5-11-2023

Using IRTrees to account for response style effects between item formats

Stephanie LeRoy
James Madison University

Using IRTrees to Account for Response Style Effects Between Item Formats

Stephanie LeRoy

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2023

FACULTY COMMITTEE:

Committee Chair: Brian Leventhal

Committee Members/ Readers:

Yu Bao

Christine DeMars

Acknowledgements

I am truly grateful to my advisor and committee chair, Dr. Brian C. Leventhal. This thesis would not have been possible without your guidance, patience, and support. Your encouragement has helped me build confidence in myself as a learner and a person. Words cannot express my gratitude towards you.

Thank you to my committee members, Dr. Christine DeMars and Dr. Yu Bao. You were both invaluable resources to me during the thesis process. Thank you for your instruction and support on this thesis.

I am also thankful for my family – my mom, dad, and sisters – who supported me every step of the way. From childhood to now, I am eternally grateful for your role in my life. Thank you, mom, dad, Michelle, and Angela, for everything. Special thanks to my cat, Mango, for his consistency. No matter how hard things were – you still meowed at me for food time. I needed that constant in my life. To Jerry, thank you for always being a rock and supporting me through the best and worst of times. I know I can always count on you.

To the friends I have made along the way – thank you. To my officemates – Mason and Kelsey – thank you for always being a chair swivel away. Your constant support and kind words were always motivating. To Josiah, Nick, and Kate thank you for your friendship and encouragement. I cherish every moment we've spent just hanging out.

To the friends I had before the program – Dylan, Ben, and Sean – thank you for being my friends and constant supports. Regardless of whether we talk once a day, or once a year, I always know I can count on you all.

Lastly, I'd like to thank the people that supported my learning journey and taught me to love learning. Thank you to Mr. Ferrucci, Ms. Sennewald, Ms. Clopton, Mr. Bogdanowicz, Dr. Lloyd, Dr. Gibbons, and to all of my other teachers.

Table of Contents

List of Tables

List of Figures

Abstract

Response styles are consistent person-traits that are defined as the tendency to systematically select responses unrelated to the construct being measured (Paulhus, 1991). Response styles introduce construct-irrelevant variance that distorts observed scores on a measure and biases interpretation of the data. The current study looks at midpoint response style (MRS) and extreme response style (ERS). MRS is the tendency to select the midpoint of a rating scale, while ERS is the tendency to select the endpoints of a rating scale. Previous research sought to either account for response style effects or prevemt them – the current study does both. To account for response style effects, the current study used IRTree models which consists of multiple IRT models layered in a decision tree format. To prevent response style effects, the current study utilized secondary data that implemented two different item formats – traditional Likert items (control) and funnel items (experimental). The MCMC procedure in SAS 9.4 software was used to estimate model parameters. The primary analyses of the IRTree models used the EAP of the differences between the control and experimental group as well as the HPD intervals of the differences. The Likert item condition presented higher difficulty levels for the majority of items for the MRS and ERS stages of the IRTree models. This suggests that funnel items are potentially related to higher cases of midpoint and extreme response selections. In other words, Likert items are potentially related to lower cases of midpoint and extreme response selections. To determine which item format to implement, the costs and benefits for each item format should be assessed.

CHAPTER 1

Introduction

Psychological measures are standardized methods of measuring particular constructs. Constructs are "concepts or characteristics that a test is designed to measure", such as attitudes, knowledge, and personality traits (American Educational Research Association et al., 2014). Psychological measures are commonly used in psychological research to measure individuals or groups, with the intent of measuring validly on the construct of interest. Validity, in this case, represents the degree to which the accumulated evidence supports the interpretation of observed scores for the construct being measured (American Educational Research Association et al., 2014). A major threat to validity is construct-irrelevant variance – the introduction of extraneous variance unrelated to the construct of interest that affect the outcomes of a measure (Downing, 2002). Some forms of construct-irrelevant variance include: testwiseness, poorly constructed items, and response styles (Downing, 2002). The focus of this thesis is on response styles as a source of construct-irrelevant variance that affects the validity of observed scores on measures.

Response Styles

Response styles are consistent person-traits that influence how individuals respond to Likert-type items, unrelated to the construct of interest (Paulhus, 1991). These traits introduce bias to individuals' observed scores on a measure and influence the way the data are interpreted. Although literature (i.e., Baumgartner & Steenkamp, 2001) outlines eight types of response styles, researchers have primarily examined four of the eight response styles: acquiescence response style (ARS; Martin, 1964; Ray, 1983),

disacquiescence response style (DARS; Couch & Keniston, 1960), extreme response style (ERS; Greenleaf, 1992b), and midpoint response style (MRS; Messick, 1968; Schuman et al., 1981). ARS is the tendency to agree with items regardless of the content being measured (Martin, 1964; Ray, 1983). DARS is the tendency to disagree with items regardless of the content being measured (Couch & Keniston, 1960). ERS is the tendency to select the endpoints of a rating scale regardless of content (Baumgartner & Steenkamp, 2001; Greenleaf, 1992b). MRS is the tendency to select the middle option of a rating scale regardless of content (Hurley, 1998; Moors, 2008). Two of the four response styes that are primarily examined in research are of interest in the present thesis: ERS and MRS.

Response styles can bias scores and lead to invalid score interpretations, especially in cross-cultural comparisons (e.g., Buckley, 2009; Chen et al., 1995; Clarke, 2000). Extreme positive, or negative, responses can increase, or decrease, observed scores – respectively (Greenleaf, 2008). For example, Angela is completing an anxiety inventory and comes across a 7-point Likert item where 1 is *strongly disagree* and 7 is *strongly agree*. The item asks for Angela's degree of agreement or disagreement toward the statement, "I am often anxious for no good reason." Angela's true attitude would be reflected by the value 5, *slightly agree*, but Angela's response is 7, *strongly agree*. The positive ERS effect in this example increased Angela's observed score compared to the score of her true attitude. A negative ERS effect, such as an observed score of 1 and a true score of 2, would represent a decrease in observed score.

The effect of midpoint responses is dependent on the mean of the measure relative to the midpoint of the rating scale (Baumgartner & Steenkamp, 2001). For example, the

effect of MRS on a 7-point item with a mean above the midpoint (4) would be negative. The majority of scores representing true attitudes would exist above the midpoint and the effect of MRS would decrease the scores. On the other hand, the effect of MRS on a 7-point item with a mean below the midpoint would be positive. The majority of scores representing true attitudes would exist below the midpoint and the effect of MRS would increase the scores.

Interpretations of these scores biased by response styles would result in inaccurate representations of the construct being measured. However, response styles are measurable and consistent traits of the individual, allowing response styles to be accounted for or prevented (e.g., Weijters et al., 2010).

Accounting for Response Styles

Methods to account for and methods to prevent response styles have been noted in previous literature. Van Vaerenbergh and Thomas (2013) list methods to account for response styles. This list includes classical methods, such as the count procedure and representative indicators, and modern methods, such as those that utilize item response theory (IRT).

The count procedure is accomplished by counting the responses indicative of a particular response style (e.g., Reynolds & Smith, 2010). For example, to measure ERS using the count procedure, extreme responses would be summed across measures. Similarly, to measure MRS, midpoint responses would be summed across measures (Van Vaerenbergh & Thomas, 2013).

The representative indicators for response styles (RIRS) method accounts for response styles by calculating response style scores from an added item set that is

maximally heterogeneous (e.g., Weijters, 2006). The RIRS method differs from the count procedure in that the count procedure models response styles with survey items that are also used for substantive purposes (Van Vaerenbergh & Thomas, 2013). These approaches are limited in that they are indifferent to the trait level of the respondent and unconcerned with the psychometric properties of the items (Bolt & Johnson, 2009).

IRT accounts for response styles by modeling the probability of selecting a particular response as a function of the underlying response style trait (e.g., Bolt & Newton, 2011). Using IRT models allows for each item to be differentially useful in measuring ERS, but has not been developed for other response styles (Van Vaerenbergh & Thomas, 2013). IRT models have properties of invariance that imply items display the same parameter estimates across groups (Asún et al., 2017).

Additional factors can be introduced into an IRT model that differentiate respondents on the same rating scale. These factors can be other traits, such as additional response styles, that influence the selection of scale responses by respondents (Bolt & Johnson, 2009). This multidimensional approach to IRT modeling for response styles is referred to as multidimensional IRT (MIRT). MIRT approaches are commonly used to directly estimate response style traits and allows for the estimation of multiple traits, such as traits of interest and response style traits (e.g., Bolt & Newton, 2011; Falk & Ju, 2020).

One MIRT model for response styles is the IRTree (e.g., Böckenholt, 2012; Böckenholt & Meiser, 2017; Leventhal, 2019; Spratto et al., 2021). IRTree models are comprised of multiple IRT models layered in a decision tree format that allow for simultaneous detection and correction of bias in observed scores caused by response style

effects (Ames & Leventhal, 2021). The MIRT model used in the current study is the IRTree that assumes a multi-stage response process to account for ERS and MRS effects.

Prevention of Response Styles

Methods to prevent response styles take preemptive measures, such as creating balanced measures, adjusting rating scale length, and altering item formatting (Van Vaerenbergh & Thomas, 2013). Some methods for preventing response style effects are limited in that they only prevent certain response style effects. One example of this is the balanced measure method in which a mixture of positively worded and negatively worded items is included on the measure (e.g., Billiet & McClendon, 2000). This balance between items is meant to prevent ARS and DARS but does not provide the same benefit to other response styles.

Another method to prevent response style effects is to alter the length of the rating scale (e.g., Kieruj & Moors, 2010; Weijters et al., 2010). Rather than using an odd-numbered rating scale, an even-numbered rating scale can be used to avoid providing a midpoint response option. Removing the choice of a midpoint response has been used as a method of mitigating MRS effects (Kieruj & Moors, 2010). To prevent ERS effects, the rating scale can be made longer as ERS effects have been found to lessen as the number of response categories increase (Weijters et al., 2010).

Another preemptive measure to prevent response style effects changes how the item is presented to the respondents. The most common item format used in psychological measures is the traditional Likert-type item with a range of responses. Item formats that have been utilized to prevent response style effects include drag-and-drop

items, where item stems are dragged and dropped into the response category selected, and funnel-formatted items, where the items are broken into sub-items (Böckenholt, 2017).

The funnel item format assumes the form of a hypothesized response process that respondents use to respond to an item. The hypothesized response process has three stages where decisions are made by the respondent. Each stage is represented by an individual sub-item and sequentially presented to the respondent. For example, Angela was previously shown an item with the statement "I am often anxious for no good reason," in conjunction with a 7-point rating scale. If the item is presented in a three-stage funnel item format, Angela would see sub-items rather than the 7-point rating scale. First, Angela would be shown the statement and asked if she has an opinion, to which she can answer "yes" or "no." If "no" then she moves on from the item. If she selects "yes" then she is moved to the next sub-item. The next sub-item asks whether she agrees or disagrees with the statement, to which she can answer "agree" or "disagree." She is then shown the next sub-item that asks for the degree to which she agrees (or disagrees) with the statement. This study utilizes different item formats to prevent response style effects. Item formats used in the current study include the traditional Likert item format, using a 5-point scale and a 7-point scale, and the funnel item format (Mellenbergh, 2011).

Current Study

In the current study, I utilized IRTrees to model ERS and MRS effects present in traditional Likert items and funnel-formatted items. The research questions that I aimed to answer through the current study were:

(1) Does one item format reduce midpoint response selections?

Specifically, if the IRTree model exhibits a higher difficulty for the MRS trait response then individuals are less likely to select the midpoint response option. If the IRTree model exhibits a lower difficulty for the MRS trait response then individuals are more likely to select the midpoint response option.

(2) Does one item format reduce extreme response selections?

Specifically, if the IRTree model exhibits a higher threshold for the ERS trait response then individuals are less likely to answer extremely. If the IRTree model exhibits a lower threshold for the ERS trait response then individuals are more likely to answer extremely. This relationship between threshold and endorsed response is similar across all traits.

CHAPTER 2

Literature Review

Response Styles

Measures in social science research have commonly used rating scales as a means to collect information on specific constructs of interest. Individuals' responses on these rating scales are assumed to provide an accurate representation of the construct being measured. This assumption that responses accurately represent the construct of interest is vital to properly interpreting observed data. However, external factors can, and often do, influence individuals' responses on these rating scales. This external variance, unrelated to the construct of interest, is construct-irrelevant variance. The rating scale associated with the items can lead to biased responses with one of the potential factors that induce biased responses in an individual being response styles (Baumgartner & Steenkamp, 2001; Cronbach, 1946).

Response styles are one potential source of construct-irrelevant variance and are the tendency to systematically select responses unrelated to the construct being measured (Paulhus, 1991). There are two types of response styles examined in the current study – ERS and MRS. ERS is the tendency to select the endpoints of a rating scale regardless of content (Baumgartner & Steenkamp, 2001; Greenleaf, 1992a). MRS is the tendency to select the middle option of a rating scale regardless of content (Hurley, 1998; Moors, 2008). Response styles can bias the observed responses on a rating scale regardless of the construct measured because they are content independent – their effect on scores does not rely on the content of the measure (Smith, 2017).

Effects of Response Styles

Response styles are a consistent tendency of an individual across administrations (e.g., Weijters et al., 2010). As a consistent person-trait, this tendency is quantifiable making it possible to take steps to understand, measure, and prevent response style effects. The presence of response styles and other sources of error imply that observed scores are not representative of just the construct being measured. Rather, observed scores are a composite of individuals' true level on the construct and construct-irrelevant variance, some of which can be attributed to response styles.

Response styles pose a potential threat to the validity of score interpretations by biasing the observed responses (Baumgartner & Steenkamp, 2001). Respondents with high ERS tendencies tend to have higher, or lower, scores than respondents with low ERS that tend to have more moderate scores (Greenleaf, 2008). Specifically, extreme positive responses would increase observed scores for positively worded items and decrease observed scores for negatively worded items. Alternatively, extreme negative responses would decrease observed scores for positively worded items and increase observed scores for negatively worded items. However, it is unlikely for respondents to completely ignore scale content meaning that ERS should bias scores in the direction of the mean of the scale, relative to the midpoint of the scale (Baumgartner & Steenkamp, 2001). If the mean of the scale is below the midpoint, ERS would make observed scores more negative. On the other hand, if the mean of the scale is above the midpoint, ERS would make observed scores more positive.

Respondents with high MRS tendencies have scores that lean towards the midpoint of the scale. The direction and magnitude of the effect of MRS on observed

scores is hypothesized to depend on the deviation of the scale's mean (average score across respondents on the scale) from the midpoint of the response scale (i.e., 3 on a 1-5 scale, 4 on a 1-7 scale; Baumgartner & Steenkamp, 2001). Consider this in the context of MRS for a 5-point response scale. If the mean of the scale were equal to 3, the midpoint of the scale, then MRS would not systematically influence the mean of scale scores but would reduce the variance of the scores. For respondents with true scores below the midpoint the bias would be positive and for respondents with true scores above the midpoint the bias would be negative. If the mean were greater than the midpoint, most true scores would be above the midpoint and MRS should decrease scores, on average. If the mean were lower than the midpoint, most true scores would be below the midpoint and MRS should increase scores, on average (Baumgartner & Steenkamp, 2001). The greater the mean of the scale deviates from the midpoint, the greater the biasing effect MRS has on observed scores.

In general, response styles pose a potential threat to the validity of score interpretations by biasing the observed responses (Baumgartner & Steenkamp, 2001). ERS is considered a potential factor in observed score differences between groups. ERS is of concern when making cross-cultural comparisons as it is related to "demographic, personality, cultural, and national variables" (Greenleaf, 2008). For example, Chen et al. (1995) found that US respondents were more likely to endorse extreme responses than Japanese, Taiwanese, and Canadian respondents. The US respondents' mean observed scores were inflated for positively worded items and deflated for negatively worded items in comparison to the other three groups. Clarke (2001) also found that culturally distinct groups exhibit varying levels of ERS tendencies. This study conducted a post hoc

statistical adjustment to minimize the bias of ERS and found that the bias of ERS alters statistical analysis in cross-cultural marketing research. Buckley (2009) used a set of ad hoc methods and a Bayesian hierarchical approach on the student questionnaire from PISA 2006 and found cross-cultural response style variation. Buckley (2009) suggests investigating potential changes in item design to mitigate issues of variation in measurement due to response styles. These biasing effects of response styles on observed scores can lead to unreliable interpretations of data.

Past research on response styles has provided methods to both account for response styles and proactively mitigate the effects of response styles on observed scores from psychological measures (e.g., Chen, Lee, & Stevenson, 1995; Stening & Everett, 1984). Each method presents their own advantages and disadvantages (Van Vaerenbergh & Thomas, 2013). Methods to account for response styles—classic and modern—will be summarized and described and followed by methods to prevent response style effects from biasing score interpretations.

Methods to Account for Response Styles

Classic Methods

Early methods to account for response styles include observed and latent variable approaches and provide the benefit of ease of use.

Count Procedure. Perhaps the simplest method to account for response styles is to count the frequency of response style indicators (Van Vaerenbergh & Thomas, 2013). If using the count procedure, ERS would be measured by counting the number of responses that represent the endpoints of the scale and MRS would be measured by counting the number of responses that represent the midpoint of the scale. Though the

count procedure is easy to utilize it still presents its own limitations. This procedure requires that the scale items measure the same construct and is not an effective method for differentiating response style effects and traits of interest scores. Reynolds and Smith (2010) quantified the presence of ERS and MRS by calculating the percentage of endpoint responses and middle category responses, respectively. By calculating percentages as values representative of ERS and MRS, Reynolds and Smith (2010) present one potential use of the count procedure.

Representative Indicators. The representative indicators for response styles (RIRS) method requires items that are maximally heterogeneous (i.e., as unrelated to the content of the measure, and each other, as possible) to be added to the survey. The total number of extreme responses to these items is used as an ERS indicator. This way of quantifying response style effects assumes that consistent response patterns, regardless of content, indicates response styles (Baumgartner & Steenkamp, 2001; Weijters, 2006). As such, these additional items are used to indicate the weight of response style tendencies (e.g., Greenleaf, 1992a). Adding RIRS is fairly comprehensive as it can be used to account for ERS, MRS, ARS, and DARS. This method is calculated similarly to the count procedure but requires additional items, unrelated to the construct of interest, to lengthen the survey.

RIRS Means and Covariance Structures. The representative indicators for response styles means and covariance structures (RIRSMAC) method is an extension of the RIRS method where the additional items serve as observed variables in a confirmatory factor analysis with response styles serving as latent variables. This method is also easily calculated and enables the use of response styles as covariates in subsequent

analyses. This method, like RIRS, requires additional items, unrelated to the construct of interest, that lengthen the survey. The RIRSMAC method requires intervention during the creation of the measure to include representative indicators of response styles making it unusable in the case of second-hand data (Van Vaerenbergh & Thomas, 2013; Weijters et al., 2008).

Modern Methods

Modern methods use multidimensional item response theory (MIRT), such as IRTrees, to measure and allow for adjustments of latent traits based on response styles traits. These methods—MIRT and IRTrees—do not require additional items to be added to a scale to model response style traits. The base of these models, item response theory (IRT), refers to a system of models that show the relationship between a respondent's trait, or ability, symbolized by theta (θ), and an item response. Responses to these items can be dichotomous (two categories) or polytomous (more than two categories). A primary function of IRT is to establish individuals' positions on an unobservable continuum under the assumptions that the unobserved trait of interest (TOI) and items on a measure are organized in that continuum (Embretson & Reise, 2000). In IRT, latent traits are considered to be unobservable characteristics, attributes, or constructs of interest. Item parameters, in addition to the latent trait, determine the value of the probability of a particular response.

To model item responses, IRT offers models for dichotomous items (e.g., 2-parameter logistic model; 2PL) as well as for polytomous items such as Samejima's (1969) Graded Response Model (GRM). To use these models, three statistical

assumptions must be met: unidimensionality, local independence, and correct model specification.

The assumption of unidimensionality means that each individual respondent has only one unobserved TOI with all other influential factors assumed to be random error. That is to say, observed scores are assumed to be a function of a continuous latent trait. Individuals can be located and compared in a unidimensional latent space. The assumption of local independence states that conditional on item and individual parameters, responses should be independent (DeMars & Jacovidis, 2016). Correct model specification, or functional form, assumes that the collected data fits the function specified by the model (De Ayala, 2009).

Dichotomous IRT models pertain to items with two categories, meaning responses can be coded as 0 or 1. The category that represents a higher level of the construct is scored 1. For example, on a dichotomous item measuring depression, the response indicating higher levels of depression would be scored 1. Dichotomous models present the probability of a score of 1. Inversely, the probability of a score of 0 is equal to one minus the probability of a score of 1. For a cognitive item, the response scored 1 would refer to the correct answer and the response scored 0 would be the incorrect answer. In an attitudinal item, such as the item measuring depression example, the response scored 1 would be the response that endorses higher levels of the construct and the response scored 0 would be the response that endorses lower levels of the construct. For simplicity's sake, the response scored 1 will be referred to as "endorsement" from here on. The probability of endorsement is expressed as a function of θ (DeMars, 2010).

This suggests that probabilities for a θ level can be interpreted as the probability of endorsement for any given examinee selected from a group with that same θ level.

The 2PL is a common model for dichotomous items and is named after the number of item parameters used in the function modeling the relationship between θ and the response (1 or 0; DeMars, 2010). The two item parameters in the 2PL model are item difficulty and item discrimination. Item difficulty, denoted as *b*, indicates the level of θ needed to be more likely than not to endorse a trait response. In other words, when a respondent's θ is equal to *b*, they have a 50% chance of scoring 1. The theoretical range for item difficulty is -∞ to +∞ but the plausible range is from -3 to +3 (DeMars, 2010). Item discrimination, denoted by *a*, indicates how well an item can differentiate individuals across θ levels. An item with higher discrimination would differentiate individuals across different θ levels and is more desirable than an item with lower discrimination. The theoretical range for item discrimination is -∞ to +∞ but the plausible range is from 0 to +4 (DeMars, 2010). The probability of endorsement using a 2PL model is given by

$$P_{ij}(\theta_j) = \frac{e^{a_i(\theta_j - b_i)}}{1+e^{a_i(\theta_j - b_i)}} \quad (1)$$

where *j* indexes individuals and *i* indexes the items. Item difficulty, for a given item *i*, is represented by b_i and the item discrimination, for a given item *i*, is represented by a_i. The latent ability of an individual is represented by θ_j.

Polytomous IRT models pertain to items with more than two categories. More commonly, these item response categories are expected to be ordered. For example, a Likert-item with a scale ranging from 1 (strongly disagree) to 5 (strongly agree) has a clear order. One such model is the GRM. The GRM models the probability of scoring in

or above a given category, or a cumulative probability. For example, consider an item scored from 0 to 2, with 0 indicating lower levels of the construct of interest and 2 indicating higher levels of the construct of interest. The probability of scoring 0 or higher (i.e., 0, 1, and 2) is 1. The probability of scoring above 2 (outside of the possible range) is 0.

Mathematically, the function for the GRM and 2PL look similar – however, the GRM function has multiple *b* parameters representative of category *k* thresholds. These *b* parameters differ from the 2PL as they represent the category boundary, or threshold, for a category, *k*, of item, *i*. This threshold represents the boundary at which examinees have a 50% chance of selecting a category lower than *k* or selecting a category *k* or higher. The cumulative probability, $P^*_{ijk}(\theta_j)$, of endorsing a particular response category, and above, as a function of latent ability using a GRM is given by

$$P^*_{ijk}(\theta_j) = \frac{e^{a_i(\theta_j - b_{ik})}}{1+e^{a_i(\theta_j - b_{ik})}} \quad (2)$$

where the asterisk denotes the cumulative aspect, *j* indexes individuals, *i* indexes the items, and *k* represents the response category. The category threshold parameter, for the k^{th} category of a given item *i*, is represented by b_{ik}, and the category discrimination, for a given item *i*, is represented by a_i. The total number of category threshold parameters is equal to $K - 1$. The latent ability of an individual is represented by θ_j.

Calculating the probability of a particular category requires the cumulative probability of selecting category *k* or above and the cumulative probability of selecting category *k+1* or above. The individual probability of an individual *j* endorsing one particular category on a given item *i* is represented by

$$P_{ijk}(\theta) = P^{*}_{ijk} - P^{*}_{ij,k+1} \quad (3)$$

Continuing the previous example of an item scored 0 to 2 – to calculate the probability of scoring 1 would be

$$P_{ij,k=1}(\theta) = P^{*}_{ij,k=1} - P^{*}_{ij,k=2} \quad (4)$$

These two IRT models, the 2PL and the GRM, model the probability correctly only when the assumption of unidimensionality is met. When this assumption is violated, such as when response styles affect endorsed responses, different methods are required to model the multidimensionality.

Multidimensional Item Response Theory. MIRT is useful when the assumption of unidimensionality in IRT is violated. When more than one trait influences an item's response, such as response style traits, MIRT can be used as the multidimensional extension of unidimensional IRT models (Leventhal & Stone, 2018). One MIRT model that has been used with response styles is the IRTree – which utilizes multiple IRT models in a decision tree format.

IRTrees. The application of IRTree models to account for response styles has become more prevalent in recent literature (e.g., Böckenholt, 2012; Böckenholt & Meiser, 2017; Leventhal, 2019; Spratto et al., 2021). This may be due to the flexibility of IRTree models to simultaneously detect and correct response style bias in observed scores (Ames & Leventhal, 2021). IRTree models can model ERS and MRS by assuming individuals respond to Likert items using a hypothesized multi-stage response process.

Each IRTree model consists of nodes and branches. The nodes are representative of separate θs. At each node a decision is made, for the respective θ, that determines which branch is followed, leading to the next node. Node-level decisions are modeled

using unidimensional IRT models. The branches represent the probability of the decision made. The probability of a particular observed response is the product of the branch probabilities associated with the path traversed by respondents (e.g., Spratto et al., 2021).

Branching probabilities in IRTree models are dependent on whether the nodes are dichotomous or polytomous. Dichotomous nodes have two branches while polytomous nodes have three or more branches a Figure 1 and Figure 2). The hypothesized multi-stage response process of interest to this study consists of three nodes modeling the θs, or traits – MRS, TOI, and ERS. The first node characterizes the θ representing MRS tendencies ($\theta_{j,MRS}$). This node corresponds with a dichotomous decision thus is modeled using the 2PL. The second node characterizes the θ representing the TOI of the measure ($\theta_{j,TOI}$). This node also corresponds with a dichotomous decision, and thus is modeled using the 2PL. The third, and final, node characterizes the θ representing ERS tendencies ($\theta_{j,ERS}$). In a 5-point scale this node corresponds with a dichotomous decision and is modeled using the 2PL. However, in a 7-point scale this node corresponds with a trichotomous decision and is modeled using the GRM (Spratto et al., 2021).

Although the 2PL and the GRM models are used at the node-level, item parameters take on modified interpretations. In the θ_{ERS} and θ_{MRS} nodes, a higher item difficulty would suggest that higher levels of the respective response style tendencies are required to endorse the particular responses that the traits represent (e.g., midpoint response for MRS node). Lower item difficulty would suggest that the individual would not require high levels of response style tendencies to endorse the item.

IRT provides many models that are helpful in quantifying the presence of response styles. Though detection, and potentially correction, of response style effects is

useful in providing less biased interpretations of scores, research has also looked toward a prevention approach.

Methods to Prevent Response Styles

To prevent response style effects, researchers must be aware of the potential sources of response styles. One source of response styles is the individual, as response style tendencies are a consistent respondent characteristic (Bachman & O'Malley, 1984). Kieruj and Moors (2010) make the argument that response styles are unpredictable across individuals, making prevention of response styles almost impossible. This is why accounting for response styles through analysis is essential. However, another source of response styles is external stimuli such as the scale format and item format.

Classic Methods

One method to prevent response styles during instrument development is to create a balanced measure with positively and negatively keyed item stems (Paulhus, 1991). There are methods that also aim to alter the rating scale rather than the item stem. One method to prevent ERS altered the rating scale by reducing the scale options to two – eliminating the problem of extreme responses. The downside to this method is that it provides less information than longer scales (Paulhus, 1991). Other methods to prevent ERS included implementation of longer scales, fully labelled scales, and the inclusion of a neutral option which resulted in lower levels of ERS (Weijters et al., 2010). Specifically, research on scale labeling suggests that endpoints are more frequently selected in scales with only the endpoints labelled (Spratto et al., 2021).

One potential method for preventing MRS, shortening the rating scale, contrasts with lengthening the rating scale for ERS prevention (Kieruj & Moors, 2010). Another

method used to prevent MRS is to remove the midpoint, however research has found that on even-numbered rating scales respondents with higher MRS tendencies selected responses located around the midpoint (Kieruj & Moors, 2010).

Item formats were also investigated for their use in preventing response style effects. Albaum et al. (2007) studied the effect of scale formatting on the proportion of extreme responses using a one- and two-stage item format. A one-stage item format was presented as a traditionally formatted Likert item while the two-stage item format was presented as two questions. The first question of the two-stage item format asked the general attitude of the individual (i.e., agree, disagree, or neutral) to the statement given. The second question asked for the degree to which the respondent agreed or disagreed with the statement given. The results suggested that the traditional Likert format items resulted in less extreme response.

Modern Methods

Böckenholt (2017) conducted a similar study to Albaum et al. (2007) where each item was presented as multiple items. The items were modeled after a hypothesized sequential response process parsed out into funnel items. IRTree models were then used to provide a framework to separate response style effects and true scores on the measure. By using multiple item formats, Böckenholt (2017) demonstrated format-dependent systematic response styles, specifically ERS, that can be measured using IRTree models. These IRTree models were found to fit the Likert and funnel formatted items better than a single stage GRM.

The Current Study

The biasing effects of response styles on the interpretability of research conclusions poses a problem for data in which response styles were not accounted for. Not taking systematic error, such as response styles, into account can lead to spurious relationships between variables and affect the impact of the research. Some methods have been put in place to detect and correct for response styles after data collection but rely heavily on predetermined conditions of the rating scales and items used. Proactively preventing response style effects in data sets is an ongoing conversation in recent literature. Böckenholt (2017) and Album et al. (2007) published promising results suggesting that item formatting may be a preemptive method to account for ERS. Currently, response style research does not provide a comprehensive method to proactively account for multiple response styles such as ERS and MRS. The current study aims to extend previous research on item formats and examine the influence of the combination of a three-stage IRTree and funnel item formatting on extreme and midpoint response styles.

CHAPTER 3
Method

The goal for this study is to evaluate the effect of item format on response style tendencies in respondents. Specifically extreme and midpoint response styles are evaluated in this study. Item formats differ between traditionally formatted Likert items and funnel-formatted items. The different item formats are implemented using a 5-point and a 7-point response scale. Observed scores are used to model response style tendencies, along with the TOI, using two IRTree models. The 5-point response scale items are modeled using a fully dichotomous IRTree model and the 7-point response scale items are modeled using a partially polytomous IRTree model. Bayesian analysis is utilized to incorporate previous literature using priors and approximating a posterior distribution of the data. After approximating the posterior distributions, model-data fit is assessed. If the model fits the data, then item parameter values are compared to answer the research questions.

The data used for this study is secondary – previously collected and repurposed for the current study. As such, artifacts of the previous study remain. For example, having two different response scales (i.e., 5-point and 7-point scale) is not essential but is included in this secondary data analysis.

Participants

The current study utilized two separate samples. For both samples, participants were undergraduate students at a mid-sized public university in the mid-Atlantic region of the United States. The control sample consisted of 3,671 incoming first-year students who completed the subscales during a mandatory university-wide assessment day during

the Fall semester of 2021. The experimental group consisted of 706 students from the university's research participant pool in the 2020-2021 academic year. These samples were purposefully selected. The incoming first-year students in the control sample would not overlap with the experimental group who already completed the assessments. These two groups were also given the same assessments and completed these assessments in a remote testing environment.

Data Collection

Data were collected from the control sample during a university-wide assessment day. On assessment day, all incoming first-year students were required to complete a battery of low-stakes assessments. Each student was randomly assigned a configuration of tests that take approximately two hours to complete. Students were not penalized for poor performance on any of the tests and students who did not complete the assessments by the deadline had holds placed on their accounts that could be removed once their required assessments were completed. The purpose of this university-wide assessment day is to facilitate pre-post data collection on student learning outcomes and developmental outcomes. From the group of incoming first-year students who completed the mandatory assessments, 3,671 students completed the assessments of interest for the current study.

Participants in the experimental group signed up for the study based on a brief description of the assessments. According to the guidelines of the research pool, only students 18 years of age and older were allowed to sign up for the study. If the student signed up to take part in the study, they were sent an email with instructions on how to participate. The instructions contained a link to a consent form and the survey.

Both samples completed the assessments online and were allowed to complete the surveys at the location of their choosing with the minimum condition being internet connection availability. Any device with access to a web browser could be used to complete the survey. Completion of the assessments was mandatory for students in the control group and voluntary for the experimental group – students were allowed to withdraw from participation at any point in the survey. Participation to be a part of research studies was voluntary for both the control and experimental group. All responses were anonymous with no identifying information linking individuals and their responses.

Measures

Participants in the control sample were given a series of surveys online with 11 total items pertaining to the current study. Items consisted of six, seven-point Likert items measuring Intellectual Overconfidence (IO) from the Intellectual Humility (IH) scale of the Critical Thinking Disposition (CTD; Sosu, 2013) assessment and five, five-point Likert items measuring Confidence in Communication (CC) from the Attitudes Towards Communication (ATC) subscale of the Test of Oral Communication Skills, Version 2 (TOCS-2; Williams, Horst, & Sundre, 2014).

Intellectual Overconfidence is the reverse of one of four distinct, yet intercorrelated, aspects of intellectual humility – lack of intellectual overconfidence. The scale has to do with the conceptualization of intellectual humility as an intrapersonal and interpersonal construct – defined as a benign awareness of one's intellectual imperfection. Outcomes associated with intellectual humility include open-mindedness

and tolerance for others (Krumrei-Mancuso & Rouse, 2016). As such, higher levels of intellectual overconfidence are considered to reflect less intellectual humility.

Confidence in Communication items were developed for the purpose of assessing affective components of communication. More specifically, items on this subscale addressed individuals' self-efficacy in communication. Overall, there is little to no conceptual overlap between these two subscales. For the purpose of this study, it is critical that different constructs are investigated as ERS and MRS traits should not be dependent on the construct being measured.

Item Format

The participants in the control group received the measures with items in a traditional Likert format. Response options ranged from "strongly disagree" to "strongly agree" on a 7- and 5-point scale, for the IO and CC subscales respectively (see example in Figure 3 and Figure 4).

The participants in the experimental group received the measures with items in a funnel format (see example in Figure 5). Each original item was separated into three sub-items. The first sub-item asked, "Do you have an opinion toward the following statement?" followed by the statement in the original item. The participants were able to select that they "have an opinion" or "do not have an opinion". If the participant responded that they did not have an opinion, they were moved on to the next set of sub-items. If the participant responded that they did have an opinion, they were moved on to the next sub-item.

The second sub-item asked, "Given that you have an opinion, do you agree or disagree with the statement," followed by the statement in the original item. The

participants were able to select that they “agree” or “disagree” with the given statement. Depending on the direction of their opinion, they were given one of two sub-items.

If the participant agreed with the given statement, the third sub-item asked, “Given that you agree, how strong is your opinion?” followed by the statement in the original item. For the seven-point scale items, the participants were able to select that the strength of their opinion was “strong,” a blank indicating moderate, and “slight” – with “slight” being removed for the five-point scale items. If the participant disagreed with the given statement, the third sub-item asked, “Given that you disagree, how strong is your opinion?” followed by the statement in the original item. For the seven-point scale items, the participants were able to select that the strength of their opinion was “strong,” a blank indicating moderate, and “slight” – with “slight” being removed for the five-point scale items.

Data Analysis

Two separate datasets were analyzed – observed scores for Likert items and observed scores for funnel items. The data in both datasets were analyzed using dichotomous and polytomous IRTree models under a Bayesian framework. IRTree models for the two datasets used the same prior distributions and initial values. This section describes the steps taken to set prior distributions, approximate the posterior distribution, and analyze model-data fit.

Dichotomous IRTree Model

In the three-stage dichotomous IRTree models, stage dependent decisions were modeled using a unidimensional 2PL model (see example in Figure 1). Each stage corresponded with one of three traits of interest: MRS, TOI, and ERS. As such, the

probability of an individual, *j*, endorsing a response on a given stage with the trait, θ, for item *i*, is

$$P_{ij}(\theta) = \frac{e^{a_i(\theta_j - b_i)}}{1+e^{a_i(\theta_j - b_i)}}, \quad (5)$$

where b_i represents stage-level difficulty and a_i represents stage-level discrimination. Stage-level difficulty is equal to the trait threshold which is the point at which the odds of endorsing the trait response becomes greater than the odds of not endorsing the trait response. For example, consider the stage representing the ERS trait. A lower difficulty, $b_{i,\,ERS}$, would suggest that respondents require lower levels of the ERS trait to select extreme responses. A higher difficulty would suggest that respondents require higher levels of the ERS trait to select extreme responses.

Polytomous IRTree Model

In the three-stage polytomous IRTree models, decisions for the first two stages, representing dichotomous decisions, were modeled using a unidimensional 2PL model. Decisions for the third stage, representing a polytomous decision, were modeled using the GRM (see example in Figure 2). The first two stages corresponded with the MRS and TOI traits and were modeled similarly to the first two stages of the dichotomous IRTree. For the third stage, corresponding with the ERS trait, the probability of an individual, *j*, selecting a category, *k*, for item *i*, is:

$$P_{ij,ERS,k}(\theta) = \begin{cases} P^*_{ij,ERS.k} & k = 2 \\ P^*_{ij,ERS.k} - P^*_{ij,ERS.k+1} & k = 1 \\ 1 - P^*_{ij,ERS.k+1} & k = 0 \end{cases} \quad (6)$$

Node-specific item parameters for ERS nodes in each model are constrained to be equal. Specifically, the item parameters for ERS nodes do not depend on whether the respondent agreed or disagreed with the item.

The model parameters were estimated using a Bayesian framework. The Bayesian framework allows for parameter estimation in complex models regardless of samples size, partly through the integration of prior knowledge using prior distributions. Four traits will be estimated for all respondents – the two TOI traits for each scale (i.e., θ_{IO}, θ_{CC}) and the two response style traits (i.e., θ_{MRS}, θ_{ERS}). Response style traits are constrained to be equal across the IO and CC scales because response styles are consistent tendencies regardless of the construct of interest.

Prior Distributions and Initial Values

The current study utilizes previous literature (i.e., Spratto et al., 2021) when selecting prior distributions and initial values. Trait estimates for θ_{IO}, $\theta_{CC,}$ θ_{MRS}, and θ_{ERS} were sampled from a multivariate normal distribution. This multivariate normal distribution had a vector of means equal to zero and a variance-covariance matrix with variances of 1. Covariances were symmetrical about the diagonal of variances. For example, $\sigma_{ERS,MRS}$ is equivalent to $\sigma_{MRS,ERS}$. Although the two subscales, IO and CC, were selected for their unrelated constructs, the covariance was freely estimated between the two traits. Covariances were estimated using an initial value of zero and the following prior:

$$N(0, var = 2, lower = -1, upper = 1) \quad (7)$$

The multivariate normal distribution is as stated below,

$$\begin{bmatrix} \theta_{IO} \\ \theta_{CC} \\ \theta_{MRS} \\ \theta_{ERS} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{IO,CC} & \sigma_{IO,MRS} & \sigma_{IO,ERS} \\ \sigma_{CC,IO} & 1 & \sigma_{CC,MRS} & \sigma_{CC,ERS} \\ \sigma_{MRS,IO} & \sigma_{MRS,CC} & 1 & \sigma_{MRS,ERS} \\ \sigma_{ERS,IO} & \sigma_{ERS,CC} & \sigma_{ERS,MRS} & 1 \end{bmatrix} \right) \quad (8)$$

The difficulty parameters for the MRS, $b_{j,MRS}$, and TOI, $b_{j\theta,TOI}$, stages in the polytomous model assumed a normal distribution with a mean value of 0 and a variance value of 3. These values may be updated as the study is run. The difficulty parameter for the ERS, $b_{j,ERS,k}$, stage in the polytomous model assumed a normal distribution with a mean value of 0 and a variance value of 3. Categories higher than 1 in the ERS stage were assumed to have a difficulty parameter with a truncated normal distribution with a lower bound of $b_{j,ERS,k-1}$. Both models assumed an informative prior for all discrimination parameters with a truncated standard normal distribution with a lower bound of 0.

Approximate Posterior

After determining the prior distributions for each parameter in the model, the posterior distribution is approximated. When a Bayesian posterior is analytically exhaustive to derive, the distribution can be approximated using simulation-based methods that randomly sample from the posterior distribution. One such simulation method is the Markov chain Monte Carlo (MCMC). To conceptualize MCMC, first consider the two aspects of MCMC, Markov chains and Monte Carlo method, separately. Monte Carlo is a method of estimating the properties of a distribution by examining random samples from the distribution (Ravenzwaaij et al., 2018). A Markov chain dictates that the random samples are generated by a special sequential process. Each random sample is a link in the Markov chain and each link is a stepping stone to generate the next random sample. Markov chains are stochastic models that generate new samples based on only the prior sample (Wohlin et al., 2003).

To run an MCMC procedure, a plausible initial value is selected and used to generate a new proposed value. The new proposed value is produced by adding random noise, generated from a proposal distribution to the initial value. The height of the posterior distribution at the new proposed value is compared against the height of the posterior distribution at the value previous to the proposed value. If the new proposed value has a higher posterior value than the previous posterior value, then the proposed value is accepted. If the new proposed value has a lower posterior value than the previous posterior value, then the proposed value is accepted or rejected probabilistically. The probability of acceptance is equal to the ratio of both posterior values. If the new proposed value is accepted it becomes the next link of the MCMC chain, otherwise, the next link is a repeat of the most recently accepted sample. This process represents one iteration and is repeated until enough iterations are run to represent the posterior distribution. This MCMC sampling method is known as the Metropolis algorithm (Ravenzwaaij et al., 2018).

Though plausible initial values are preferable, these values may be some arbitrary value that fits the constraints of the prior. Since the initial values selected may be incorrect, the beginning portion of the MCMC chain should be discarded as burn-in. For MCMC chains with less iterations, not discarding the burn-in phase may lead to incorrect reflections of the posterior distribution. When initial values are disparate, a new chain can be run with a better initial value or a lengthened burn-in phase to remove early samples from the non-stationary portion of the chain – where equilibrium is not met.

PROC MCMC. MCMC was used to estimate model parameters in SAS 9.4 software through the MCMC procedure. This procedure is designed to fit Bayesian

models. The MCMC procedure requires at least one PARMS statement, PRIOR statements equal to the number of parameters in the PARMS statements, and at least one MODEL statement. The PARMS statements are declarations of the parameters in the model and are used to assign initial values to the parameters. The PRIOR statements are where the prior distributions of the parameters are stated. The MODEL statements specify the likelihood functions of the outcome variables. Values for some of the options available in the PROC MCMC statement, such as number of iterations and amount of burn-in, are determined using previous literature.

Iterations and Burn-In. Leventhal et al. (2022), estimated IRTree models in a Bayesian framework by implementing the MCMC procedure in SAS (SAS Institute Inc., 2018). The MCMC procedure in Leventhal et al. (2022) consisted of 20,000 iterations in the burn-in phase and 80,000 iterations in the post burn-in phase for a total of 100,000 iterations. Based on previous literature, the current study's iterative MCMC procedure will initially sample from the posterior distribution 100,000 times. The initial 20,000 iterations will be discarded as burn-in and the following 80,000 iterations will be retained for estimation.

Check Convergence

Prior to conducting posterior inference, it is essential that convergence of the chain is assessed. If the chain of any parameter in the model has not fully converged, valid inferences cannot be made. Methods to check convergence include visual inspection of trace plots and statistical diagnostics tests.

Trace Plots. A trace plot will be developed for every parameter estimated in the model to visually inspect chain convergence. A visual inspection of these trace plots

should show stability in the mean and variance of the chain. If the visual inspection of the chain does not show convergence, then the iterations in the MCMC procedure will be increased by 5,000 – 10,000 iterations in the burn-in phase and 40,000 iterations in the post burn-in phase. Iterations are increased until all parameters in the model show visual convergence in the trace plots.

Visual convergence is determined by the speed in which the Markov chain traverses the parameter space. If the chain rapidly traverses the parameter space and has stabilized about a particular value of the parameter of interest, then there is evidence of convergence. If the chain slowly traverses the parameter space or takes small steps, then the chain has not converged. Even if a chain appears to have converged to a stable target distribution, there is a chance that the chain converged locally. Local convergence is when the chain appears to have converged during a visual check, usually for a smaller number of iterations, but if run for a larger number of iterations the chain shifts and converges to another location in the parameter space.

Convergence Diagnostic. After visual convergence is assessed, convergence is formally tested using the Geweke diagnostic (Geweke, 1992) by comparing running means of two portions of a chain to identify potential differences (Depaoli & van de Schoot, 2017). A *z*-test is used, comparing the first 10% and last 50% of the chain to see if the means for both sections of the chain are similar. If the two portions of the chain significantly differ, then full chain convergence was not met suggesting local convergence is an issue and a longer burn-in phase is necessary. The Geweke diagnostic is repeated, as needed, until the two portions of the chain do not significantly differ, suggesting that full chain convergence was met, and local convergence is not an issue.

Kernel Density Estimation Plot. A kernel density plot will be developed for each parameter to determine the approximate shape of the posterior distribution and visually analyze modality in the model. This density plot is a visual representation that uses a kernel density estimate to present a probability density function of the parameter. A satisfactory kernel density plot would present with smoother lines that are bell-shaped. This would suggest that there are enough iterations providing an adequate number of random samples. When the kernel density plot of the posterior is lumpy and not smooth, the chain may need more iterations to create a more reasonable summary of the posterior distribution.

Autocorrelation. Autocorrelation refers to a pattern of correlation in the chain, with sequential pulls of a parameter from the conditional distribution being correlated. The level of autocorrelation in a chain will often reduce as lags in the chain increases. Lag refers to the distance between two points of a chain. The autocorrelations for each parameter at varying amounts of lag will be checked using the posterior autocorrelations table in SAS. If autocorrelation does not reduce with an increased number of lags in the chain, there may be an issue with the model. One method to decrease autocorrelation in a chain uses a process called "thinning" – where every t^{th} sample from the chain (when $t > 1$) is selected to create the post burn-in sample in an attempt to lessen dependency in the posterior. Past research thinned the post burn-in sample to include every 15^{th} iteration (i.e., Spratto et al., 2021). If autocorrelation in the chain does not decrease after burn-in is discarded, the current study expects to mirror the thinning value in Spratto et al. (2021).

Research Questions

Difficulty values for the dichotomous models and category threshold values for the polytomous models were compared between item formats. The values compared were the Expected A Posteriori (EAP) and the highest posterior density (HPD) intervals. The term "Expected" in EAP refers to an expected value. The term "A Posteriori" in EAP refers to a posterior probability distribution of latent TOI scores. In sum, EAP is the expected value of the posterior distribution of a trait – or a point estimate of a parameter. An HPD interval is often used to refer to the smallest region of values that contain 95% of the posterior probability. Another interpretation of the HPD interval states that there is a 95% chance that the parameter's value is within the interval's range.

A higher EAP estimate for the difficulty or category threshold parameters would suggest a higher trait level is required to endorse the trait response. Inversely, a lower EAP estimate for the difficulty or category threshold parameters would suggest a lower trait level is required to endorse the trait response. EAP estimates of the differences are calculated by averaging the posterior mean differences across each iteration. Posterior mean differences are calculated by subtracting the draw from the posterior of the experimental condition's *b*-parameter from the draw from the posterior of the control condition's *b*-parameter, at each iteration. These calculated differences represent the posterior distribution of differences. Afterwards, I calculate the EAP and HPD interval for these differences. The EAP estimate of the differences between groups represents how much higher, or lower, the difficulty parameter is for the control group over the experimental group.

If the HPD intervals of the differences include 0, there would be a possibility that the difference between the difficulty parameters of the control and experimental group was not significant. If this range does not include 0, it is more likely that the difference between the difficulty parameters for the control group and the experimental group was significant. Comparing these EAP value differences and HPD intervals of the residuals between the two samples for each item will aid in answering the research questions: (1) Does one item format reduce midpoint response selections? (2) Does one item format reduce extreme response selections?

CHAPTER 4

Results

The research questions pertain to differing response style effects between item formats. Prior to interpreting the Bayesian results to answer the research questions, I interpreted descriptive statistics to get basic information about the variables and potential relationships among variables in the observed data. Then, I assessed convergence to the posterior distribution. After finding evidence of convergence for each of the item parameters in the model, I reviewed the analyses to answer the research questions.

Preliminary Analyses

A preliminary check of the data was done to assess any abnormalities in the data and to understand the distribution of responses. I calculated descriptive statistics for extreme and midpoint option selections in both control (Likert format) and experimental (funnel format) conditions for both subscales (see Table 1). For the preliminary analysis, I examined results related to extreme responses, followed by results related to midpoint responses. I conducted an independent samples *t*-test for both midpoint and extreme response types to test whether there was a significant difference in the mean number of response type selections and whether there was a significant difference in the mean differences of response type selections for the IO and CC subscales separately. Additionally, visual representations of the percentage of response type by scale for the control and experimental groups can be found in Figure 6 for midpoint responses and in Figure 7 for extreme responses.

Midpoint Responses

On average, the experimental group had a higher number of midpoint response selections for the IO subscale but not the CC subscale (see Table 1). The experimental group presents a higher percentage of midpoint responses in comparison to the control group for the majority of the items on the IO subscale, but not the CC subscale (see Figure 6). The mean midpoint response count for the IO scale was significantly higher for the experimental group (M = 2.16, SD = 1.75) than the control group (M = 1.83, SD = 1.59), $t(940.52) = -4.86$, $p < .001$, with a small effect size (d = 0.20). However, the mean midpoint response count for the CC scale did not significantly differ between control group (M = 1.29, SD = 1.27) and treatment group (M = 1.24, SD = 1.33), $t(4375) = 0.96$, $p = .337$, with an effect size of d = 0.04.

Extreme Responses

On average, the experimental group had a higher number of extreme response selections for both the IO and CC subscales (see Table 1). The experimental group presents a higher percentage of midpoint responses in comparison to the control group for all items except item 1 (see Figure 7). Specifically, for the IO subscale, the mean extreme response count was significantly higher for the experimental group (M = 1.02, SD = 1.19) than the control group (M = 0.62, SD = 1.18), $t(4375) = -8.28$, $p < .001$, with a small effect size (d = 0.34). For the CC subscale, the mean extreme response count was significantly higher for the experimental group (M = 1.44, SD = 1.25) than the control group (M = 0.98, SD = 1.38), $t(1064.8) = -8.78$, $p < .001$, with a small effect size (d = 0.35).

Convergence

It is vital to assess the convergence of the Markov chain before conducting any posterior inference. To reliably interpret the parameter estimates, evidence of convergence for each parameter in the model is required. There are two effective methods of assessing convergence: visually inspecting the trace plots and using statistical convergence diagnostics, such as Geweke's (1992) convergence diagnostic. I used a combination of visual and statistical convergence diagnostics to build evidence of convergence (see Figure 8).

Trace Plots

I conducted a visual check of the trace plots for each estimated parameter as a base check of convergence. Each trace plot traversed the parameter space quickly and efficiently, resulting in plots that resembled a "fuzzy caterpillar" (see Figure 8). Each trace plot stabilized about the mean for the parameter and varied within a set range of values from that mean. The trace plots did not exhibit any signs of local convergence. In other words, the chain shows evidence of convergence.

Convergence Diagnostic

Geweke's (1992) diagnostic was calculated as an additional check of convergence. Geweke's diagnostic tests the equality of the mean for the first 10% and the last 50% of iterations of the Markov chain after burn-in. The test statistic is the difference between the two means divided by the estimated standard error – or a standard *z*-score. For most of the parameters, the *z*-statistics that I calculated suggested that there were no significant differences between the means of the different portions of the chain. However, I found that seven of the 154 item parameters had a significant *p*-value using

Geweke's (1992) diagnostic (see Appendix A). However, when visually examining the trace plots of these parameters, I did not find any meaningful difference between the first 10% and the last 50% of iterations.

Kernel Density Estimation Plot

Kernel density estimation plots were utilized as a way to visualize the shape and modality of the data. The kernel density plots of each parameter exhibited a smooth normal bell-shape with a unimodal distribution (see Figure 8).

Autocorrelation

The MCMC procedure output autocorrelation plots for each parameter to specify the amount of autocorrelation for each of the posterior samples. Initial MCMC parameters resulted in autocorrelation concerns. When thinned by 20, the concerns were resolved. After thinning, autocorrelation for each of the parameters quickly moved to zero – within 10 lags – and remained trivial for higher lags (see Figure 8).

Convergence Determination

A visual check of the trace plots suggested convergence of the Markov chain. Although the Geweke's (1992) diagnostic values suggested convergence for the majority of the item parameters, a visual check of the trace plots suggested no significant difference in the two compared portions of the chain. The kernel density estimation plots exhibited smooth unimodal normal distributions and autocorrelation was low within 10 lags for all item parameters after thinning. Overall, I am confident that the MCMC algorithm converged to a stable posterior for all item parameters.

Primary Analyses

Interpretations are made using the posterior distributions after finding evidence of convergence. I used the posterior mean differences and HPD intervals of the differences to answer the research questions: (1) Does one item format reduce midpoint response selections? and (2) Does one item format reduce endpoint response selections?

I calculated the posterior mean differences between groups for each iteration in the posterior dataset. Individual posterior mean differences for the difficulty, or *b*, parameters were calculated by subtracting the draw from the posterior of the experimental condition's *b*-parameter value from the draw from the posterior of the control condition's *b*-parameter value for each individual iteration. Then a post-processing macro was run to calculate the mean and HPD intervals of the posterior differences across iterations. Posterior mean differences are the resulting EAP, or mean, from the post-processing macro run on the differences. The HPD intervals of the differences are the resulting HPD intervals from the post-processing macro run on the differences.

Dichotomous IRTree Model

Posterior mean differences and HPD intervals for the residuals were examined in the dichotomous IRTree for the MRS and ERS stages. EAP values and HPD intervals of the MRS stage difficulty parameters for each item can be found in Table 2. EAP values and HPD intervals of the ERS stage difficulty parameters for each item can be found in Table 3. EAP values and HPD intervals of the differences in difficulty parameters for each item can be found in Table 4 for the MRS stage and Table 5 for the ERS stage. A visualization of the EAP values and HPD intervals of the differences in difficulty

parameters for each item can be found in Figure 9 for the MRS stage and Figure 10 for the ERS stage. The dichotomous IRTree is represented by items 7 through 11 – or the CC subscale items.

MRS Stage. In the MRS stage of the dichotomous IRTree, the value of $b_{j,MRS}$ is indicative of the MRS trait level at which respondents have a .5 probability of choosing the midpoint response option. For example, $b_{7,MRS}$ had an EAP value of 1.37 for the control group and an EAP value of 1.51 for the experimental group. This suggests that in order to have ≥ 50% chance of selecting the midpoint, respondents need a $\theta_{i,MRS}$ level of ≥ 1.37 if in the control group and a $\theta_{i,MRS}$ level of ≥ 1.51 if in the experimental group. In other words, respondents need a higher MRS level to select the midpoint if given funnel items for item 7. EAP values and HPD intervals of the MRS stage difficulty parameters for each item can be found in Table 2.

Posterior mean differences for the MRS stage difficulty parameters were positive values for a majority of the items – the exception was item seven (see Figure 9). This means that the EAP values of the difficulty parameters for the majority of items in the MRS stage were greater for the control group than for the experimental group. In other words, a higher level of the MRS trait is required in the control group, than the experimental group, in order to have a .50 probability of endorsing the midpoint response option (see Figure 9).

The majority of the HPD intervals of the differences for the MRS stage in the dichotomous model included the value of 0 in their ranges. The only items that did not include 0 in their HPD interval range were items 9 and 10. In other words, the posterior mean difference of zero – or no difference – is not within the highest density range of

differences that contains 95% of the posterior distribution for two of the items (see Figure 9). This suggests in the MRS stage, there are meaningful differences in the *b* parameters between the control and experimental conditions for some of the items.

ERS Stage. In the ERS stage of the dichotomous IRTree, the value of $b_{7,ERS}$ is indicative of the ERS trait level at which respondents have a .5 probability of choosing the extreme response option. For example, $b_{7,ERS}$ had an EAP value of 1.24 for the control group and an EAP value of 1.15 for the experimental group. This suggests that in order to have ≥ 50% chance of selecting the extreme response, respondents need a $\theta_{i,ERS}$ level of ≥ 1.24 if in the control group and a $\theta_{i,ERS}$ level of ≥ 1.15 if in the experimental group. In other words, respondents need a higher ERS level to select the extreme response if given Likert items for item 7. EAP values and HPD intervals of the ERS stage difficulty parameters for each item can be found in Table 3.

Posterior mean differences for the ERS stage difficulty parameters were positive values for all of the items (see Figure 10). This means that the EAP values of the difficulty parameters for all of the items in the ERS stage were greater for the control group than for the experimental group. This suggests that a higher level of the ERS trait is required in the control group, than the experimental group, in order to have a .50 probability of endorsing the extreme response option.

Most of the HPD intervals of the differences for the ERS stage in the dichotomous model did not include the value of 0 in their ranges – the exceptions were items seven and 10. In other words, the posterior mean difference of zero – or no difference – is not within the highest density range of differences that contains 95% of the posterior distribution for the majority of the items (see Figure 10). This suggests, in the ERS stage,

there are meaningful differences in the *b* parameters between the control and experimental conditions.

Polytomous IRTree Model

Posterior mean differences and HPD intervals for the residuals were examined in the polytomous IRTree for the MRS and ERS stages. EAP values and HPD intervals of the MRS stage difficulty parameters for each item can be found in Table 2. EAP values and HPD intervals of the ERS stage difficulty parameters for each item can be found in Table 3. EAP values and HPD intervals of the differences in difficulty parameters for each item can be found in Table 4 for the MRS stage and Table 5 for the ERS stage. A visualization of the EAP values and HPD intervals of the differences in difficulty parameters for each item can be found in Figure 9 for the MRS stage and Figure 10 for the ERS stage. The polytomous IRTree is represented by items 1 through 6 – or the IO subscale items.

MRS Stage. Similar to the dichotomous IRTree, in the polytomous IRTree the value of $b_{j,MRS}$ is indicative of the MRS trait level at which respondents have a .5 probability of choosing the midpoint response option. For example, $b_{1,MRS}$ had an EAP value of 0.19 for the control group and an EAP value of 0.35 for the experimental group. This suggests that in order to have ≥ 50% chance of selecting the midpoint, respondents need a $\theta_{i,MRS}$ level of ≥ 0.19 if in the control group and a $\theta_{i,MRS}$ level of ≥ 0.35 if in the experimental group. In other words, respondents need a higher MRS level to select the midpoint if given funnel items for item 1.

Posterior mean differences for the MRS stage difficulty parameters were positive values for a majority of the items – the exception was item 1 (see Figure 9). This

indicates that the EAP values of the difficulty parameters for the majority of items in the MRS stage were greater for the control group than for the experimental group. This suggests that a higher level of the MRS trait is required in the control group, than the experimental group, in order to have a .50 probability of endorsing the midpoint response option.

Most of the HPD intervals of the differences for the MRS stage in the polytomous model did not include the value of 0 in their ranges – the exceptions were items 1 and 4. In other words, a posterior mean difference of zero – or no difference – is not within the highest density range of differences that contains 95% of the posterior distribution for the majority of items (see Figure 9). This suggests in the MRS stage, there are meaningful differences in the *b* parameters between the control and experimental conditions for some of the items.

ERS Stage. In the polytomous IRTree, there are two *b* parameters that can be interpreted in the ERS stage: $b_{j,k=1}$, and $b_{j,k=2}$. The $b_{j,k=1}$ parameter can be interpreted as the level of $\theta_{i,ERS}$ required to have a 50% chance of selecting a one or greater – or selecting one and two. However, the *b* parameter of interest is $b_{j,k=2}$, which can be interpreted as the level of $\theta_{i,ERS}$ required to have a 50% chance of selecting a two or greater – with category two referring to the most extreme option (see Figure 2). The $b_{j,k=2}$ parameter is the focus because it separates the extreme selection (k=2) from the non-extreme selections (k=0 and k=1), while the other *b* parameter just separates different non-extreme selections.

For example, $b_{1,ERS,k=1}$ had an EAP value of -0.31 for the control group and an EAP value of -0.59 for the experimental group. This suggests that in order to have a ≥

50% chance of selecting the response representing category one, or greater, respondents need a $\theta_{i,ERS}$ level of ≥ -0.31 if in the control group and $\theta_{i,ERS}$ level of ≥ -0.59 if in the experimental group. In other words, respondents need a higher ERS level to select the response representing category one, or greater, if given Likert items for item 1. However, this interpretation is of little interest when examining extreme response style in the polytomous IRTree.

For example, with the *b* parameter of interest, $b_{1,ERS,k=2}$ had an EAP value of 1.86 for the control group and an EAP value of 3.20 for the experimental group. This suggests that in order to have a ≥ 50% chance of selecting the response representing category two, or greater, respondents need a $\theta_{i,ERS}$ level of ≥ 1.86 if in the control group and $\theta_{i,ERS}$ level of ≥ 3.20 if in the experimental group. In other words, respondents need a higher ERS level to select the response representing category two, or greater, if given funnel items for item 1.

Posterior mean differences for the ERS stage difficulty parameters of interest were positive values for the majority of the items – the exception being item 1 (see Figure 10). This means that the EAP values of the difficulty parameters of interest for most of the items in the ERS stage were greater for the control group than for the experimental group. This suggests that a higher level of the ERS trait is required in the control group, than the experimental group, in order to have a .50 probability of endorsing the extreme response option, or greater.

Half of the HPD intervals of the differences of interest for the ERS stage in the polytomous model did not include the value of 0 in their ranges – the exceptions were items 2, 4, and 5. In other words, a posterior mean difference of zero – or no difference –

is not within the highest density range of differences that contains 95% of the posterior distribution for half of the items (see Figure 10). This suggests in the ERS stage, there are meaningful differences in the $b_{i,ERS,k=2}$ parameters between the control and experimental conditions for some of the items.

CHAPTER 5
Discussion

The purpose of administering a psychological instrument is to collect participants' responses to a construct of interest. The responses collected from these instruments are assumed to be accurate representations of the construct being measured. Proper interpretation of observed data is dependent on this assumption. Response styles disrupt this assumption by inducing biased responses from respondents (Baumgartner & Steenkamp, 2001; Cronbach, 1946). This study offers insight into methods used to account for and prevent response style effects – specifically, IRTrees and funnel item formatting. The purpose of the current thesis was to answer two research questions while using these methods. First, I wanted to investigate if the Likert item format or funnel item format related to fewer midpoint response selections. Second, I wanted to investigate if the Likert item format or the funnel item related to fewer extreme response selections. After collecting data using these methods, I modeled the responses with IRTrees. I then used difficulty parameter EAPs and HPD intervals of the differences to answer the research questions. Descriptive statistics are also helpful in understanding general trends in the observed data.

Research Question 1: Does one item format reduce midpoint response selection?

MRS – the tendency to select the midpoint regardless of the construct being measured – is hypothesized to affect observed scores dependent on difference of scale mean from the midpoint of the scale (Baumgartner & Steenkamp, 2001). Previous literature has altered item formatting as a way of preventing response style effects (Albaum et al., 2007; Böckenholt, 2017). The current study examined both Likert items

and funnel-format items, similar to Böckenholt (2017). I wanted to investigate if one item format was related to lower midpoint response selections over the other. In addition to enacting methods to prevent response style effects, I also aimed to account for them as if their effects are still likely to persist even with prevention methods.

Based on previous literature, the current study implemented IRTrees to account for response style effects (e.g., Böckenholt, 2012; Böckenholt & Meiser, 2017; Leventhal, 2019; Spratto et al., 2021). The first stage of the IRTree model corresponds with the MRS trait and is modelled using a unidimensional 2PL model. Item difficulty, or the *b* parameter, for the MRS stage indicates the level of MRS needed to be more likely than not to endorse the midpoint option.

There was some uniformity regarding which condition had higher stage-level difficulty values within each model for the MRS stage. This could be determined by looking at the sign of the EAP of the differences. The differences were calculated by subtracting the funnel item condition *b* parameter from the Likert item condition *b* parameter. This means that the EAP of the differences is positive when the Likert item condition has a higher difficulty level and negative if the funnel item condition has a higher difficulty level. Specifically, all items, with the exception of items 1 and 7, had higher difficulty values in the Likert item format condition than the funnel-item format condition.

Items 1 and 7 are the first items for the IO and CC subscale, respectively. These items may have had higher difficulty values in the funnel item format condition because of the novelty of the item type. Participants may have used the initial item to become familiar with the item format and to see all the stages. To do so, they must not select

neutral on the first item, potentially resulting in the item having a higher MRS stage threshold. After familiarizing themselves with the item they may have potentially began to select the midpoint to rush through the items. In the remaining items, that did not have HPD intervals of the differences overlapping with 0, the Likert item condition had higher difficulty parameter values than the funnel item condition.

This pattern indicates that a higher level of MRS is necessary to select midpoint responses in the Likert item format condition than in the funnel-item format condition. This suggests that midpoint response selection should be higher in the funnel-item format condition. This is supported by the pattern of midpoint selection presented in Table 1. The funnel-item format condition presented a higher number of midpoint response selections, on average, compared to the Likert item format for the IO subscale, but not the CC scale. This discrepancy between subscales may be caused by testing fatigue, as the CC subscale was administered after the IO subscale. Respondents in both testing conditions may have begun to select the midpoint as a neutral option to rush through the measures. The funnel item condition may have initially participated in this satisficing to a lesser degree because of the novelty of their item format (Cristofaro et al., 2022; Kieruj & Moors, 2010). This pattern of midpoint selection suggests that funnel items are potentially related to higher instances of midpoint response selection. In other words, Likert items are potentially related to lower instances of midpoint response selection.

However, results may not be solely directed by format differences. One potential influence on the results is testing scenario differences. The incoming first year students who completed the measures in Likert item format may have been more inclined to put effort into completing the subscales. This effort could have translated into fewer midpoint

option selections compared to the students from the university's research pool (Herzog & Bachman, 1981).

Students from the university's research pool who completed the measures in funnel-item format may have been less inclined to put effort into completing the subscales. These students recognize that they just have to complete the survey, not try, in order to receive course credit. These students may have been more motivated to rush through the items as a form of satisficing – engagement in substandard decision-making tactics to conserve cognitive effort (Barge & Gehlbach, 2012). To optimally rush through the funnel items, the neutral option can be continuously selected to only view one sub-item per item, rather than three sub-items per item if they were to express an opinion. This could be one reason why students in the funnel-item format condition exhibited higher average midpoint response selections in comparison to students in the Likert item format condition. This is especially relevant after the first item was administered as participants most likely began to understand the item format's structure.

Research Question 2: Does one item format reduce extreme response selection?

The current study examined two different item formats, Likert items and funnel-format items (Böckenholt, 2017). I wanted to investigate if one item format was related to lower extreme response selections over the other. This is because ERS – the tendency to select the extreme option regardless of the construct being measured – influences observed scores to be more positive, or more negative, depending on the direction of the scale mean (Baumgartner & Steenkamp, 2001).

Similar to previous literature, I utilized IRTrees to model ERS effects (e.g., Böckenholt, 2012; Böckenholt & Meiser, 2017; Leventhal, 2019; Spratto et al., 2021).

The third stage of the IRTree model corresponds with the ERS trait and is modelled using a unidimensional 2PL model for dichotomous options or the GRM for polytomous options. Item difficulty, or the *b* parameter, in the 2PL for the ERS stage indicates the level of ERS needed to be more likely than not to endorse the extreme option. Category threshold parameters, or b_{ik}, represent the boundary at which respondents have a 50% chance of selecting a particular category, *k*, or higher. For the current study, I primarily investigated the $b_{i,k=2}$ parameters in the GRM for the ERS stage.

Similar to the MRS stage, there was some uniformity regarding which condition had higher stage-level difficulty values within each model for the ERS stage. This could be determined by looking at the sign of the EAP of the differences. The differences were calculated by subtracting the funnel item condition *b* parameter from the Likert item condition *b* parameter. This means that the EAP of the differences is positive when the Likert item condition has a higher difficulty level and negative if the funnel item condition has a higher difficulty level. Specifically, all items with the exception of item 1 where $k = 2$ had higher difficulty values in the Likert item format condition than the funnel-item format condition. In most of the items that did not have HPD intervals of the differences overlapping with 0, the Likert item condition had higher difficulty parameter values than the funnel item condition.

This pattern indicates that a higher level of ERS is necessary to select extreme responses in the Likert item format condition than in the funnel-item format condition. This suggests that extreme response selection should be higher in the funnel-item format condition. This is supported by the pattern of extreme response selection presented in Table 1. The funnel-item format condition presented a higher number of extreme

response selections, on average, compared to the Likert item format for both the IO and CC subscales. This suggests that funnel items are potentially related to higher instances of extreme response selection. In other words, Likert items are potentially related to lower instances of extreme response selection.

Implications of Results

These results may have occurred for multiple potential reasons. One is that funnel items could result in more midpoint and extreme response selections than Likert items. Another is that funnel items may allow respondents to represent their trait levels more accurately through their response pattern.

Funnel items have been used in previous research to mitigate the effects of response styles (e.g., Albaum et al., 2007; Böckenholt, 2017). Böckenholt (2017) found that funnel items showed reduced response style effects in comparison to Likert items. However, the current study found that the funnel item condition exhibited lower difficulties for response style stages than the Likert item condition. This could be interpreted as a higher level of MRS and ERS effects because lower difficulties translate to lower levels of the trait required to be more likely than not to endorse that trait. In turn, these item parameter differences show up as more midpoint and extreme responses for funnel items as opposed to traditional Likert items. This result does not align with the reviewed research, but it can be argued that the funnel items potentially allow for more accurate representations of a respondent's true response.

This is because the funnel items were modelled after the hypothesized response process respondents navigate in order to reach a decision on an item (Böckenholt, 2012). The flexibility of the IRTree model allows for it to be consistent with this hypothesized

response process (Böckenholt, 2017). This can be suggestive of the IRTree modelling an accurate representation of respondents' true attitudes and that it matches data collection for the funnel item.

The decision to utilize Likert or funnel items is dependent on subjective weighting of the pros and cons. Funnel items are novel and may present results that are more indicative of the respondents' true attitudes in comparison to Likert items. This is because the funnel format is based on a response process that can be mirrored using IRTree models (Böckenholt, 2017). However, funnel items triple the length of the measure – three sub-items per item – and potentially increase the time taken to complete the measure which can lead to testing fatigue (Ben-Nun, 2008). Likert items are more commonly used and well-researched. They can also be implemented in a paper-and-pencil format whereas the funnel item requires presentation of items based on logic only a digital presentation (e.g., computer, tablet, phones) can provide. Pros and cons of both item formats can be further developed and researched to solidify the decision on which one to use.

Limitations and Future Research

This study, like others, is not without limitations. The current study lacks random assignment to conditions, assesses only one chain in the MCMC, and fails to completely investigate model-data fit. Future research can improve upon these limitations and further the research done in the current study.

To reiterate, the current study is not a true experiment despite using the term "experimental" as the participants were not randomly assigned to conditions. This was not feasible as the data used for analyses was secondary and collected for the purpose of a

different study. Future studies should randomly assign participants to different conditions.

Another limitation is due to using only one chain as a reference to determine convergence due to time and computational power constraints. Ideally, once visual convergence is met, the MCMC procedure would be run again with double the number of iterations (Depaoli & van de Schoot, 2017). This secondary MCMC procedure would be used as a secondary check of convergence to ensure local convergence is not an issue. A single Markov chain may not expose all of the potential issues with convergence, such as multi-modal distributions. For example, to explore the potential of multiple modes existing in the posterior distribution it is suggested that multiple chains are implemented for each model parameter (Depaoli & van de Schoot, 2017). Doubling the iterations of the MCMC procedure would show stability of the full-length chain and assess the chain for local convergence. Future research could conduct a secondary check of convergence on the new chain using trace plots, a convergence diagnostic, and with a computation of relative deviation.

Finally, the current study assumed model-data fit rather than formally testing it. Ideally in the Bayesian framework, model-data fit would be assessed using posterior predictive model checking (PPMC) to ensure the model minimizes discrepancies between the observed data and data simulated under the model. In future studies, after evidence of convergence is found, model-data fit could be assessed to ensure the model minimizes discrepancies between observed and simulated data. If there is good model-data fit, the simulated values generated with the model should be relatively similar to the observed values. For example, future research should focus on investigating mean and standard

deviation discrepancy statistics for this type of study to assure adequate recovery of parameters related to total score.

Conclusion

The results of this study suggest that items implemented in the traditional Likert item format are related to lower midpoint and extreme response selections. Despite producing higher midpoint and extreme response selections, funnel-format items may more accurately represent the respondents' response processes – representing their own response styles more accurately. Though this is one potential interpretation of the results, there are multiple ways the results can be interpreted.

As researchers search for methods to reduce negative response style effects, they may speculate about whether a Likert item or funnel item would be most beneficial for their measures. Though not a definitive answer to that dilemma, the results of this study should clarify some of the concerns that come up in that search. Continued research on this topic will continue to move us towards interpretations of results that better account for these response style effects.

Table 1

Descriptive Statistics for Response Type Counts for Control and Experimental Groups

Scale	Response Type	Group	N	Mean	Std. Dev.	*t*-value	*p*	*d*
IO	Extreme	Control	3671	0.62	1.18	-8.28	<.001	0.34
		Experimental	706	1.02	1.19			
	Midpoint	Control	3671	1.83	1.59	-4.86	<.001	0.20
		Experimental	706	2.16	1.75			
CC	Extreme	Control	3671	0.98	1.38	-8.20	<.001	0.35
		Experimental	706	1.44	1.25			
	Midpoint	Control	3671	1.29	1.27	0.96	0.337	0.04
		Experimental	706	1.24	1.33			

Table 2

MRS Stage b *parameter EAP and HPD Intervals for Control and Experimental Groups*

		Control				Experimental			
				HPD				HPD	
Subscale	Item	Mean	SD	Lower	Upper	Mean	SD	Lower	Upper
IO	1	0.19	0.04	0.11	0.27	0.35	0.10	0.16	0.56
	2	0.78	0.05	0.68	0.88	0.31	0.08	0.15	0.46
	3	0.98	0.06	0.88	1.10	0.55	0.09	0.39	0.72
	4	1.06	0.06	0.93	1.17	0.92	0.10	0.74	1.13
	5	1.58	0.08	1.43	1.75	0.77	0.08	0.61	0.94
	6	1.00	0.06	0.88	1.12	0.45	0.08	0.30	0.60
CC	7	1.37	0.10	1.18	1.57	1.51	0.15	1.23	1.83
	8	1.20	0.08	1.05	1.36	1.00	0.12	0.77	1.23
	9	1.60	0.10	1.41	1.81	0.78	0.13	0.55	1.03
	10	1.44	0.11	1.24	1.65	0.79	0.09	0.62	0.97
	11	1.67	0.11	1.46	1.89	1.37	0.13	1.13	1.63

Table 3

ERS Stage b *parameter EAP and HPD Intervals for Control and Experimental Groups*

			Control				Experimental			
					HPD				HPD	
Subscale	*k*	Item	Mean	SD	Lower	Upper	Mean	SD	Lower	Upper
IO	1	1	-0.31	0.05	-0.42	-0.21	-0.59	0.22	-1.03	-0.19
		2	-0.86	0.06	-0.99	-0.74	-1.49	0.37	-2.25	-0.86
		3	0.01	0.04	-0.07	0.09	-1.00	0.22	-1.44	-0.59
		4	0.22	0.05	0.13	0.31	-1.26	0.33	-1.90	-0.68
		5	-0.36	0.04	-0.45	-0.28	-1.54	0.40	-2.35	-0.88
		6	0.22	0.04	0.14	0.30	-0.44	0.14	-0.71	-0.17
	2	1	1.86	0.10	1.66	2.06	3.20	0.60	2.13	4.40
		2	1.26	0.07	1.12	1.39	1.21	0.29	0.69	1.78
		3	2.32	0.12	2.10	2.55	1.58	0.28	1.10	2.13
		4	2.35	0.13	2.11	2.60	1.70	0.38	1.06	2.46
		5	2.00	0.10	1.80	2.19	1.40	0.35	0.82	2.11
		6	2.31	0.11	2.10	2.53	1.60	0.25	1.17	2.11
CC	-	7	1.24	0.07	1.11	1.39	1.15	0.31	0.60	1.76
		8	1.44	0.07	1.30	1.58	1.00	0.19	0.67	1.37
		9	0.89	0.04	0.81	0.98	0.38	0.22	-0.01	0.79
		10	0.94	0.05	0.85	1.03	0.82	0.18	0.49	1.16
		11	0.87	0.04	0.78	0.96	0.46	0.15	0.18	0.75

Table 4

MRS Stage b *parameter EAP and HPD Intervals of the Differences*

		Differences			
				HPD	
Subscale	Item	Mean	SD	Lower	Upper
IO	1	-0.16	0.11	-0.39	0.05
	2	0.47	0.09	0.28	0.65
	3	0.43	0.10	0.21	0.63
	4	0.13	0.12	-0.11	0.37
	5	0.81	0.12	0.57	1.03
	6	0.54	0.10	0.34	0.73
CC	7	-0.14	0.18	-0.52	0.20
	8	0.20	0.14	-0.08	0.48
	9	0.82	0.16	0.51	1.14
	10	0.65	0.14	0.38	0.92
	11	0.30	0.17	-0.02	0.64

Table 5

ERS Stage b *parameter EAP and HPD Intervals of the Differences*

			Differences			
					HPD	
Subscale	*k*	Item	Mean	SD	Lower	Upper
IO	1	1	0.28	0.22	-0.13	0.73
		2	0.63	0.38	-0.01	1.39
		3	1.01	0.23	0.60	1.47
		4	1.48	0.33	0.90	2.14
		5	1.18	0.40	0.49	1.97
		6	0.66	0.15	0.38	0.95
	2	1	-1.34	0.61	-2.56	-0.24
		2	0.05	0.30	-0.55	0.58
		3	0.74	0.30	0.13	1.28
		4	0.65	0.40	-0.14	1.36
		5	0.60	0.36	-0.13	1.25
		6	0.71	0.27	0.16	1.22
CC	-	7	0.09	0.32	-0.54	0.65
		8	0.44	0.20	0.03	0.79
		9	0.51	0.22	0.09	0.92
		10	0.12	0.19	-0.24	0.46
		11	0.41	0.15	0.11	0.70

Figure 1

Example of a Fully Dichotomous IRTree Model

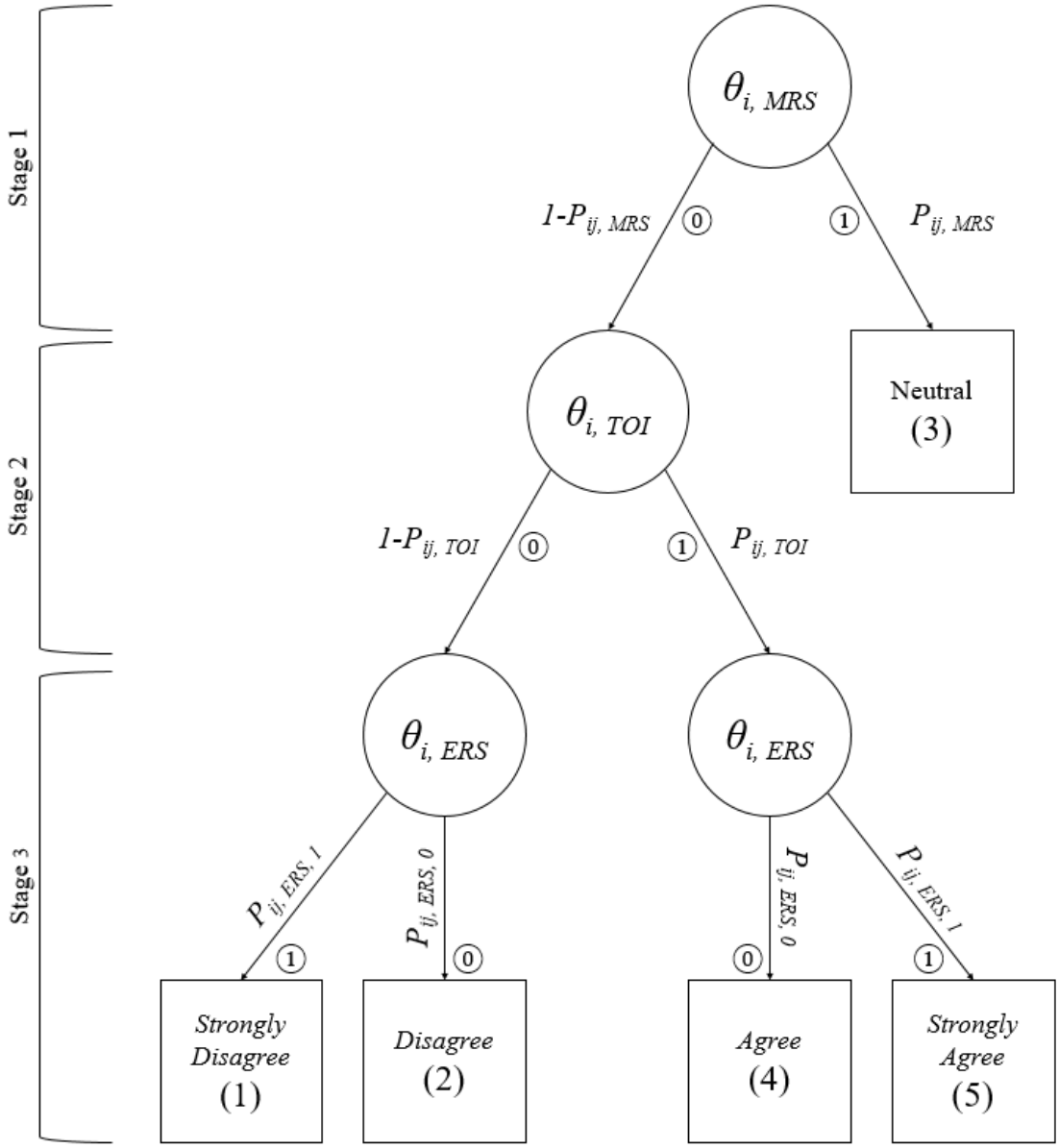


Figure 2

Example of a Partially Polytomous IRTree Model

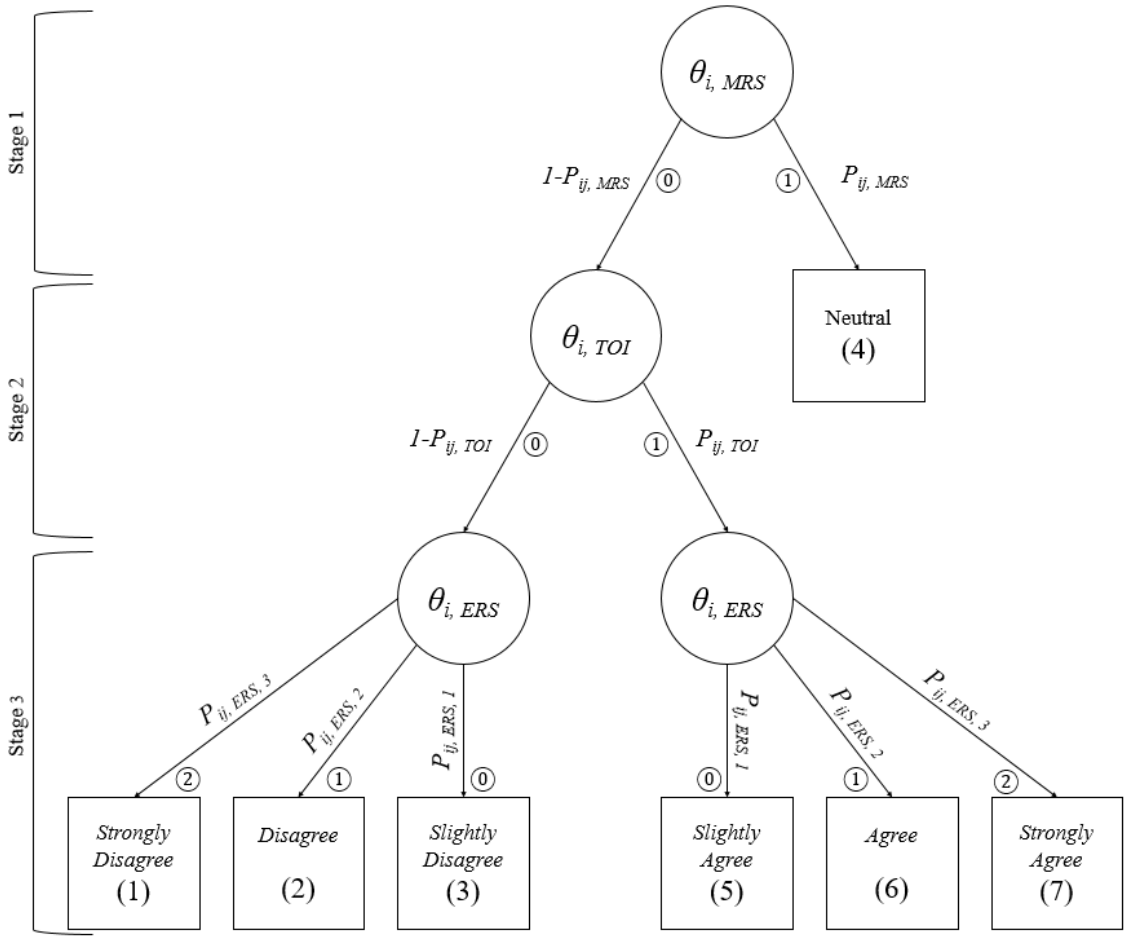


Figure 3

Example 7-point Likert Item

Please indicate the extent to which you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Disagree Slightly	Neither Agree nor Disagree	Agree Slightly	Agree	Strongly Agree
My ideas are usually better than other people's ideas.	○	○	○	○	○	○	○

Figure 4

Example 5-point Likert Item

Please indicate the extent to which you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I find it easy to speak up during class discussions.	○	○	○	○	○

Figure 5

Example Funnel Item

IO1_1. Do you have an opinion toward the following statement?

My ideas are usually better than other people's ideas.

◯ I have an opinion (1)

◯ I do not have an opinion (2)

If IO1_1 = 1, then display:

IO1_2. Given that you have an opinion, do you agree or disagree with the statement?

My ideas are usually better than other people's ideas.

◯ Agree (1)

◯ Disagree (2)

If IO1_2 = 1, then display:

IO1_3. Given that you agree, how strong is your opinion?

My ideas are usually better than other people's ideas.

◯ Strong (1)

◯ (2)

◯ Slight (3)

If IO1_2 = 2, then display:

IO1_4. Given that you disagree, how strong is your opinion?

My ideas are usually better than other people's ideas.

◯ Strong (1)

◯ (2)

◯ Slight (3)

Figure 6

Percentage of Midpoint Responses by Scale for the Control and Experimental Groups

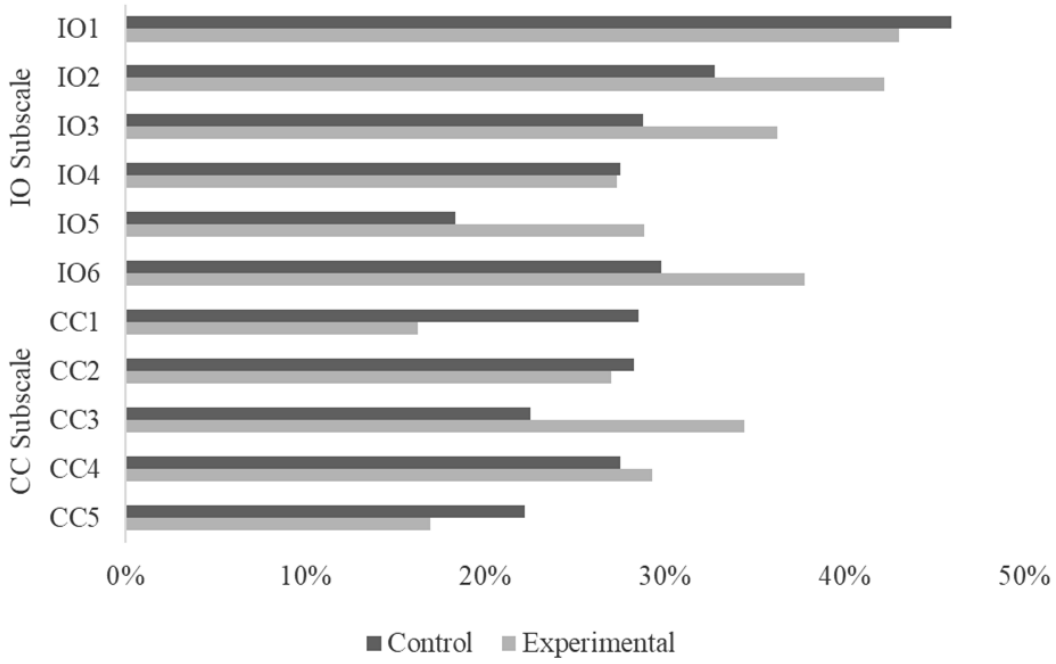


Figure 7

Percentage of Extreme Responses by Scale for the Control and Experimental Groups

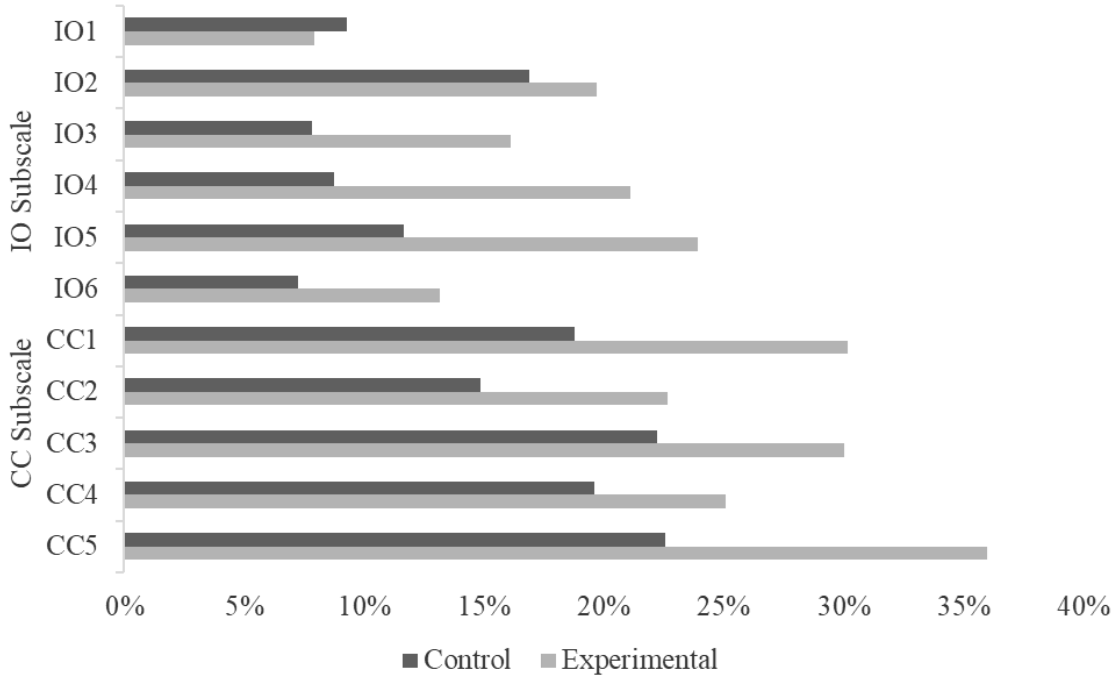


Figure 8

Example Visual Convergence Plots

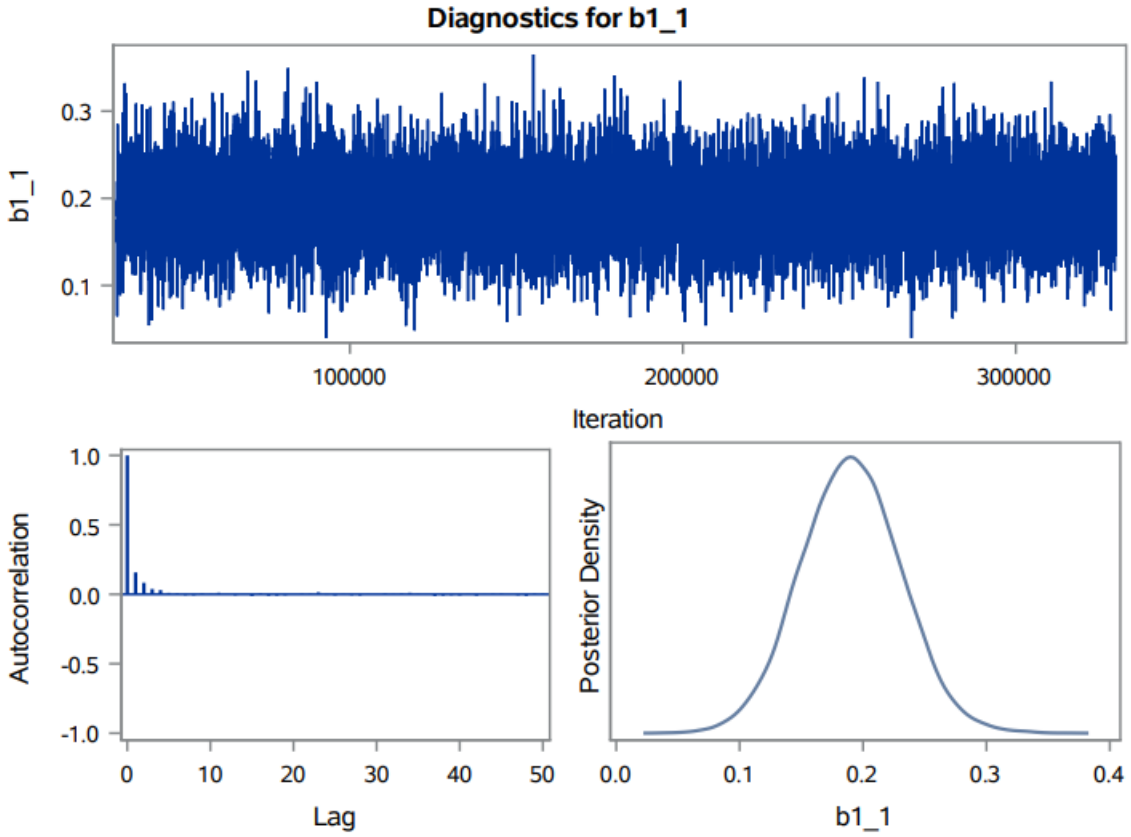
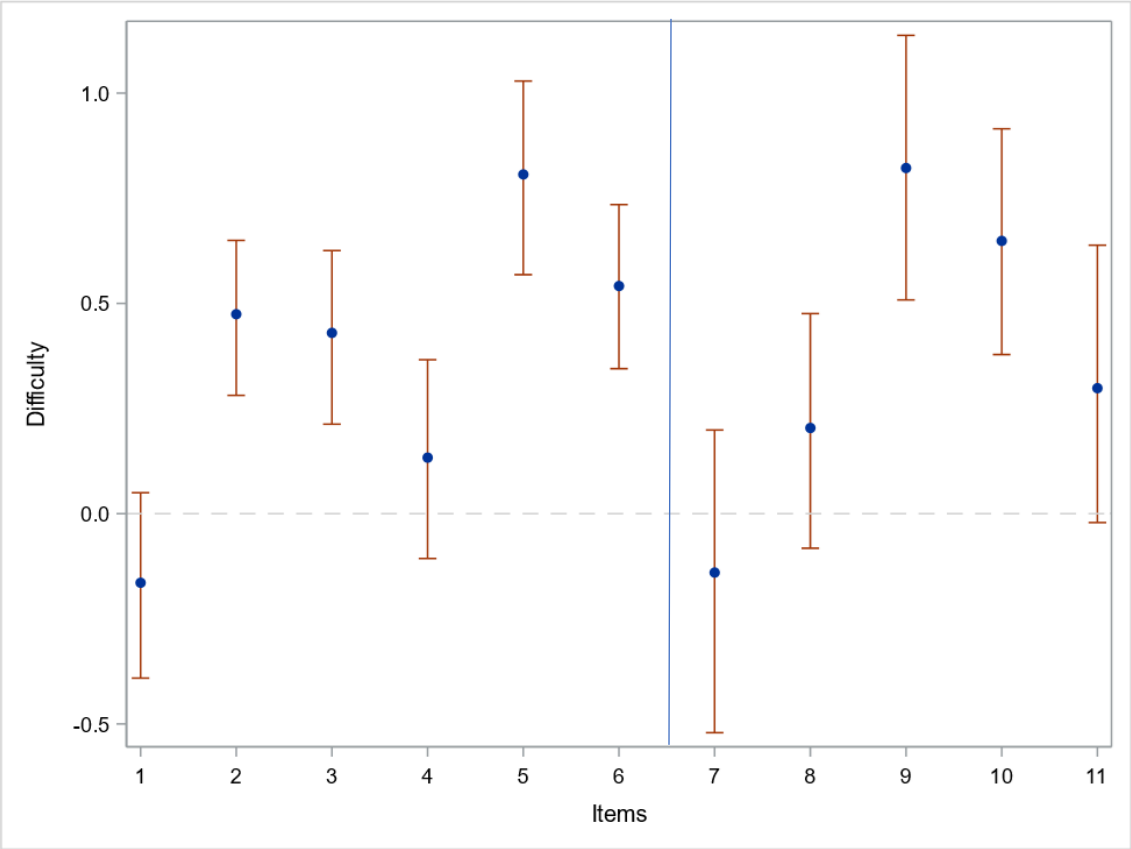


Figure 9

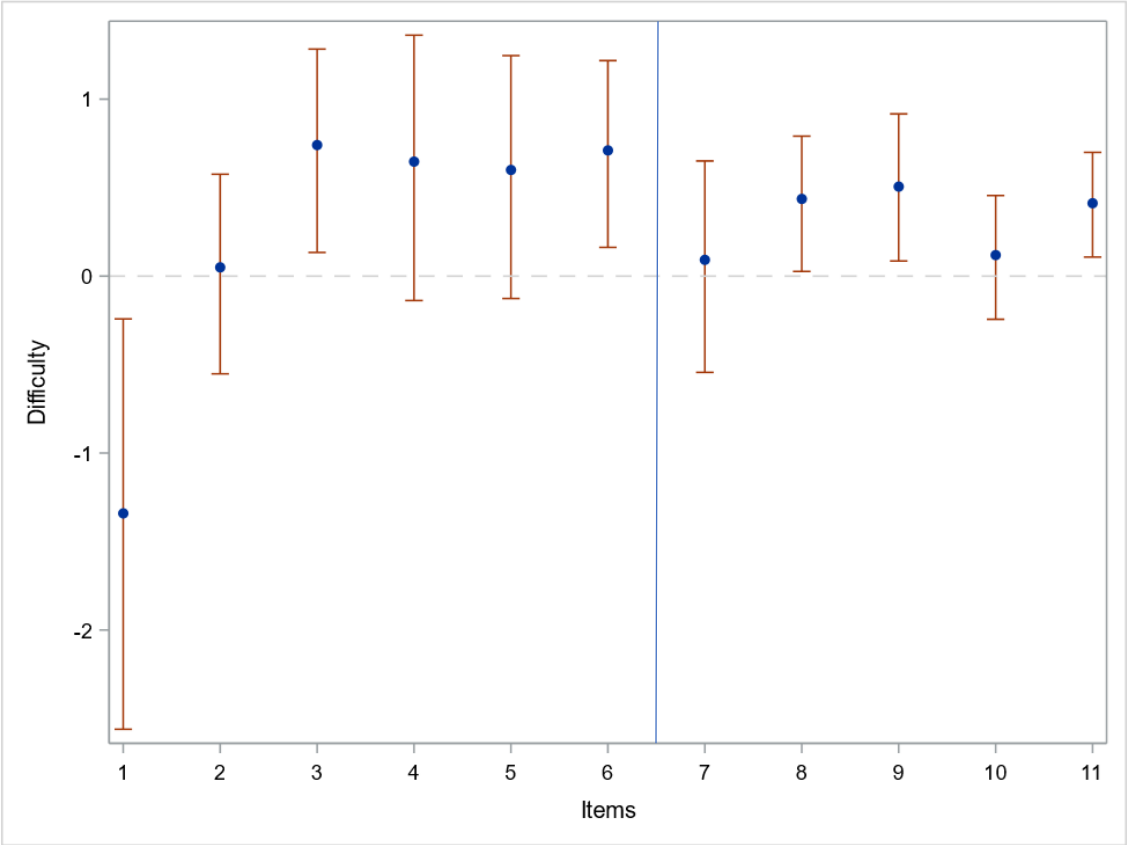
MRS Stage Difference in b *parameters and HPD Intervals of the Differences*



Note. The first six items are part of the IO subscale. The last five items are part of the CC subscale. These two subscales are divided in the figure by a solid line.

Figure 10

ERS Stage Difference in b *parameters and HPD Intervals of the Differences*



Note. The first six items are part of the IO subscale. The last five items are part of the CC subscale. These two subscales are divided in the figure by a solid line. The *b* parameters shown for the IO subscale are specifically the $b_{i,k=2}$ parameters.

Appendix A

Geweke's (1992) Convergence Diagnostic Values

Stage	Group	Item	Parameter	*z*-score	*p*-value
MRS	Control	1	a	1.71	0.09
			b	-0.08	0.94
		2	a	0.99	0.32
			b	-1.44	0.15
		3	a	1.39	0.16
			b	-0.86	0.39
		4	a	0.08	0.94
			b	-1.09	0.27
		5	a	-0.56	0.58
			b	0.45	0.66
		6	a	-0.42	0.68
			b	0.18	0.86
		7	a	0.94	0.35
			b	-1.15	0.25
		8	a	-0.90	0.37
			b	0.78	0.44
		9	a	-0.25	0.80
			b	0.11	0.91
		10	a	-1.29	0.20
			b	1.65	0.10
		11	a	-0.51	0.61
			b	0.47	0.64
	Experimental	1	a	-0.58	0.56
			b	-0.85	0.39
		2	a	-1.50	0.13
			b	-1.03	0.30

Stage	Group	Item	Parameter	*z*-score	*p*-value
		3	a	0.38	0.71
			b	-0.92	0.36
		4	a	0.68	0.50
			b	-0.76	0.45
		5	a	0.13	0.89
			b	-0.50	0.62
		6	a	-0.05	0.96
			b	-0.46	0.64
		7	a	0.01	0.99
			b	-0.60	0.55
		8	a	0.63	0.53
			b	-0.81	0.42
		9	a	1.65	0.10
			b	-0.77	0.44
		10	a	1.32	0.19
			b	-0.88	0.38
		11	a	-0.22	0.82
			b	-0.42	0.68
TOI	Control	1	a	0.00	1.00
			b	-0.11	0.91
		2	a	-0.07	0.95
			b	-0.12	0.90
		3	a	0.56	0.58
			b	-0.33	0.74
		4	a	1.24	0.21
			b	-0.62	0.54
		5	a	0.61	0.54
			b	-0.48	0.63
		6	a	1.02	0.31

Stage	Group	Item	Parameter	*z*-score	*p*-value
			b	0.77	0.44
		7	a	0.90	0.37
			b	0.58	0.56
		8	a	0.33	0.74
			b	0.74	0.46
		9	a	1.11	0.27
			b	1.53	0.13
		10	a	-0.53	0.60
			b	-0.05	0.96
		11	a	-0.98	0.33
			b	0.08	0.93
	Experimental	1	a	-0.15	0.88
			b	0.62	0.54
		2	a	1.00	0.32
			b	-1.31	0.19
		3	a	0.18	0.86
			b	0.19	0.85
		4	a	-0.81	0.42
			b	-0.36	0.72
		5	a	0.06	0.96
			b	-0.38	0.71
		6	a	0.24	0.81
			b	1.24	0.22
		7	a	-0.64	0.52
			b	0.56	0.58
		8	a	2.54	0.01
			b	-2.01	0.04
		9	a	1.74	0.08
			b	1.70	0.09

Stage	Group	Item	Parameter	*z*-score	*p*-value
		10	a	-0.78	0.43
			b	-1.00	0.32
		11	a	-0.61	0.54
			b	-1.50	0.13
ERS	Control	1	a	1.63	0.10
			b1	0.21	0.83
			b2	-2.02	0.04
		2	a	1.62	0.10
			b1	0.35	0.73
			b2	-2.22	0.03
		3	a	1.27	0.20
			b1	-0.34	0.73
			b2	-1.62	0.10
		4	a	-0.83	0.40
			b1	-0.89	0.37
			b2	0.25	0.80
		5	a	1.31	0.19
			b1	0.27	0.79
			b2	-1.33	0.18
		6	a	-0.53	0.59
			b1	-1.64	0.10
			b2	-0.29	0.77
		7	a	0.11	0.91
			b	-0.41	0.68
		8	a	0.57	0.57
			b	-0.69	0.49
		9	a	-0.44	0.66
			b	-0.19	0.85
		10	a	-0.25	0.80

Stage	Group	Item	Parameter	*z*-score	*p*-value
			b	-0.94	0.35
		11	a	0.32	0.75
			b	-0.72	0.47
	Experimental	1	a	1.87	0.06
			b1	1.37	0.17
			b2	-1.74	0.08
		2	a	0.79	0.43
			b1	0.60	0.55
			b2	0.55	0.59
		3	a	0.19	0.85
			b1	0.43	0.67
			b2	-0.24	0.81
		4	a	-0.72	0.47
			b1	-0.66	0.51
			b2	1.08	0.28
		5	a	-0.32	0.75
			b1	0.06	0.95
			b2	0.98	0.33
		6	a	-1.32	0.19
			b1	-0.41	0.68
			b2	2.01	0.04
		7	a	-0.40	0.69
			b	0.22	0.83
		8	a	1.78	0.07
			b	-0.69	0.49
		9	a	0.22	0.83
			b	-0.04	0.97
		10	a	0.15	0.88
			b	0.66	0.51

Stage	Group	Item	Parameter	*z*-score	*p*-value
		11	a	-0.98	0.33
			b	0.51	0.61

References

Albaum, G., Roster, C., Yu, J. H., & Rogers, R. D. (2007). Simple rating scale formats: Exploring extreme response. *International Journal of Market Research, 49*(5), 1–10. https://doi.org/10.1177/147078530704900508

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ames, A. J., & Leventhal, B. C. (2021). Application of a Longitudinal IRTree Model: Response Style Changes Over Time. *Assessment*, https://doi.org/10.1177/10731911211042932**.**

Asún, R., Rdz-Navarro, K. & Alvarado, J. (2017). The sirens' call in psychometrics: The invariance of IRT models. *Theory & Psychology*, *27*(3), 389-406. https://doi.org/10.1177/0959354317706272

Bachman, J. G., & O'Malley, P. M. (1984). Black-White differences in self-esteem: Are they affected by response styles? *American Journal of Sociology, 90*(3), 624–639. https://doi.org/10.1086/228120

Barge, S., & Gelhbach, 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., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Ben-Nun, P. (2008). Respondent fatigue. In *Encyclopedia of survey research methods.* Sage Publications, Inc., https://doi.org/10.4135/9781412963947

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*(4), 608–628. https://doi.org/10.1207/S15328007SEM0704_5

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*(1), 69–83. https://doi.org/10.1037/met0000106

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *The British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181. https://doi.org/10.1111/bmsp.12086

Bolt D. M., & Johnson T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*(5), 335-352. https://doi.org/10.1177/0146621608329891

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*(5), 814–833. https://doi.org/10.1177/0013164410388411

Buckley J. (2009). *Cross-national response styles in international educational assessments: Evidence from PISA 2006*. Retrieved from: https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf

Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among east Asian and north American students. *Psychological Science*, *6*(3), 170–175. http://www.jstor.org/stable/40063010

Clarke, I. III. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior & Personality, 15*(1), 137–152.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology, 60*(2), 151–174. https://doi.org/10.1037/h0040372

Cristofaro, M., Giardino, P. L., Malizia, A. P., & Mastrogiorgio, A. (2022). Affect and cognition in managerial decision making: A systematic literature review of neuroscience evidence. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.762993

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494. https://doi.org/10.1177/001316444600600405

de Ayala, R. J. (2009). *The theory and practice of item response theory* (1st ed.). New York: Guilford Press.

DeMars, C. (2010). *Item response theory.* Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

DeMars, C. E. & Jacovidis, J. N. (2016, April). *Multilevel IRT: When is local independence violated?* Electronic board presented at the annual meeting of the National Council on Measurement and Education, Washington, DC.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods, 22*(2), 240–261. https://doi.org/10.1037/met0000065

Downing, S. M. (2002). *Threats to validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation, 7*, 235-241. https://doi.org/10.1023/A:1021112514626

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press. https://doi.org/10.4324/9781410605269

Falk, C. F., & Ju, U. (2020). Estimation of response styles using the multidimensional nominal response model: A tutorial and comparison with sum scores. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.00072

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. https://doi.org/10.21034/sr.148Greenleaf, E. A. (1992a). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research, 29*(2), 176–188. https://doi.org/10.2307/3172568

Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly, 56*(3), 328–351. https://doi.org/10.1086/269326

Greenleaf, E. (2008). Extreme response style. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 257-276). Sage Publications, Inc., https://dx.doi.org/10.4135/9781412963947.n173

Herzog, A. R. & Bachman, J. G. (1981) Effects of questionnaire length on response quality. *Public Opinion Quarterly, 45*(4). https://doi.org/10.1086/268687

Hurley, J. R. (1998). Timidity as a response style to psychological questionnaires. *The Journal of Psychology, 132*(2), 201-210. https://doi.org/10.1080/00223989809599159

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: Macmillan

Kieruj, N. D. & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International Journal of Public Opinion Research, 22*(3), 320-342. https://doi.org/10.1093/ijpor/edq001

Krumrei-Mancuso, E. J., &; Rouse, S. V. (2015). The development and validation of the comprehensive intellectual humility scale. *Journal of Personality Assessment, 98*(2), 209–221. https://doi.org/10.1080/00223891.2015.1068174

Leventhal, B. C. (2019) Extreme response style: A simulation study comparison of three multidimensional item response models. *Applied Psychological Measurement, 43*(4), 322-335. https:// doi.org/10.1177/0146621618789392

Leventhal, B. C., Gregg, N., & Ames, A. J. (2022). Accounting for response styles: Leveraging the benefits of combining response process data collection and response process analysis methods. *Measurement: Interdisciplinary Research and Perspectives, 20*(3), 151-174. https:// doi.org/10.1080/15366367.2021.1953315

Leventhal, B. C., & Stone, C. A. (2018). Bayesian analysis of multidimensional item response theory models: A discussion and illustration of three response style models. *Measurement: Interdisciplinary Research and Perspectives, 16*(2), 114–128. https://doi.org/10.1080/15366367.2018.1437306

Martin, J. (1964). Acquiescence—measurement and theory. *British Journal of Social and Clinical Psychology*, *3*(3), 216–225. https://doi.org/10.1111/j.2044-8260.1964.tb00430.x

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics*. Amsterdam, the Netherlands: Eleven International.

Messick, S. (1968). Response sets. In D. F. Sills (Ed.), *International encyclopedia of the social sciences* (Vol. 13, pp. 492-496). Macmillan.

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity*, *42*(6), 779–794. https://doi.org/10.1007/s11135-006-9067-x

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov chain Monte–Carlo sampling. *Psychonomic Bulletin & Review, 25*(1), 143–154. https://doi.org/10.3758/s13423-016-1015-8

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *The Journal of Social Psychology, 121*, 81–96. https://doi.org/10.1080/00224545.1983.9924470

Reynolds, N., & Smith, A. (2010). Assessing the impact of response styles on cross-cultural service quality evaluation: A simplified approach to eliminating the problem. *Journal of Service Research*, *13*(2), 230–243. https://doi.org/10.1177/1094670509360408

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97. https://doi.org/10.1007/bf03372160

SAS Institute Inc. (2018). *SAS/STAT® 9.4 User's Guide,* Cary, NC: SAS Institute Inc.

Schuman, H., Presser, S., & Ludwig, J. (1981). Context effects on survey responses to questions about abortion. *Public Opinion Quarterly*, *45*(2), 216–223. https://doi.org/10.1086/268652

Smith, P. B. (2017). Response style. In *The SAGE encyclopedia of communication research methods.* https://dx.doi.org/10.4135/9781483381411

Sosu, E. M. (2013). The development and psychometric validation of a critical thinking disposition scale. *Thinking Skills and Creativity*, *9*, 107–119. https://doi.org/10.1016/j.tsc.2012.09.002

Spratto, E. M., Leventhal, B. C., & Bandalos, D. L. (2021). Seeing the forest and the trees: Comparison of two IRTree models to investigate the impact of full versus endpoint-only response option labeling. *Educational and psychological measurement*, *81*(1), 39–60. https://doi.org/10.1177/0013164420918655

Stening, B. W., & Everett, J. E. (1984). Response styles in a cross-cultural managerial study. *The Journal of Social Psychology*, *122*(2), 151–156. https://doi.org/10.1080/00224545.1984.9713475

Vaerenbergh, Y.V., & Thomas, T.D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(3), 195-217. https://doi.org/10.1093/ijpor/eds021

Weijters, B. (2006). *Response styles in consumer research* (Publication No. 4100284) [Doctoral dissertation, Ghent University]. UGent Academic Bibliography. http://hdl.handle.net/1854/LU-4100284

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: the number of response categories and response category labels.

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

Williams, L. M., Horst, S. J., & Sundre. D. L. (2014). Test of oral communication skills, version 2: TOCS-II test manual. Harrisonburg, VA: Center for Assessment and Research Studies and Madison Assessment. http://www.madisonassessment.com/uploads/TOCS2%20Manual%20(2014).pdf

Wohlin, C., Höst, M., Runeson, P., Wesslén, A., & Meyers, R. (2003). Software reliability. In Meyers, R. A. (Ed.), *Encyclopedia of Physical Science and Technology* (3rd ed., pp. 25–39). Academic Press.