.06**	[.01]
RR$_C$	30%	50%	**16.43**	[16.02]	**.06**	[.06]	**.11**	[.03]	**148.36**	[59.10]	**.12**	[.03]	**.10**	[.03]
RR$_C$	30%	100%	**6.16**	[4.27]	**.03**	[.04]	**.07**	[.02]	**99.13**	[36.68]	**.09**	[.03]	**.07**	[.02]
MRR$_C$	5%	20%	**9.82**	[5.32]	**.03**	[.03]	**.05**	[.01]	**45.53**	[13.57]	**.03**	[.03]	**.04**	[.01]
MRR$_C$	5%	50%	**9.54**	[5.75]	**.03**	[.03]	**.06**	[.02]	**70.17**	[31.66]	**.06**	[.03]	**.05**	[.01]
MRR$_C$	5%	100%	**6.95**	[4.11]	**.03**	[.03]	**.07**	[.01]	**80.44**	[29.52]	**.08**	[.03]	**.06**	[.01]
MRR$_C$	15%	20%	**13.28**	[9.05]	**.04**	[.04]	**.06**	[.01]	**50.24**	[16.81]	**.04**	[.03]	**.05**	[.01]
MRR$_C$	15%	50%	**12.75**	[10.06]	**.05**	[.04]	**.08**	[.02]	**96.83**	[47.11]	**.09**	[.04]	**.07**	[.02]
MRR$_C$	15%	100%	**6.67**	[4.66]	**.03**	[.04]	**.07**	[.01]	**88.91**	[25.92]	**.09**	[.02]	**.06**	[.01]
MRR$_C$	30%	20%	**17.30**	[12.60]	**.06**	[.05]	**.06**	[.01]	**51.78**	[17.86]	**.04**	[.03]	**.05**	[.01]
MRR$_C$	30%	50%	**15.45**	[11.96]	**.06**	[.05]	**.10**	[.03]	**112.97**	[51.72]	**.10**	[.04]	**.09**	[.03]
MRR$_C$	30%	100%	**6.40**	[3.92]	**.03**	[.03]	**.07**	[.01]	**76.66**	[20.27]	**.07**	[.02]	**.06**	[.01]
IR$_C$	5%	20%	**12.03**	[6.61]	**.04**	[.03]	**.06**	[.01]	**59.93**	[21.28]	**.05**	[.03]	**.05**	[.01]
IR$_C$	5%	50%	**10.63**	[7.61]	**.04**	[.04]	**.07**	[.02]	**103.65**	[44.74]	**.09**	[.03]	**.06**	[.02]
IR$_C$	5%	100%	**5.22**	[3.49]	**.02**	[.03]	**.05**	[.01]	**70.45**	[25.05]	**.07**	[.03]	**.04**	[.01]
IR$_C$	15%	20%	**32.17**	[21.22]	**.10**	[.06]	**.08**	[.02]	**113.16**	[47.20]	**.10**	[.03]	**.07**	[.01]
IR$_C$	15%	50%	**10.09**	[8.94]	**.04**	[.04]	**.11**	[.03]	**187.48**	[73.30]	**.14**	[.04]	**.11**	[.03]
IR$_C$	15%	100%	**4.64**	[2.92]	**.02**	[.03]	**.05**	[.01]	**120.28**	[48.27]	**.11**	[.03]	**.05**	[.01]
IR$_C$	30%	20%	**75.35**	[42.52]	**.19**	[.07]	**.10**	[.07]	**190.02**	[67.01]	**.15**	[.03]	**.09**	[.02]
IR$_C$	30%	50%	**7.35**	[6.24]	**.02**	[.03]	**.14**	[.03]	**232.06**	[72.40]	**.17**	[.03]	**.14**	[.03]
IR$_C$	30%	100%	**5.87**	[3.59]	**.03**	[.03]	**.05**	[.02]	**183.88**	[71.97]	**.14**	[.03]	**.04**	[.01]
***M* Baseline Model Fit Values**			**9.51**	[5.41]	**.02**	[.03]	**.05**	[.01]	**42.75**	[11.21]	**.03**	[.02]	**.04**	[.01]

Note. CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. Values in brackets represent the mean standard deviations of the model fit values. Values for each condition represent mean values averaged across 100 replications.

Table 36. Model fit value across all simulation conditions (200 participants/60 items)

			IRT Model Fit			CFA Model Fit		
Type	Prevalence	Severity	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
CR$_C$	5%	20%	**2107.40** [296.99]	**.04** [.01]	**.06** [.01]	**2588.50** [220.27]	**.05** [.01]	**.06** [.01]
CR$_C$	5%	50%	**2559.05** [696.00]	**.05** [.02]	**.08** [.01]	**3082.82** [397.08]	**.06** [.01]	**.08** [.01]
CR$_C$	5%	100%	**1475.27** [111.60]	**.00** [.01]	**.08** [.01]	**3191.90** [339.27]	**.07** [.01]	**.07** [.01]
CR$_C$	15%	20%	**3155.31** [581.41]	**.07** [.01]	**.08** [.01]	**2957.83** [242.21]	**.06** [.01]	**.07** [.01]
CR$_C$	15%	50%	**4295.68** [1083.71]	**.09** [.02]	**.11** [.01]	**3939.23** [366.36]	**.08** [.01]	**.11** [.02]
CR$_C$	15%	100%	**1477.75** [114.71]	**.00** [.01]	**.08** [.01]	**3548.15** [320.99]	**.07** [.01]	**.08** [.01]
CR$_C$	30%	20%	**4299.47** [1112.41]	**.09** [.02]	**.09** [.01]	**3282.40** [342.63]	**.07** [.01]	**.09** [.01]
CR$_C$	30%	50%	**3685.39** [804.01]	**.02** [.02]	**.14** [.01]	**4399.17** [301.51]	**.09** [.00]	**.15** [.02]
CR$_C$	30%	100%	**1496.59** [150.40]	**.01** [.01]	**.08** [.01]	**3532.55** [269.13]	**.07** [.01]	**.08** [.01]
RR$_C$	5%	20%	**1829.93** [146.37]	**.03** [.01]	**.06** [.00]	**2425.58** [160.72]	**.04** [.01]	**.05** [.00]
RR$_C$	5%	50%	**2309.77** [314.75]	**.05** [.01]	**.07** [.01]	**2889.56** [272.73]	**.06** [.01]	**.07** [.01]
RR$_C$	5%	100%	**1723.69** [136.93]	**.02** [.01]	**.08** [.01]	**3224.96** [297.11]	**.07** [.01]	**.08** [.01]
RR$_C$	15%	20%	**2622.87** [544.87]	**.06** [.01]	**.07** [.01]	**2745.57** [226.19]	**.05** [.01]	**.07** [.01]
RR$_C$	15%	50%	**4373.36** [1060.14]	**.09** [.02]	**.10** [.01]	**3542.99** [318.48]	**.07** [.01]	**.10** [.01]
RR$_C$	15%	100%	**1971.43** [191.74]	**.04** [.01]	**.09** [.01]	**3658.27** [300.56]	**.08** [.01]	**.09** [.01]
RR$_C$	30%	20%	**3039.76** [803.78]	**.07** [.02]	**.08** [.01]	**2859.66** [293.38]	**.06** [.01]	**.07** [.01]
RR$_C$	30%	50%	**6234.00** [2241.01]	**.12** [.03]	**.12** [.01]	**4011.92** [340.73]	**.08** [.01]	**.12** [.01]
RR$_C$	30%	100%	**2142.67** [326.06]	**.04** [.01]	**.09** [.01]	**3540.54** [232.28]	**.07** [.00]	**.09** [.01]
MRR$_C$	5%	20%	**1791.04** [115.18]	**.03** [.01]	**.06** [.00]	**2404.51** [145.95]	**.04** [.00]	**.05** [.00]
MRR$_C$	5%	50%	**2233.67** [273.26]	**.05** [.01]	**.07** [.01]	**2887.40** [260.53]	**.06** [.01]	**.07** [.01]
MRR$_C$	5%	100%	**2016.95** [127.52]	**.04** [.01]	**.08** [.01]	**3558.66** [352.12]	**.07** [.01]	**.08** [.01]
MRR$_C$	15%	20%	**2148.80** [254.55]	**.04** [.01]	**.07** [.01]	**2525.50** [167.84]	**.05** [.01]	**.06** [.00]
MRR$_C$	15%	50%	**3746.35** [659.97]	**.08** [.01]	**.09** [.01]	**3315.54** [268.49]	**.07** [.01]	**.09** [.01]
MRR$_C$	15%	100%	**2208.95** [124.04]	**.05** [.01]	**.10** [.01]	**3971.89** [318.63]	**.08** [.01]	**.09** [.01]
MRR$_C$	30%	20%	**2315.21** [377.16]	**.05** [.01]	**.07** [.01]	**2540.56** [174.28]	**.05** [.01]	**.06** [.01]
MRR$_C$	30%	50%	**5393.55** [1116.07]	**.11** [.02]	**.11** [.01]	**3403.13** [269.86	**.07** [.01]	**.10** [.01]
MRR$_C$	30%	100%	**2257.78** [130.99]	**.05** [.00]	**.10** [.01]	**3558.22** [213.77]	**.07** [.00]	**.08** [.01]
IR$_C$	5%	20%	**2360.93** [335.73]	**.05** [.01]	**.07** [.01]	**2750.68** [221.61]	**.05** [.01]	**.06** [.01]
IR$_C$	5%	50%	**2918.11** [704.19]	**.06** [.02]	**.08** [.01]	**3423.35** [395.12]	**.07** [.01]	**.08** [.01]
IR$_C$	5%	100%	**1344.64** [73.28]	**.00** [.00]	**.07** [.01]	**3309.99** [399.30]	**.07** [.01]	**.06** [.01]
IR$_C$	15%	20%	**4795.10** [846.58]	**.10** [.01]	**.09** [.01]	**3660.47** [335.96]	**.08** [.01]	**.09** [.01]
IR$_C$	15%	50%	**4018.11** [728.83]	**.09** [.01]	**.13** [.01]	**4665.40** [433.90]	**.09** [.01]	**.14** [.01]
IR$_C$	15%	100%	**1508.51** [100.65]	**.01** [.01]	**.08** [.02]	**4411.13** [513.99]	**.09** [.01]	**.07** [.01]
IR$_C$	30%	20%	**7623.87** [1283.27]	**.14** [.01]	**.12** [.01]	**4705.55** [415.26]	**.09** [.01]	**.12** [.01]
IR$_C$	30%	50%	**6163.51** [1403.93]	**.12** [.02]	**.16** [.01]	**5417.64** [460.22]	**.10** [.01]	**.16** [.01]
IR$_C$	30%	100%	**1970.19** [162.85]	**.04** [.01]	**.07** [.01]	**5846.15** [638.98]	**.11** [.01]	**.08** [.01]
***M* Baseline Model Fit Values**			**1598.32** [56.39]	**.01** [.01]	**.05** [.01]	**2170.53** [104.57]	**.04** [.00]	**.05** [.00]

Note CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. Values in brackets represent the mean standard deviations of the model fit values. Values for each condition represent mean values averaged across 100 replications.

Table 37. Model fit value across all simulation conditions (500 participants/10 items)

			IRT Model Fit			CFA Model Fit		
Type	Prevalence	Severity	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
CR$_C$	5%	20%	**11.91** [10.33]	**.03** [.03]	**.04** [.01]	**67.19** [31.80]	**.04** [.02]	**.03** [.01]
CR$_C$	5%	50%	**14.34** [9.69]	**.04** [.03]	**.06** [.02]	**141.52** [70.26]	**.07** [.03]	**.05** [.02]
CR$_C$	5%	100%	**4.74** [3.16]	**.01** [.02]	**.06** [.01]	**163.70** [99.24]	**.08** [.03]	**.05** [.02]
CR$_C$	15%	20%	**40.47** [41.34]	**.08** [.05]	**.05** [.01]	**114.55** [73.42]	**.06** [.03]	**.05** [.02]
CR$_C$	15%	50%	**31.15** [40.02]	**.07** [.06]	**.09** [.02]	**286.20** [110.42]	**.12** [.03]	**.08** [.02]
CR$_C$	15%	100%	**5.70** [3.95]	**.02** [.02]	**.07** [.02]	**237.11** [156.13]	**.10** [.04]	**.06** [.03]
CR$_C$	30%	20%	**75.41** [64.65]	**.12** [.07]	**.06** [.02]	**159.39** [101.53]	**.08** [.04]	**.06** [.02]
CR$_C$	30%	50%	**22.71** [25.38]	**.05** [.05]	**.12** [.03]	**406.63** [137.07]	**.14** [.03]	**.11** [.03]
CR$_C$	30%	100%	**6.98** [7.41]	**.02** [.03]	**.08** [.03]	**273.76** [154.46]	**.11** [.04]	**.07** [.03]
RR$_C$	5%	20%	**9.48** [6.14]	**.02** [.02]	**.03** [.01]	**54.28** [17.36]	**.03** [.02]	**.03** [.01]
RR$_C$	5%	50%	**10.46** [7.41]	**.03** [.03]	**.05** [.01]	**95.23** [46.74]	**.05** [.02]	**.04** [.01]
RR$_C$	5%	100%	**5.58** [3.39]	**.01** [.02]	**.07** [.02]	**189.78** [99.27]	**.09** [.03]	**.06** [.02]
RR$_C$	15%	20%	**30.47** [34.22]	**.06** [.05]	**.05** [.02]	**87.97** [53.83]	**.05** [.03]	**.04** [.01]
RR$_C$	15%	50%	**25.69** [29.11]	**.06** [.04]	**.08** [.02]	**205.59** [102.81]	**.10** [.03]	**.07** [.02]
RR$_C$	15%	100%	**7.73** [7.72]	**.02** [.03]	**.08** [.03]	**241.04** [129.08]	**.10** [.04]	**.07** [.03]
RR$_C$	30%	20%	**38.91** [50.93]	**.07** [.06]	**.05** [.02]	**99.05** [66.80]	**.05** [.03]	**.04** [.02]
RR$_C$	30%	50%	**44.03** [49.21]	**.09** [.06]	**.10** [.03]	**314.56** [151.30]	**.12** [.03]	**.09** [.03]
RR$_C$	30%	100%	**17.42** [22.83]	**.05** [.05]	**.08** [.03]	**242.10** [139.45]	**.10** [.03]	**.08** [.03]
MRR$_C$	5%	20%	**9.22** [5.64]	**.02** [.02]	**.04** [.01]	**53.38** [19.03]	**.03** [.02]	**.03** [.01]
MRR$_C$	5%	50%	**10.05** [7.46]	**.03** [.03]	**.05** [.01]	**93.19** [45.69]	**.05** [.02]	**.04** [.01]
MRR$_C$	5%	100%	**6.28** [3.89]	**.02** [.02]	**.06** [.02]	**149.29** [75.85]	**.08** [.02]	**.05** [.02]
MRR$_C$	15%	20%	**18.97** [18.76]	**.04** [.04]	**.04** [.01]	**65.59** [29.33]	**.04** [.02]	**.04** [.01]
MRR$_C$	15%	50%	**21.18** [17.42]	**.06** [.04]	**.07** [.02]	**158.31** [82.69]	**.08** [.03]	**.06** [.02]
MRR$_C$	15%	100%	**6.60** [4.00]	**.02** [.02]	**.08** [.03]	**177.44** [92.38]	**.09** [.03]	**.06** [.02]
MRR$_C$	30%	20%	**24.22** [22.56]	**.05** [.05]	**.05** [.01]	**69.03** [33.43]	**.04** [.02]	**.04** [.01]
MRR$_C$	30%	50%	**24.32** [20.91]	**.06** [.04]	**.08** [.03]	**193.18** [95.31]	**.09** [.03]	**.07** [.02]
MRR$_C$	30%	100%	**10.23** [9.74]	**.03** [.03]	**.07** [.02]	**136.99** [62.05]	**.07** [.02]	**.06** [.02]
IR$_C$	5%	20%	**15.66** [14.55]	**.04** [.03]	**.04** [.01]	**85.97** [45.81]	**.05** [.02]	**.04** [.01]
IR$_C$	5%	50%	**17.59** [14.65]	**.05** [.04]	**.06** [.02]	**190.09** [100.96]	**.09** [.03]	**.06** [.02]
IR$_C$	5%	100%	**4.65** [3.48]	**.01** [.02]	**.06** [.01]	**179.09** [111.98]	**.08** [.04]	**.05** [.02]
IR$_C$	15%	20%	**72.19** [66.84]	**.12** [.07]	**.07** [.02]	**202.10** [138.50]	**.09** [.04]	**.06** [.02]
IR$_C$	15%	50%	**24.96** [40.88]	**.05** [.06]	**.11** [.02]	**399.32** [161.25]	**.14** [.03]	**.11** [.03]
IR$_C$	15%	100%	**4.35** [2.99]	**.01** [.02]	**.06** [.02]	**352.30** [206.84]	**.13** [.05]	**.06** [.03]
IR$_C$	30%	20%	**150.37** [106.72]	**.19** [.08]	**.09** [.02]	**351.92** [220.30]	**.13** [.05]	**.08** [.02]
IR$_C$	30%	50%	**12.77** [24.86]	**.02** [.04]	**.14** [.02]	**521.80** [154.65]	**.16** [.03]	**.13** [.03]
IR$_C$	30%	100%	**5.37** [3.98]	**.01** [.02]	**.07** [.03]	**556.62** [298.73]	**.16** [.05]	**.07** [.03]
***M* Baseline Model Fit Values**			**7.69** [4.80]	**.01** [.02]	**.03** [.00]	**46.69** [14.38]	**.02** [.02]	**.03** [.00]

Note. CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. Values in brackets represent the mean standard deviations of the model fit values. Values for each condition represent mean values averaged across 100 replications.

Table 38. Model fit value across all simulation conditions (500 participants/60 items)

			IRT Model Fit			CFA Model Fit		
Type	Prevalence	Severity	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
CR$_C$	5%	20%	**2723.48** [610.40]	**.04** [.01]	**.05** [.01]	**2983.04** [361.90]	**.04** [.01]	**.04** [.01]
CR$_C$	5%	50%	**3461.37** [632.77]	**.05** [.01]	**.06** [.01]	**3996.11** [467.55]	**.05** [.01]	**.06** [.01]
CR$_C$	5%	100%	**1499.86** [79.14]	**.00** [.00]	**.06** [.01]	**4032.25** [546.93]	**.05** [.01]	**.06** [.01]
CR$_C$	15%	20%	**5982.49** [1554.21]	**.07** [.01]	**.07** [.01]	**4116.89** [589.03]	**.05** [.01]	**.07** [.01]
CR$_C$	15%	50%	**6895.75** [1929.81]	**.08** [.01]	**.10** [.01]	**6293.80** [600.62]	**.07** [.01]	**.11** [.01]
CR$_C$	15%	100%	**1555.51** [128.40]	**.01** [.01]	**.07** [.01]	**4692.10** [540.47]	**.06** [.01]	**.06** [.01]
CR$_C$	30%	20%	**9026.91** [1742.70]	**.10** [.01]	**.09** [.01]	**5007.27** [511.00]	**.06** [.00]	**.08** [.01]
CR$_C$	30%	50%	**6186.68** [1670.73]	**.08** [.01]	**.13** [.01]	**7251.58** [581.17]	**.08** [.00]	**.14** [.01]
CR$_C$	30%	100%	**1669.75** [185.41]	**.01** [.01]	**.07** [.01]	**4823.84** [543.30]	**.06** [.01]	**.06** [.01]
RR$_C$	5%	20%	**2282.48** [394.95]	**.03** [.01]	**.04** [.01]	**2715.34** [334.73]	**.03** [.01]	**.04** [.01]
RR$_C$	5%	50%	**3156.35** [582.86]	**.05** [.01]	**.06** [.01]	**3654.79** [577.33]	**.05** [.01]	**.06** [.01]
RR$_C$	5%	100%	**1774.37** [139.10]	**.02** [.01]	**.06** [.01]	**4117.94** [537.11]	**.05** [.01]	**.06** [.01]
RR$_C$	15%	20%	**3778.69** [845.35]	**.05** [.01]	**.05** [.01]	**3285.94** [370.36]	**.04** [.01]	**.05** [.01]
RR$_C$	15%	50%	**7768.94** [1950.10]	**.09** [.01]	**.09** [.01]	**5222.46** [555.19]	**.06** [.01]	**.09** [.01]
RR$_C$	15%	100%	**2429.99** [353.20]	**.03** [.01]	**.08** [.01]	**5018.99** [588.28]	**.06** [.01]	**.08** [.01]
RR$_C$	30%	20%	**5206.02** [1345.59]	**.07** [.01]	**.07** [.01]	**3763.73** [462.23]	**.05** [.01]	**.06** [.01]
RR$_C$	30%	50%	**10396.78** [3868.17]	**.10** [.02]	**.11** [.01]	**6461.20** [673.66]	**.07** [.01]	**.11** [.01]
RR$_C$	30%	100%	**2728.20** [476.05]	**.04** [.01]	**.08** [.01]	**4672.99** [605.13]	**.06** [.01]	**.08** [.01]
MRR$_C$	5%	20%	**2077.44** [251.85]	**.03** [.01]	**.04** [.01]	**2569.89** [212.97]	**.03** [.00]	**.04** [.00]
MRR$_C$	5%	50%	**2746.32** [380.46]	**.04** [.01]	**.05** [.01]	**3409.29** [335.29]	**.04** [.00]	**.05** [.01]
MRR$_C$	5%	100%	**1993.38** [113.83]	**.02** [.00]	**.06** [.01]	**4238.49** [486.99]	**.05** [.01]	**.06** [.01]
MRR$_C$	15%	20%	**3053.32** [566.60]	**.04** [.01]	**.05** [.01]	**2907.55** [270.49]	**.04** [.00]	**.05** [.01]
MRR$_C$	15%	50%	**6149.46** [1451.03]	**.08** [.01]	**.08** [.01]	**4421.62** [469.73]	**.06** [.01]	**.08** [.01]
MRR$_C$	15%	100%	**2193.12** [131.62]	**.03** [.00]	**.08** [.01]	**4689.95** [467.10]	**.06** [.00]	**.07** [.01]
MRR$_C$	30%	20%	**3512.50** [742.56]	**.05** [.01]	**.06** [.01]	**3009.76** [292.22]	**.04** [.00]	**.05** [.01]
MRR$_C$	30%	50%	**10168.48** [2409.17]	**.10** [.02]	**.09** [.01]	**4798.41** [423.13]	**.06** [.00]	**.09** [.01]
MRR$_C$	30%	100%	**2281.95** [131.79]	**.03** [.00]	**.08** [.01]	**4003.70** [304.12]	**.05** [.00]	**.06** [.01]
IR$_C$	5%	20%	**3584.59** [626.30]	**.05** [.01]	**.05** [.01]	**3510.65** [393.16]	**.05** [.00]	**.05** [.01]
IR$_C$	5%	50%	**4358.97** [869.94]	**.06** [.01]	**.07** [.01]	**4843.02** [582.65]	**.06** [.01]	**.07** [.01]
IR$_C$	5%	100%	**1390.01** [65.78]	**.00** [.00]	**.06** [.01]	**4387.57** [655.53]	**.06** [.01]	**.05** [.01]
IR$_C$	15%	20%	**10081.90** [2174.94]	**.10** [.01]	**.09** [.01]	**5858.60** [779.74]	**.07** [.01]	**.08** [.01]
IR$_C$	15%	50%	**7197.41** [1670.51]	**.08** [.01]	**.12** [.01]	**7674.28** [773.92]	**.08** [.01]	**.13** [.01]
IR$_C$	15%	100%	**1586.20** [84.99]	**.01** [.01]	**.07** [.02]	**6443.84** [918.67]	**.07** [.01]	**.06** [.01]
IR$_C$	30%	20%	**17168.87** [3109.81]	**.14** [.01]	**.12** [.01]	**8419.98** [1000.27]	**.09** [.01]	**.11** [.01]
IR$_C$	30%	50%	**10565.91** [3557.80]	**.11** [.01]	**.15** [.01]	**9221.88** [864.71]	**.09** [.01]	**.15** [.01]
IR$_C$	30%	100%	**2219.44** [188.53]	**.03** [.00]	**.06** [.01]	**8812.19** [1148.66]	**.09** [.01]	**.07** [.01]
***M* Baseline Model Fit Values**			**1581.59** [51.83]	**.01** [.00]	**.03** [.00]	**2093.92** [95.00]	**.02** [.00]	**.03** [.00]

Note. CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. Values in brackets represent the mean standard deviations of the model fit values. Values for each condition represent mean values averaged across 100 replications.

Appendix G - Plots for Significant IRT Model Fit Interactions

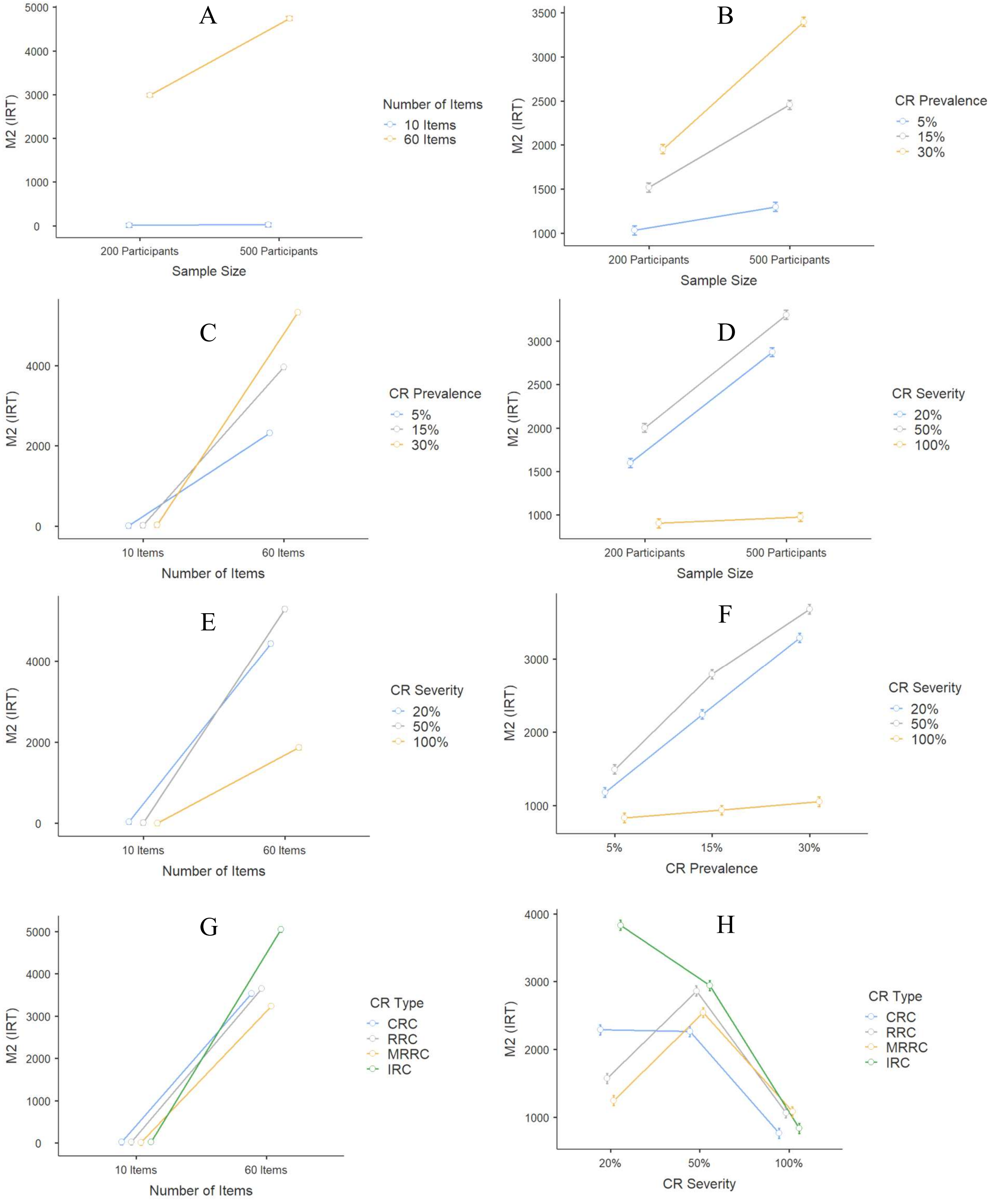


Figure 8. Plots of meaningful interactions for the IRT *M2* index

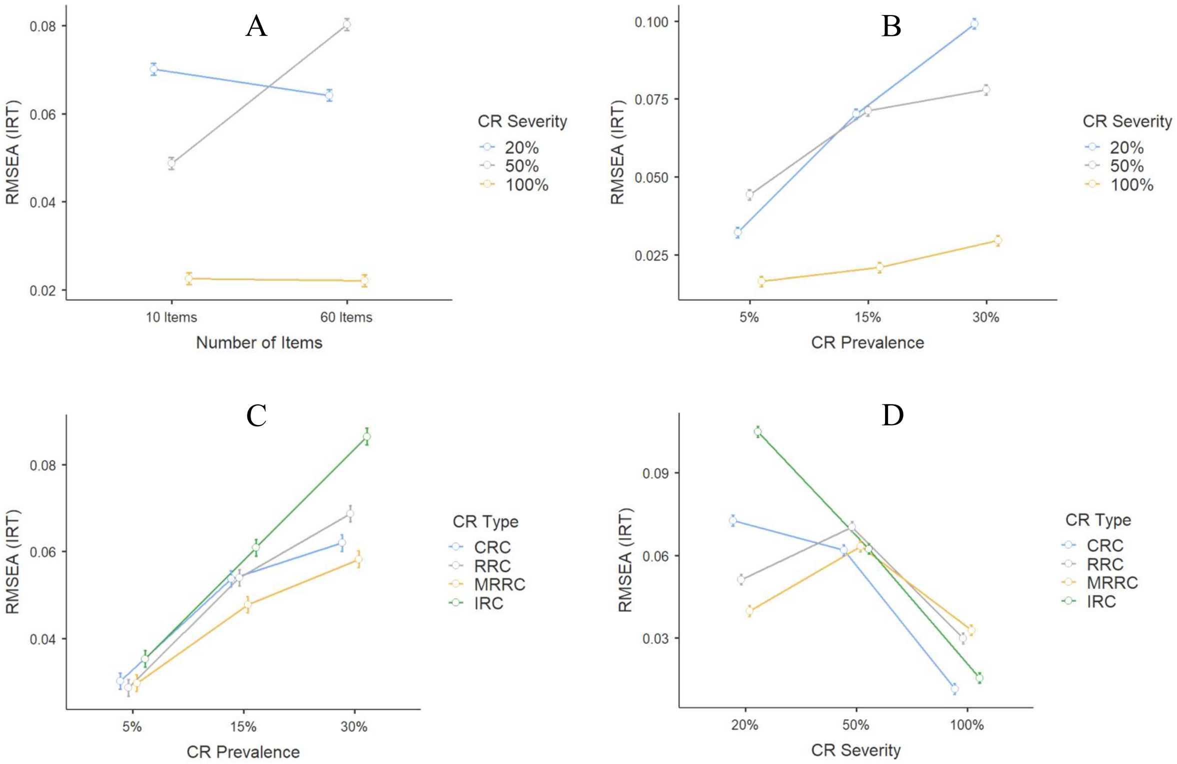


Figure 9. Plots of meaningful interactions for the IRT *RMSEA* index

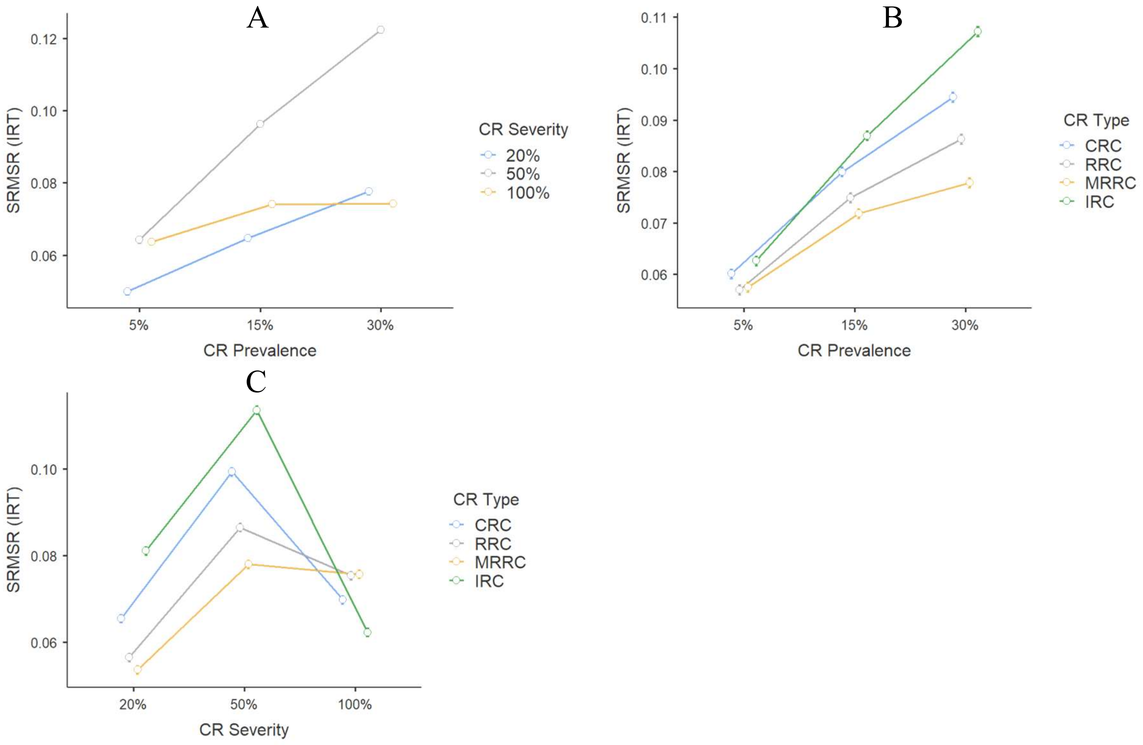


Figure 10. Plots of meaningful interactions for the IRT *SRMSR* index

Appendix H - Plot for Significant CFA Model Fit Interactions

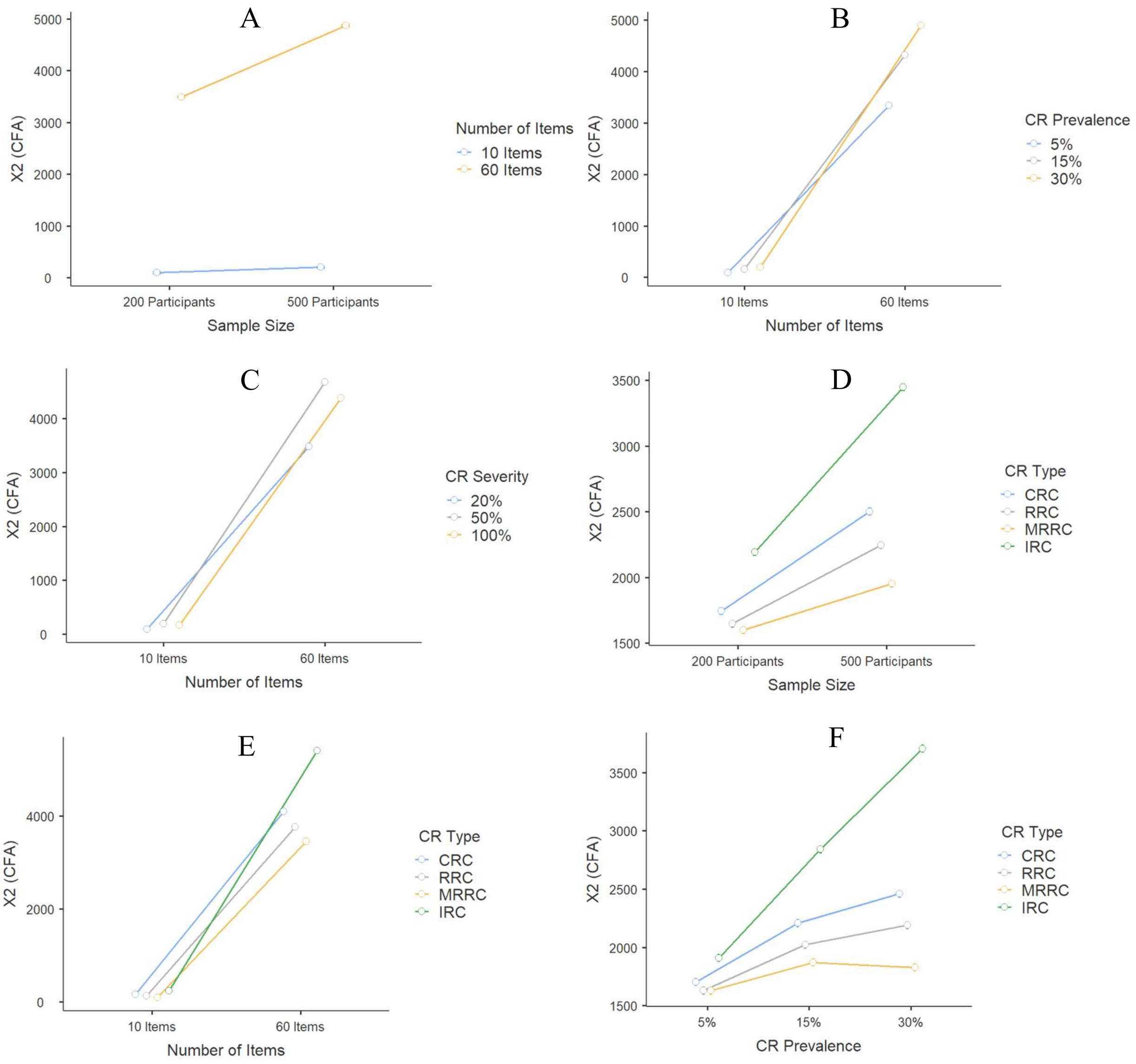


Figure 11. Plots of meaningful interactions for the CFA χ^2 index

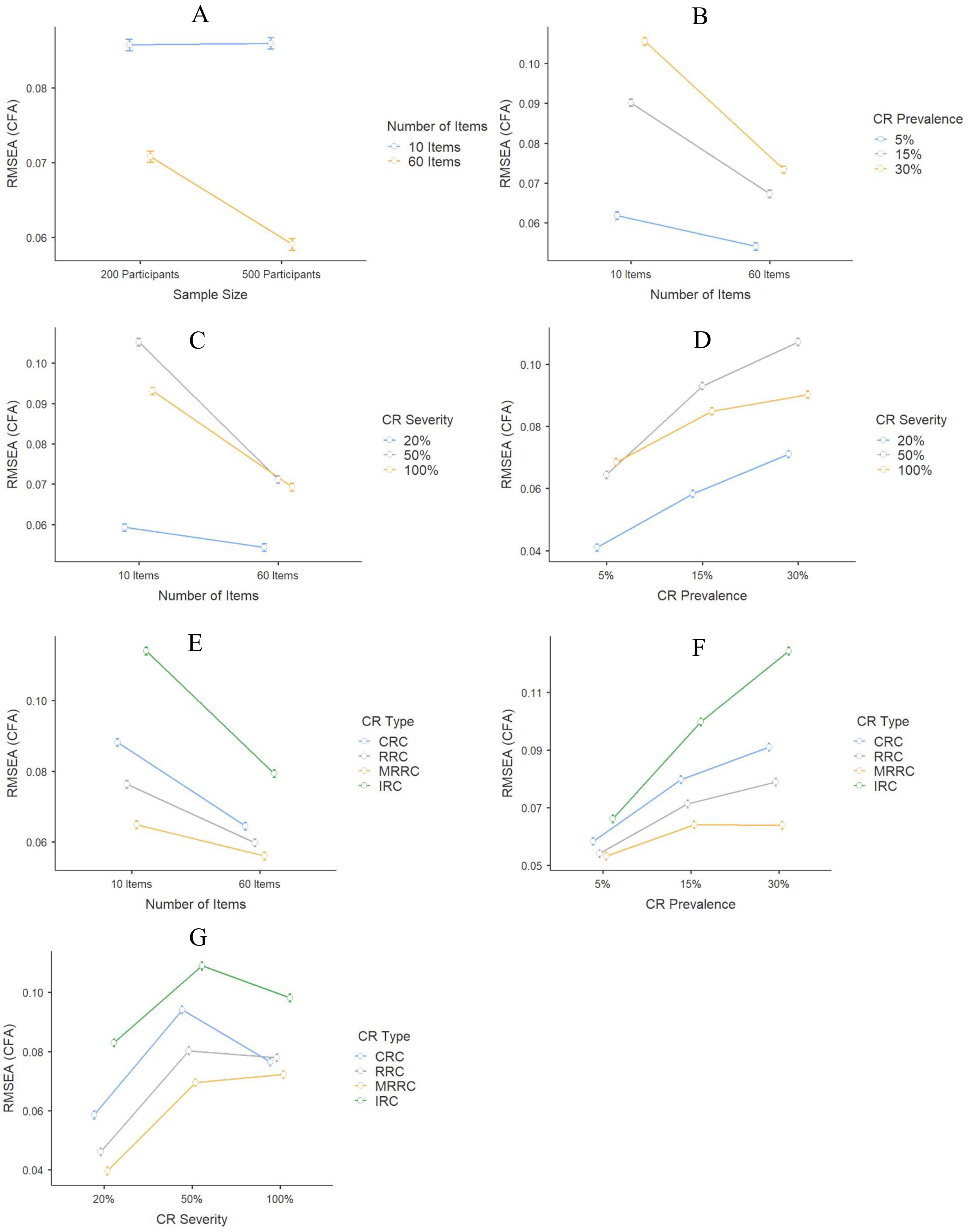


Figure 12. Plots of meaningful interactions for the CFA *RMSEA* index

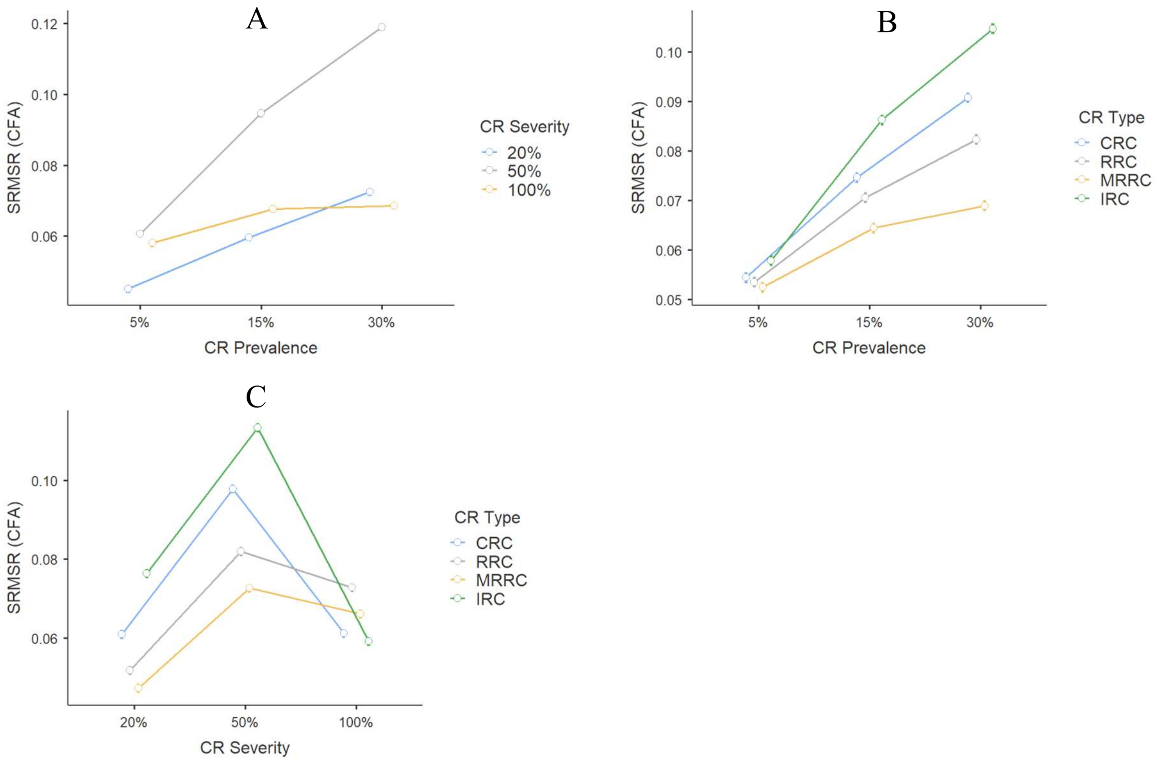


Figure 13. Plots of meaningful interactions for the CFA *SRMSR* index

Appendix I - Model Fit Bias for All Conditions

Table 39. Summary of all model fit bias values (200 participants/10 items)

			IRT Bias			CFA Bias		
Type	Prevalence	Severity	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
CR$_C$	5%	20%	1.13	.01	.00	7.39	.01	.00
CR$_C$	5%	50%	1.12	.02	.02	43.15	.05	.03
CR$_C$	5%	100%	-4.20	.00	.01	26.55	.03	.01
CR$_C$	15%	20%	11.86	.05	.01	27.27	.03	.02
CR$_C$	15%	50%	6.06	.04	.05	115.85	.10	.06
CR$_C$	15%	100%	-5.09	.00	.02	48.53	.06	.01
CR$_C$	30%	20%	30.45	.10	.03	51.76	.06	.03
CR$_C$	30%	50%	4.25	.02	.08	149.29	.12	.08
CR$_C$	30%	100%	-4.89	.00	.02	51.53	.06	.01
RR$_C$	5%	20%	0.58	.01	.00	3.04	.00	.00
RR$_C$	5%	50%	1.55	.02	.01	29.64	.04	.01
RR$_C$	5%	100%	-2.91	.00	.01	30.62	.04	.01
RR$_C$	15%	20%	5.85	.03	.01	14.29	.02	.01
RR$_C$	15%	50%	4.73	.04	.04	74.14	.07	.04
RR$_C$	15%	100%	-3.93	.00	.02	52.23	.06	.02
RR$_C$	30%	20%	13.87	.05	.01	20.27	.03	.02
RR$_C$	30%	50%	6.92	.04	.06	105.61	.09	.06
RR$_C$	30%	100%	-3.35	.01	.02	56.38	.06	.03
MRR$_C$	5%	20%	0.31	.01	.00	2.78	.00	.00
MRR$_C$	5%	50%	0.03	.01	.01	27.42	.03	.01
MRR$_C$	5%	100%	-2.56	.01	.02	37.69	.05	.02
MRR$_C$	15%	20%	3.77	.02	.01	7.49	.01	.01
MRR$_C$	15%	50%	3.24	.03	.03	54.08	.06	.03
MRR$_C$	15%	100%	-2.84	.01	.02	46.16	.06	.02
MRR$_C$	30%	20%	7.79	.04	.01	9.03	.01	.01
MRR$_C$	30%	50%	5.94	.04	.05	70.22	.07	.05
MRR$_C$	30%	100%	-3.11	.01	.02	33.91	.04	.02
IR$_C$	5%	20%	2.52	.02	.01	17.18	.02	.01
IR$_C$	5%	50%	1.12	.02	.02	60.90	.06	.02
IR$_C$	5%	100%	-4.29	.00	.00	27.70	.04	.00
IR$_C$	15%	20%	22.66	.08	.03	70.41	.07	.03
IR$_C$	15%	50%	0.58	.02	.06	144.73	.11	.07
IR$_C$	15%	100%	-4.87	.00	.00	77.53	.08	.01
IR$_C$	30%	20%	65.84	.17	.05	147.27	.12	.05
IR$_C$	30%	50%	-2.16	.00	.09	189.31	.14	.10
IR$_C$	30%	100%	-3.64	.01	.00	141.13	.11	.00
Mean Model Bias			**4.29**	**.03**	**.02**	**57.57**	**.06**	**.03**

Note. CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. Values for each condition represent mean values averaged across 100 replications.

Table 40. Summary of all model fit bias values (200 participants/60 items)

			IRT Bias			CFA Bias		
Type	Prevalence	Severity	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
CR$_C$	5%	20%	509.08	.03	.01	417.97	.01	.01
CR$_C$	5%	50%	960.73	.04	.03	912.29	.02	.03
CR$_C$	5%	100%	-123.05	-.01	.03	1021.37	.03	.02
CR$_C$	15%	20%	1556.99	.06	.03	787.30	.02	.02
CR$_C$	15%	50%	2697.36	.08	.06	1768.70	.04	.06
CR$_C$	15%	100%	-120.57	-.01	.03	1377.62	.03	.03
CR$_C$	30%	20%	2701.15	.08	.04	1111.87	.03	.04
CR$_C$	30%	50%	2087.07	.01	.09	2228.64	.05	.10
CR$_C$	30%	100%	-101.73	.00	.03	1362.02	.03	.03
RR$_C$	5%	20%	231.61	.02	.01	255.05	.00	.00
RR$_C$	5%	50%	711.45	.04	.02	719.03	.02	.02
RR$_C$	5%	100%	125.37	.01	.03	1054.43	.03	.03
RR$_C$	15%	20%	1024.55	.05	.02	575.04	.01	.02
RR$_C$	15%	50%	2775.04	.08	.05	1372.46	.03	.05
RR$_C$	15%	100%	373.11	.03	.04	1487.74	.04	.04
RR$_C$	30%	20%	1441.44	.06	.03	689.13	.02	.02
RR$_C$	30%	50%	4635.68	.11	.07	1841.39	.04	.07
RR$_C$	30%	100%	544.35	.03	.04	1370.01	.03	.04
MRR$_C$	5%	20%	192.72	.02	.01	233.98	.00	.00
MRR$_C$	5%	50%	635.35	.04	.02	716.87	.02	.02
MRR$_C$	5%	100%	418.63	.03	.03	1388.13	.03	.03
MRR$_C$	15%	20%	550.48	.03	.02	354.97	.01	.01
MRR$_C$	15%	50%	2148.03	.07	.04	1145.01	.03	.04
MRR$_C$	15%	100%	610.63	.04	.05	1801.36	.04	.04
MRR$_C$	30%	20%	716.89	.04	.02	370.03	.01	.01
MRR$_C$	30%	50%	3795.23	.10	.06	1232.60	.03	.05
MRR$_C$	30%	100%	659.46	.04	.05	1387.69	.03	.03
IR$_C$	5%	20%	762.61	.04	.02	580.15	.01	.01
IR$_C$	5%	50%	1319.79	.05	.03	1252.82	.03	.03
IR$_C$	5%	100%	-253.68	-.01	.02	1139.46	.03	.01
IR$_C$	15%	20%	3196.78	.09	.04	1489.94	.04	.04
IR$_C$	15%	50%	2419.79	.08	.08	2494.87	.05	.09
IR$_C$	15%	100%	-89.81	.00	.03	2240.60	.05	.02
IR$_C$	30%	20%	6025.55	.13	.07	2535.02	.05	.07
IR$_C$	30%	50%	4565.19	.11	.11	3247.11	.06	.11
IR$_C$	30%	100%	371.87	.03	.02	3675.62	.07	.03
Mean Model Bias			**1390.98**	**.05**	**.04**	**1323.29**	**.03**	**.04**

Note. CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. Values for each condition represent mean values averaged across 100 replications.

Table 41. Summary of all model fit bias values (500 participants/10 items)

			IRT Bias			CFA Bias		
Type	Prevalence	Severity	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
CR$_C$	5%	20%	4.22	.02	.01	20.50	.02	.00
CR$_C$	5%	50%	6.65	.03	.03	94.83	.05	.02
CR$_C$	5%	100%	-2.95	.00	.03	117.01	.06	.02
CR$_C$	15%	20%	32.78	.07	.02	67.86	.04	.02
CR$_C$	15%	50%	23.46	.06	.06	239.51	.10	.05
CR$_C$	15%	100%	-1.99	.01	.04	190.42	.08	.03
CR$_C$	30%	20%	67.72	.11	.03	112.70	.06	.03
CR$_C$	30%	50%	15.02	.04	.09	359.94	.12	.08
CR$_C$	30%	100%	-0.71	.01	.05	227.07	.09	.04
RR$_C$	5%	20%	1.79	.01	.00	7.59	.01	.00
RR$_C$	5%	50%	2.77	.02	.02	48.54	.03	.01
RR$_C$	5%	100%	-2.11	.00	.04	143.09	.07	.03
RR$_C$	15%	20%	22.78	.05	.02	41.28	.03	.01
RR$_C$	15%	50%	18.00	.05	.05	158.90	.08	.04
RR$_C$	15%	100%	0.04	.01	.05	194.35	.08	.04
RR$_C$	30%	20%	31.22	.06	.02	52.36	.03	.01
RR$_C$	30%	50%	36.34	.08	.07	267.87	.10	.06
RR$_C$	30%	100%	9.73	.04	.05	195.41	.08	.05
MRR$_C$	5%	20%	1.53	.01	.01	6.69	.01	.00
MRR$_C$	5%	50%	2.36	.02	.02	46.50	.03	.01
MRR$_C$	5%	100%	-1.41	.01	.03	102.60	.06	.02
MRR$_C$	15%	20%	11.28	.03	.01	18.90	.02	.01
MRR$_C$	15%	50%	13.49	.05	.04	111.62	.06	.03
MRR$_C$	15%	100%	-1.09	.01	.05	130.75	.07	.03
MRR$_C$	30%	20%	16.53	.04	.02	22.34	.02	.01
MRR$_C$	30%	50%	16.63	.05	.05	146.49	.07	.04
MRR$_C$	30%	100%	2.54	.02	.04	90.30	.05	.03
IR$_C$	5%	20%	7.97	.03	.01	39.28	.03	.01
IR$_C$	5%	50%	9.90	.04	.03	143.40	.07	.03
IR$_C$	5%	100%	-3.04	.00	.03	132.40	.06	.02
IR$_C$	15%	20%	64.50	.11	.04	155.41	.07	.03
IR$_C$	15%	50%	17.27	.04	.08	352.63	.12	.08
IR$_C$	15%	100%	-3.34	.00	.03	305.61	.11	.03
IR$_C$	30%	20%	142.68	.18	.06	305.23	.11	.05
IR$_C$	30%	50%	5.08	.01	.11	475.11	.14	.10
IR$_C$	30%	100%	-2.32	.00	.04	509.93	.14	.04
Mean Model Bias			**15.70**	**.04**	**.04**	**156.51**	**.07**	**.03**

Note. CR$_C$ = Respondent-derived careless responding. RR$_C$ = Respondent-derived random responding. MRR$_C$ = Mathematical random responding. IR$_C$ = Invariant responding. Values for each condition represent mean values averaged across 100 replications.

Table 42. Summary of all model fit bias values (500 participants/60 items)

			IRT Bias			CFA Bias		
Type	Prevalence	Severity	*M2*	*RMSEA*	*SRMSR*	χ^2	*RMSEA*	*SRMSR*
CR$_{\text{C}}$	5%	20%	1141.89	.03	.02	889.12	.02	.01
CR$_{\text{C}}$	5%	50%	1879.78	.04	.03	1902.19	.03	.03
CR$_{\text{C}}$	5%	100%	-81.73	-.01	.03	1938.33	.03	.03
CR$_{\text{C}}$	15%	20%	4400.90	.06	.04	2022.97	.03	.04
CR$_{\text{C}}$	15%	50%	5314.16	.07	.07	4199.88	.05	.08
CR$_{\text{C}}$	15%	100%	-26.08	.00	.04	2598.18	.04	.03
CR$_{\text{C}}$	30%	20%	7445.32	.09	.06	2913.35	.04	.05
CR$_{\text{C}}$	30%	50%	4605.09	.07	.10	5157.66	.06	.11
CR$_{\text{C}}$	30%	100%	88.16	.00	.04	2729.92	.04	.03
RR$_{\text{C}}$	5%	20%	700.89	.02	.01	621.42	.01	.01
RR$_{\text{C}}$	5%	50%	1574.76	.04	.03	1560.87	.03	.03
RR$_{\text{C}}$	5%	100%	192.78	.01	.03	2024.02	.03	.03
RR$_{\text{C}}$	15%	20%	2197.10	.04	.02	1192.02	.02	.02
RR$_{\text{C}}$	15%	50%	6187.35	.08	.06	3128.54	.04	.06
RR$_{\text{C}}$	15%	100%	848.40	.02	.05	2925.07	.04	.05
RR$_{\text{C}}$	30%	20%	3624.43	.06	.04	1669.81	.03	.03
RR$_{\text{C}}$	30%	50%	8815.19	.09	.08	4367.28	.05	.08
RR$_{\text{C}}$	30%	100%	1146.61	.03	.05	2579.07	.04	.05
MRR$_{\text{C}}$	5%	20%	495.85	.02	.01	475.97	.01	.01
MRR$_{\text{C}}$	5%	50%	1164.73	.03	.02	1315.37	.02	.02
MRR$_{\text{C}}$	5%	100%	411.79	.01	.03	2144.57	.03	.03
MRR$_{\text{C}}$	15%	20%	1471.73	.03	.02	813.63	.02	.02
MRR$_{\text{C}}$	15%	50%	4567.87	.07	.05	2327.70	.04	.05
MRR$_{\text{C}}$	15%	100%	611.53	.02	.05	2596.03	.04	.04
MRR$_{\text{C}}$	30%	20%	1930.91	.04	.03	915.84	.02	.02
MRR$_{\text{C}}$	30%	50%	8586.89	.09	.06	2704.49	.04	.06
MRR$_{\text{C}}$	30%	100%	700.36	.02	.05	1909.78	.03	.03
IR$_{\text{C}}$	5%	20%	2003.00	.04	.02	1416.73	.03	.02
IR$_{\text{C}}$	5%	50%	2777.38	.05	.04	2749.10	.04	.04
IR$_{\text{C}}$	5%	100%	-191.58	-.01	.03	2293.65	.04	.02
IR$_{\text{C}}$	15%	20%	8500.31	.09	.06	3764.68	.05	.05
IR$_{\text{C}}$	15%	50%	5615.82	.07	.09	5580.36	.06	.10
IR$_{\text{C}}$	15%	100%	4.61	.00	.04	4349.92	.05	.03
IR$_{\text{C}}$	30%	20%	15587.28	.13	.09	6326.06	.07	.08
IR$_{\text{C}}$	30%	50%	8984.32	.10	.12	7127.96	.07	.12
IR$_{\text{C}}$	30%	100%	637.85	.02	.03	6718.27	.07	.04
Mean Model Bias			**3164.32**	**.04**	**.05**	**2776.38**	**.04**	**.04**

Note. CR$_{\text{C}}$ = Respondent-derived careless responding. RR$_{\text{C}}$ = Respondent-derived random responding. MRR$_{\text{C}}$ = Mathematical random responding. IR$_{\text{C}}$ = Invariant responding. Values for each condition represent mean values averaged across 100 replications.

Appendix J - Bias Visualizations for Select Conditions

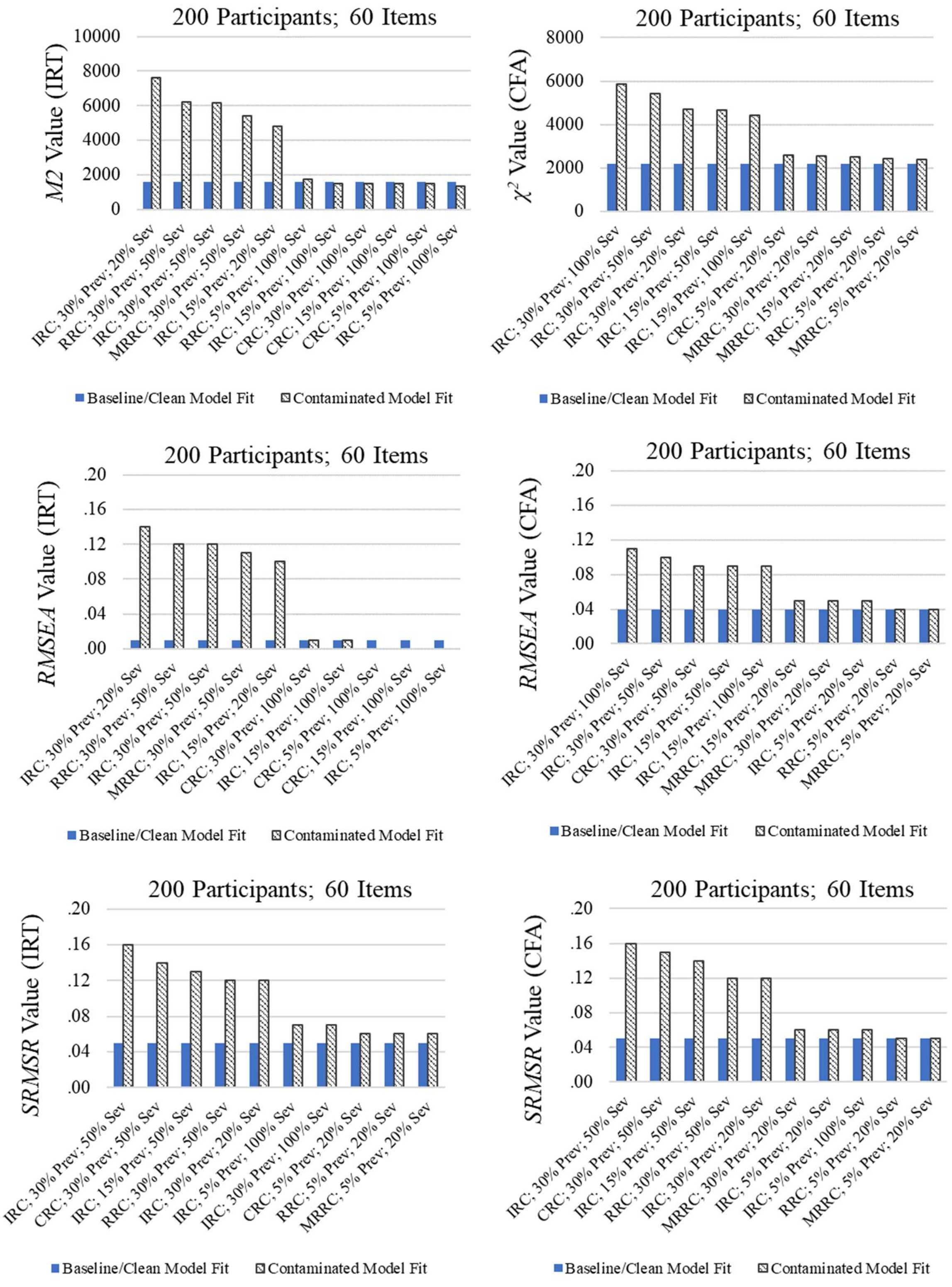


Figure 14. Visual representation of model fit/bias (200 participants/60 items)

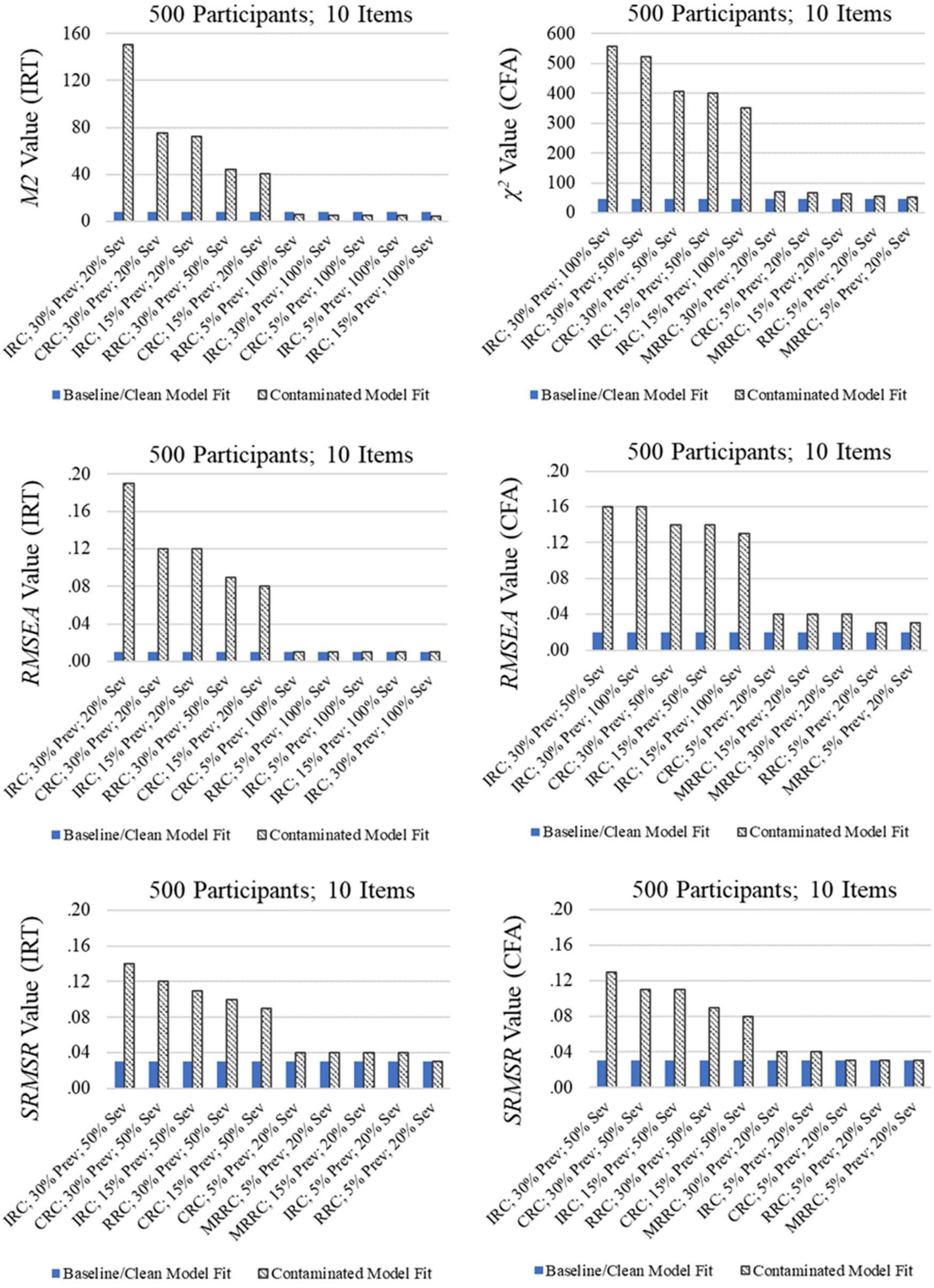


Figure 15. Visual representation of model fit/bias (500 participants/10 items)

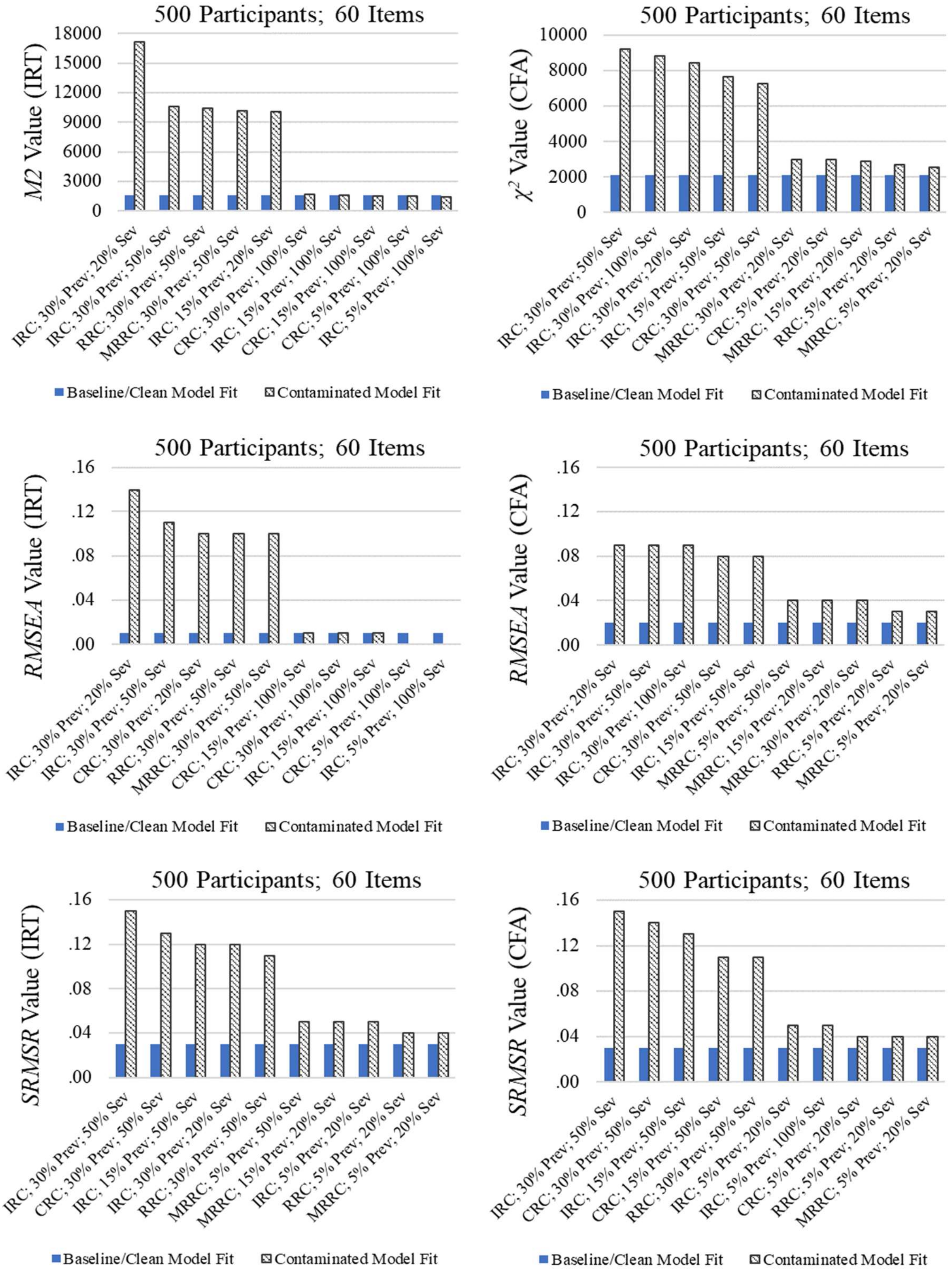


Figure 16. Visual representation of model fit/bias (500 participants/60 items)