4-15-2020

Can Ratings of Item Location Enhance Statistical Item Parameter Estimation? Extending the Feasibility of Unfolding IRT Models

Michael McKenna

University of Missouri-St. Louis, mgmn7d@mail.umsl.edu

Can Ratings of Item Location Enhance Statistical Item Parameter Estimation? Extending the Feasibility of Unfolding IRT Models

Michael G. McKenna Jr.

University of Missouri-St. Louis

Dissertation submitted to the Graduate College of University of Missouri at St. Louis in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology with an emphasis in Industrial & Organizational Psychology

May 2020

Committee:

Stephanie Merritt, PhD, Advisor

Garett Foster, PhD

John Meriac, PhD

Cody Ding, PhD

Abstract

Research and development of modern psychometric methods such as item response theory have drastically changed the way we understand and carry out the measurement of psychological constructs. Despite this, there has been relatively little adoption by psychological researchers to incorporate these methods into their research. While multiple explanations are surely valid, one oft stated reason is the large sample size requirements of these methods. The sample size requirements of item response theory are needed so that effective estimation of item parameters can be carried out. In an attempt to make these modern measurement methods more accessible and feasible to psychological researchers, this study investigated the extent to which subject matter experts and trained novices could effectively rate the location parameter of items to use as starting parameters in the item parameter estimation process. Rather than starting with random values, as is the default approach, starting with more accurate item locations was hypothesized to result in just as accurate item parameters that do not require typical sample sizes for these models. A pseudo-simulation process was carried out to estimate parameter recovery at various sample sizes when using SME and trainee ratings of item locations as starting parameters. Results suggest that while SMEs and trainees were not able to perfectly align item location parameters with statistical estimations, person estimates derived when using these as starting parameters yielded quite similar results to the parameters from the default MML procedure. Similar results were uncovered across sample sizes. Additionally, as sample size decreased from 500 to 200, recovery results became less stable indicating that even with SME and trainee estimates of item location

used as starting parameters, sample size issues still remained when estimating item parameters.

Acknowledgements

I would like to express my deepest gratitude to Dr. Garett Foster for his supervision and encouragement on this research project. I would also like to acknowledge and thank my committee, Drs. Merritt, Meriac, and Ding for their feedback and continued support through this process. In addition, I would like to acknowledge Dr. Carol Shoptaugh for her continued dedication to me and my career throughout my time at both Missouri State University and the University of Missouri-St. Louis. I would not be where I am today without her guidance, motivation, and friendship. Finally, I would like to thank my mother, Laurie McKenna. Her voice of reason and her insight were more helpful during my academic career than she will ever know.

This dissertation is dedicated to my wife, Lisa.

Can Ratings of Item Location Replace Statistical Item Parameter Estimation? Extending the Feasibility Unfolding IRT Models

To understand psychological phenomena, researchers typically develop tools that require participants to respond to a number of various questions or statements. Psychometricians have devised different ways to model the way in which a person responds to different types of statements. One type of model that has emerged as particularly useful for personality and attitude assessment is known as ideal point models. While extremely useful, there appears to be a lack of adoption of this modeling framework within typical psychological research. Empirical investigations that require self-report psychological measurement typically utilize a classical test theory framework which is much simpler to employ but is commonly critiqued as a vast oversimplification of the process by which a person responds to a personality or attitude statement. One commonly cited potential reason for this lack of adoption of ideal point models (and other item response theory models) is because of the large sample sizes needed to accurately estimate both item and person parameters. The goal of this dissertation was to investigate an alternative item estimation technique that would potentially reduce the need for such large sample sizes and make these ideal point models more accessible and feasible for all psychological researchers.

A Brief History of Psychological Measurement

Since the inception of psychology, the field has depended on the ability to accurately measure psychological constructs. While various forms of measurement were developed and utilized over the past century, none has been more impactful than self-report responses to various forms of stimuli. This measurement method has allowed

researchers and practitioners to easily ascertain insights about a person that cannot be objectively observed. Self-report methods, while providing a unique window into human behavior, are not without their issues unfortunately. Among numerous other substantive issues (e.g., cognitive biases, response distortion, etc.), they require a large amount of attention and development to ensure that they are both reliable and accurately measure the psychological construct intended. While other research should continue to examine ways to mitigate the other issues, the focus of this investigation was enhancing the feasibility and utility of these methods. To ensure these, psychometricians have attempted to develop a number of different models with different assumptions over the years to better explain and understand the way people respond to different types of psychological items.

The Early Years. Louis Thurstone developed one of the first techniques to measure psychological constructs (1927; 1928; 1929). He proposed that a person would endorse a statement that they felt closely aligned with their perception of their level on the construct or attitude. He argued that if the statement was a perfect measure of the trait and the person was thorough, they would appropriately respond to an item. Using his notation, suppose there were N_1 people with a specific value, S_1, on a psychological construct. Thurstone argued that in reality, only n_1 people would be expected to agree with a statement with a value of S_1, due to various issues, where $n_1 < N_1$. Additionally, those who did not endorse the statement with a value S_1 would be expected to endorse a statement with a different value, S_2, and the probability of endorsement is inversely related to the absolute value of the distance between S_2 and S_1. This simply means that as a statement gets further away in either direction from the person's level on the trait, there

is a decreasing probability that they will endorse that statement. His scaling technique required statements measuring the full range of a psychological construct. As explained by Drasgow, Chernyshenko, and Stark (2010), Thurstone used a 6-item scale measuring militarism (with pacificism at the other extreme). Figure 1 shows the location of the six items used. Figure 1 also indicates the distribution of militarism in some population. Thurstone argued that a person who is just below moderate on militarism (on the side of pacifism) would only endorse statements in the range of *d* and *e*. This person might reject items to the left of *d* because they indicate very strong level of pacifism. And they would also reject the entire range of militaristic options. While intuitive, the scoring for this procedure was fairly complex for the time period. At the time, it required a rather large amount of computational power when large numbers of statements and participants were used.

A much simpler alternative to Thurstone scaling was developed by Likert (1932) only a few years later. Likert (1932) found that one could use the entire range of the construct as the response options for each item by requiring endorsement of one of five responses with “Strongly Disapprove” and “Strongly Approve” on either end. Likert found that for statements that represented very low levels of the construct, one could reverse code those items and then simply take the mean or sum of the item scores, which could be used as a person’s score or level of the trait in question. This was a pivotal development and became the de facto method for self-report scaling and continues to be to this day. Later research began to focus more attention at better explaining the way in which a person actually responds to a statement and how that can be used to understand their standing on the trait.

Unfolding and Dominance Models. In 1964, Coombs coined the term dominance process models which included the scaling technique described by Likert. Dominance models assume a monotonic relationship between a person's standing on the trait and their probability of statement endorsement (Carter et al., 2014). Figure 2 illustrates a typical item response function assuming a dominance approach, for a dichotomously scored item. The Greek letter theta (θ) is typically used to represent the trait or construct of interest in item response theory (IRT) terminology. As you can see, as theta, in other words the trait of interest, increases, the probability of a positive response also increases. Coombs (1964) is also credited with coining the term unfolding response process. This process describes what Thurstone postulated decades earlier. Coombs decided on the term unfolding because the probability of endorsement decreases, or unfolds, in both directions from the individual's ideal point (Drasgow et al., 2010). In other words, statements measuring higher and lower locations, than the individual's ideal point, on the latent trait continuum have a decreasing probability of being endorsed the further they are from the individual's location on the trait. Thus, unfolding models assume a nonmonotonic relationship between the person's level of the trait and the probability of statement endorsement (see Figure 3).

Test and Item Based Theories. Only a few years after Coombs coined these terms, a seminal piece of work was published in Lord and Novick's *Statistical Theories of Mental Test Scores* (1968) based on the work of Allen Birnbaum. The latter part of the book presents and explains Birnbaum's logistic function which was proposed to model the relationship between an underlying trait, θ, and the probability of endorsing an item. This work is one of the earliest descriptions of a model that falls within a collection of

models known as IRT or sometimes referred to as latent trait models. These models are considered the "new age" of psychological measurement and are much more complex than their counterpart, classical test theory (CTT). Classical test theory has been the mainstay for test development for most of the 20th century (Embretson & Resie, 2000) and focuses specifically on the test as a whole, rather than the individual items within the test. Item response theory models that were conceived directly from Birnbaum's work – the 1-, 2-, and 3-parameter logistic models – fall within the framework of dominance models described earlier alongside all techniques within CTT.

Explosion of IRT Models. After the publication of Lord and Novick's (1968) text, there was an abundance of published literature that developed new estimation techniques and proposed new IRT models (Foster et al., 2017). This explosion of models led researchers to develop computer programs that were capable of estimating item parameters as well as person-level parameters such as BILOG (Mislevy & Bock, 1990), MULTILOG (Thissen, 1991), LOGIST (Wingersky, Barton, & Lord, 1982), NORMOG (Kolakowski & Bock, 1973), and PARSCALE (Muraki & Bock, 1991). These programs made it easier for psychometricians to utilize these models in research and applied settings. To psychometricians, it became clear that IRT was superior to CTT for a variety of reasons. Most notably, the parameters of an item are considered to be invariant across subpopulations (Guion & Ironson, 1983). Developers of tests used in applied settings (e.g., employee selection and educational testing) have recognized this and other advantages and have utilized IRT during test development for decades. However, psychological research seems to have failed to fully embrace these advantages (Foster et al., 2017). Additionally, IRT underlies an important assessment technique known as

computerized adaptive testing (CAT). This technique uses IRT to optimally select items that are most appropriate for an examinee, given the current estimate of their latent trait, or θ (Embretson & Reise, 2000).

Some models, such as the 1-, 2-, and 3-PL models, were developed for tests that used dichotomously scored items. In other words, items that had a right answer and that resulted in responses that were binary (e.g., 0 = incorrect, 1 = correct). Other models were also developed to score items that utilize polytomous responses like that of Likert's scaling. The graded response model (GRM; Samejima, 1969) appears to be the most widely used polytomous IRT model (Foster et al., 2017), but the generalized partial credit model (GPCM; Muraki, 1992) is another polytomous model that is available and commonly referenced in IRT texts.

Reemergence of Unfolding Research. All of the IRT models referenced thus far assume a dominance response process (Carter et al., 2014). They all assume that as the latent trait (θ) increases, so too does the probability of positively responding to an item. In the early 1990's however, unfolding response models crept back into the literature with the publication of Andrich's (1993) application of the hyperbolic cosine to latent trait modeling. He explained it as a symmetric function that effectively reflects the two reasons one might disagree with a statement. For ordered response categories, such as Likert's response format, if a person was presented with an item that represented a moderate level of the underlying trait, they could disagree with the statement for two reasons. Andrich (1993) eloquently argued this point using the example item *I don't believe in capital punishment, but I am not sure it isn't necessary* (p. 254). He explained that when considering a simple agree-disagree dichotomous response, a person can either

(1) agree with the statement, (2) disagree because they are very much for capital punishment (i.e., disagree from above), or (3) disagree because they are very much against capital punishment (i.e., disagree from below). Andrich realized that as the distance between the person and item increases, the probability of a positive response decreases in both directions (Andrich, 1993). Thus, unfolding models are also referred to as ideal point modeling because the closer an item gets to the person's location on the trait, or ideal point, the more likely they are to agree with the item. Andrich applied the hyperbolic cosine to mathematically capture this unfolding process. As Figure 3 illustrates, for a moderate item, only those with a theta level around 0 would be likely to endorse. Those with very low levels of theta would likely disagree because the item represented too much of the trait and those with very high levels of theta would disagree because the item represented too little of the trait.

The important difference between dominance response processes like Likert's scaling and unfolding models concerns the utility of these moderately worded items (Drasgow et al., 2010). A key metric of item quality proposed by Likert was an item's correlation with the total score of the test or scale. These intermediate or moderately worded items yield poor item-total correlations (Chernyshenko, Stark, Drasgow, & Roberts, 2007) and thus were suggested for removal or avoidance by Likert (Drasgow et al., 2010). Contemporary researchers, however, realized the important contribution that moderate items could provide to accurate scoring. As Drasgow and colleagues argue (Drasgow et al., 2010), most people do not fall towards the extremes of attitudes or traits if you assume they are normally distributed in the population. Thus, the ability to effectively measure those who are moderate should yield higher reliability and validity

(Drasgow et al., 2010) and more precise rank ordering. Andrich's (1993) application of the hyperbolic cosine to the unfolding process was utilized in a model that has gained some recognition today.

The Generalized Graded Unfolding Model

The generalized graded unfolding model (GGUM) proposed by Roberts and colleagues (Roberts, Donoghue, & Laughlin, 1999; Roberts, Donoghue, & Laughlin, 2000) has become the go-to unfolding or ideal point model and is an extension of the partial credit model and GPCM (Muraki, 1992). Roberts and colleagues (Roberts et al., 1999) designed the GGUM to handle polytomous data and expressed the formula as:

$$P(z_i = z \mid \theta_j) = \frac{\exp\{\alpha_i[z(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}]\} + \exp\{\alpha_i[(M - z)(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}]\}}{\sum_{w=0}^{C}\left\{\exp\{\alpha_i[w(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}]\} + \exp\{\alpha_i[(M - z)(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}]\}\right\}}, \quad (1)$$

which states that the probability of person *j* giving response *z* as their observed response to item *i* is a function of the distance between the person location parameter (θ_j) and the item location parameter (δ_i) on the latent trait continuum ($\theta_j - \delta_i$). This formula models the probability of a given observed response ($z = 0$ for the strongest level of disagreement to $z = C$ for the strongest level of agreement) as a function of the corresponding subjective response (i.e. agreeing because the item is close, disagreeing because the item is higher on the continuum than the person, or disagreeing because the item is lower on the continuum than the person, with M representing the strongest level of agreement from above the item and $M = 2C + 1$). This is what allows the GGUM to discern the meaning of non-endorsement of moderate items. The GGUM is argued to be more general than other ideal point IRT models because it allows items to vary in their discrimination (α_i) and threshold (τ_{ik}) parameters (Roberts, Donoghue, & Laughlin, 1999). The discrimination

parameter resembles α parameters in dominance processes, in other words, the ability of the item to differentiate between persons close to one another on the underlying trait. The τ parameters signify the point at which a person will change response options (e.g., the point at which someone will select strongly disagree rather than disagree).

Roberts, Donoghue, and Laughlin (1999; 2000) proposed a marginal maximum likelihood (MML) method for estimating item level parameters (δ_i, α_i, τ_{ik}) and an expected a posteriori (EAP) technique for estimating person parameters (θ_j). Multiple simulation studies were conducted by Roberts and his colleagues (Roberts, Donoghue, & Laughlin, 2002) to ascertain the performance of these techniques under varying conditions. A main goal of the study was to understand the sample size requirements using these estimation procedures and how they fared as sample size decreased. As suggested previously by the authors (Roberts, Donoghue, & Laughlin, 1998), accuracy estimates stabilized when about 750 simulated examinees were included. Additionally, they found that δ_i were easier to estimate than α_i, which in turn were easier to estimate than τ_{ik}.

Research aimed at advancing the utility and functioning of the GGUM was carried on by Roberts. In 2008, Roberts extended an item fit statistic that was developed by Orlando & Thissen (2000) to the GGUM. And Roberts and Thompson (2011) developed a new technique to estimate item parameters that utilizes a marginal maximum a posteriori (MMAP) estimation. While the authors argue that the MMAP approach combines the efficiency of MML estimation with Bayesian prior distributions, the advantages over MML were only evident for items with extreme δ parameters that had few response category options (i.e., two or three per item; Roberts & Thompson, 2011).

These advances of the GGUM and the understanding of ideal point models more generally have led to a number of studies that have proposed that an ideal point approach may be more appropriate for measuring noncognitive constructs such as personality, attitudes, or interests (Carter et al., 2014; Drasgow et al., 2010; Stark, Chernyshenko, Drasgow, & Williams, 2006). While the work of Thurstone, Andrich and Roberts and colleagues focused mainly on the measurement of attitudes, a bulk of the work since then has applied these models to personality constructs. Researchers have argued that personality statements are essentially attitudes about oneself (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Drasgow et al., 2010). Thus, ideal point models should fit personality items relatively well. In 2010, by way of a focal article in *Industrial and Organizational Psychology*, Drasgow, Stark, and Chernyshenko laid out the utility of ideal point models and called on organizational researchers and practitioners to begin utilizing these models to more appropriately measure psychological constructs. This article aimed to explain unfolding models and highlight why appropriate measurement is important.

GGUM in Research. One of the first investigations of unfolding models by Industrial and Organizational Psychology (I/O) researchers was by Stark and colleagues (Stark et al., 2006). Their research examined the assumptions of item responses for personality scales. They tested whether personality data fit an ideal point model better than a dominance one, using the GGUM. The major finding was that for the 16 Personality Factor Questionnaire (Conn & Rieke, 1994), ideal point models provided better model-data fit and lead to increased item information compared to dominance models. The important takeaway was that ideal point models should be considered for

personality modeling because they can fit monotonically increasing item response functions but is not a requirement or an assumption. Further research added to this evidence showing the scale or test development process could benefit from the flexibility added by adopting an ideal point approach from the early stages of the development process (Chernyshenko et al., 2007). This research showed that the development of discriminable moderate or neutral items was possible when the GGUM was used to model responses. When these items were analyzed using a CTT or dominance IRT approaches, they were removed due to low item-total correlations and small α, or discrimination, parameters.

In their 2010 focal article, Drasgow et al. (2010) put forth their arguments and evidence to support the claim that ideal point models are more appropriate than dominance for modeling personality items. While the commentary surrounding the focal article did provide some dissenting views (Reise, 2010; Spector & Brannick, 2010) overall there were a lot of endorsements or clarifications of the arguments. This could be argued to be a turning point for the measurement of personality in applied psychology. It likely opened researchers' eyes to the potential that these alternative models provide to non-cognitive measurement. Drasgow et al.'s (2010) arguments focused on research, but other researchers have examined how this model may affect applied decisions.

Findings from applied research have also revealed the importance of appropriately modeling personality data. In an organizational sample, Carter et al. (2014) compared curvilinear relationships between the personality construct of conscientiousness and job performance, a relationship that is typically assumed to be linear and monotonic. The difference was based was the modeling process used to measure scores of

conscientiousness, either an ideal point process (i.e., GGUM) or a dominance process (i.e., a sum score). Their findings suggested that curvilinear relationships were more likely to be found when conscientiousness was measured using an ideal point process. This also directly affected potential applied decisions. This was shown when focusing on the top 100 scorers on the conscientiousness measure, which is a common predictor used in employee selection. When the ideal point method was used to measure conscientiousness, it would have resulted in less undesirable employee attrition. These two studies suggest that personality measurement may be most appropriate when assuming an ideal point response process. The remaining literature on the GGUM has focused on its ability to model forced choice pairwise preference items (Stark, Chernyshenko, & Drasgow, 2005; Chernyshenko et al., 2009).

GGUM in Practice. The research support for the utility of the GGUM that was described in the previous sections seems to be recognized in applied settings. The applications of the GGUM in these settings, however, differs slightly compared to the response process explained earlier. One of the most impressive aspects of the GGUM is that is can fit both binary and graded responses (Chernyshenko et al., 2007). This allows the GGUM to fit pairwise-preference forced choice items in addition to standard graded responses like Likert's scaling (Chernyshenko et al., 2007). These forced-choice items require that respondents select one statement of two that are presented that most closely resembles the way they feel. While this is considered a special case of the GGUM, pairwise-preference forced choice item types are argued to have advantages over Likert response formats (Chernyshenko et al., 2007; Drasgow et al., 2010) and is a popular technique for CAT. Because of these advantages, large consulting firms and public

institutions have developed assessments using the GGUM in combination with forced choice formats.

Institutions on both sides of the applied spectrum have applied the GGUM in high stakes settings. The first was the Tailored Adaptive Personality Assessment System (TAPAS; Drasgow, Chernyshenko, & Stark, 2010). This assessment was developed to support United States Army personnel in making selection and classification decisions. Another application of the GGUM to forced choice personality measurement was done by a large consulting firm that focuses on employee selection and assessment (Boyce, Conway, & Caputo, 2016). A third instance of this application was conducted by the Educational Testing Service. As a part of their employee assessment suite, they developed an assessment for job fit based on personality traits (Naemi, Seybert, Robbins, & Kyllonen, 2014). These applications of the GGUM within measurement settings with very high stakes support the wealth of research around the GGUM and its effectiveness at modeling personality data. Whether the response format is forced choice CAT or utilizes a straightforward Likert format, the GGUM holds real potential for modeling personality, as well as other psychological constructs that require introspection (e.g., attitudes, interests, etc.).

Despite the evidence for the advantages associated with using the GGUM to measure constructs like personality by measurement researchers and applied institutions, there is lack of utilization of unfolding models, and the GGUM more specifically, in psychological research. Foster et al.'s (2017) review shows that the GGUM seems to be the popular choice for modeling the unfolding process. It further reveals that since Roberts and colleague's initial publication in 2000, it has only been utilized in a research

context eight times across 17 of the field of I/O's most prominent and popular journals. While similar studies across other areas of the psychological sciences seem to be absent, it stands to reason that this trend continues across the span of psychological research.

A Lack of Measurement Precision in Psychological Research

Broadly speaking, it has been argued that psychological research has failed to integrate the advances of psychometrics over the past few decades (Borsboom, 2006). In the preface of *Item Response Theory for Psychologists*, Embretson and Reise (2000) specifically state that they intended for the content to be geared toward a general psychological audience; an audience that only measures constructs like personality or cognitive ability as a part of broader goals to test complex hypotheses. Four years later, it was noted by Embretson (2004) that the large majority of psychological tests were still being constructed with a CTT approach rather than an IRT one. There are many potential explanations for this issue and Borsboom eloquently lays out some of them in an article published in 2006 titled *Attack of the Psychometricians*. He argues that psychological research rarely focuses on developing a model structure to relate an observed score to an underlying, theoretical attribute (à la IRT). Rather, most research assumes the true score, in CTT terms, is in fixed relation to the observed score on some measure. In a reply commentary to Borsboom's article, Clark (2006) yields similar concerns and even takes this argument one step further. She argues that Borsboom does not go far enough in that,

> "He does not criticize what likely are thousands of published studies in which the outcome of an experimental manipulation or the difference between two naturally occurring groups is assessed with an instrument or procedure developed for that particular study, with the resulting scores treated as a psychological construct

> (i.e., attribute), with no apparent thought given to the measurement issues involved" (2006, p. 448).

This claim rings even more true, today, as we see the field of psychology descending into a period wrought with reproducibility issues and a lack of credibility. Aside from Borsboom's (2006) abstract arguments about the way psychological researchers think and behave, he also posits several pragmatic factors to explain this lack of integration. A lack of training in the areas of modern test theory / IRT, non-inclusion of these modeling techniques in popular psychological software (i.e., SPSS), and feasibility of large sample sizes are all suggested as logical reasons for this disconnect. While Borsboom's focal article generated a good deal of counterpoints (Clark, 2006; Merenda, 2007; Heiser; 2006), it does seem that even those who dissented did not disagree with the focal claim that a lack of measurement precision is evident in psychology. Sentiments similar to Borsboom's (2006) have been argued by researchers and practitioners within the field of I/O in recent years. The importance of accurate measurement is not lost on the average psychological researcher, but the criticality may be more apparent to those in this subfield because measurement is often used to make important decisions (Carter et al., 2014).

If Borsboom's and other's arguments are valid and non-psychometrically focused psychologists are not attempting to harness the potential of these more modern techniques, the question is why? This question is quite general could be applied to any psychometric model. However, models based on dominance assumptions such as the 1-, 2-, and 3-PL models and the GPCM tend to not fit certain types of constructs well. In psychology, a very large proportion of the literature uses measurements of personality, attitudes, and interests in researching various human behavior. These constructs, as

previously explained, are better modeled with an ideal point framework (Chernyshenko et al., 2001; Stark et al., 2006; Chernyshenko et al., 2007). Thus, rather than focusing on IRT in general, it makes more sense to focus on why IRT models that are specifically geared towards modeling these kinds of constructs are not being used more in psychological research. The GGUM is one such model and has been argued to hold great promise for the field of organizational research (Foster et al., 2017). Using forced-choice pairwise preference items, it has garnered notable applications in the applied employee assessment field from large consulting firms and public institutions (Houston, Borman, Farmer & Bearden, 2006; Boyce, Conway, & Caputo, 2016). While these applications and the little bit of research that continues to be done is useful and necessary, it appears the GGUM has failed to gain widespread acclaim from psychological researchers as of today. This begs the question: Why? The question may have multiple answers.

GGUM's Failed Adoption in Psychological Research

Several reasons could explain why the GGUM has not yet been adopted by broader psychological researcher. The first may be due to the fact that there are a limited number of available, intuitive software programs capable of carrying out item and person parameter estimation (Foster et al., 2017; Lui & Chalmers, 2018). However, the GGUM2004 program (Roberts, Fang, Cui, & Wang, 2006) is freely available, designed specifically for item and person parameter estimation and has been around for almost 15 years. However, the GGUM2004 program is not very user-friendly and requires knowledge of Fortran programming language for data input. More recently, multiple R packages have been developed to carry out GGUM estimation and GGUM estimation has also been added to a general IRT estimation package (Tendeiro & Castro-Alvarez, 2018;

King & Roberts, 2015, Lui & Chalmers, 2018), all of which are also freely available on the Comprehensive R Archive Network (R Core Team, 2018). Therefore, while the GGUM2004 program may be slow and complex to operate and the R packages may not yet be well known, it seems that there is a more likely explanation for the lack of utilization in psychological research.

The second potential reason for the lack of GGUM application in research may be due to the perceived complexity of the model (Foster et al., 2017). Researchers may be intimidated by the complexity or feel like they do not have an adequate understanding of the model. Similarly, they may not feel they have the necessary expertise to correctly utilize it. Foster et al. (2017) report that, of those who reported not using IRT, 21% explained it was due to their lack of relevant education and training surrounding IRT in general. While no psychological researcher would advocate utilizing a model without proper knowledge of its properties and assumptions, there is plenty of research noting the benefits of properly modeling item responses (Stark et al., 2006; Carter et al., 2014). Thus, as research and information on ideal point models continue to grow, psychological researchers would undoubtedly benefit from learning the ins and outs of these models. While this explanation for the lack of research utilizing the GGUM may hold true and future research may be fruitful, another explanation may also require empirical investigation.

The final, more likely reason for the slow acceptance of IRT and the GGUM by researchers is the large sample sizes needed to effectively estimate stable person and item parameters. As Hambleton and Jones (1994) report, “Sample size ranks as one of the most important factors that affect the item calibration task” (p. 268). Various researchers

have reported this as one of the main barriers to utilizing the most appropriate measurement methodologies (Borsboom, 2006; Dalal, Withrow, Gibby, & Zickar, 2010; de le Torre & Hong, 2010; Stark, Chernyshenko, & Guenole, 2011; Foster et al., 2017; Sahin & Anil, 2017). In their review of IRT, Foster and colleagues (Foster et al., 2017) uncover that the GGUM research consistently utilizes much larger samples compared to CTT based research. They explain this is one of the hindrances of the use of IRT models, and the GGUM more specifically. Dalal et al. (2010) answer potential practitioner questions regarding ideal point models, one of which focused on sample size. They note that the sample size requirements for ideal point models are likely larger than participant pools that most practitioners have at their disposal. Ideal point IRT models even require a larger sample size than their dominance IRT counterparts (Dalal et al., 2010). For IRT models more generally, the sample size requirements have been argued to be a real burden to their utilization in psychological research (Borsboom, 2006). This hinderance has been cited across many domains of psychological research, but the vast majority comes from educational (for example, see Sahin & Anil, 2017) and applied (for example, see de le Torre & Hong, 2010 or Dalal et al., 2010) domains.

It is well known that the required sample size for IRT models is related to the number of parameters that are estimated in the model (Reise & Embretson, 2000; Dalal et al., 2010). Therefore, it is not surprising that the GGUM requires very large sample sizes (Roberts et al., 2002; Foster et al., 2017). For polytomous item formats, the GGUM estimates item location (δ), discrimination (α), subjective category threshold (τ), and the person parameters (θ; Roberts et al., 2000). Thus, the GGUM requires a large number of parameters to be estimated. Also, depending on the number of response options used, the

number of estimated parameters increases even more. To accurately estimate the large number of parameters, Roberts et al.'s (2000) research suggested a minimum sample size of 750. Generalized graded unfolding model researchers rightfully collect a large amount of data, as evidenced by Foster and colleagues'

findings (2017). Their results showed that the median sample size used in research was just over 600. While the GGUM was the smallest median sample size observed, it is still large compared to most psychology research with an average sample size of around 200 (Marszalek, Barber, & Kohlhart, 2011). They go on to argue that new ways of estimating IRT parameters will allow for more widespread adoption in psychological research (Foster et al., 2017). Furthermore, one of the few empirical investigations that utilized the GGUM (Carter et al., 2014) ended with the limitation that sample size for the GGUM will always concern researchers.

Alternative Parameter Estimation. In recent years, calls have been made for investigations into alternative ways to estimate unfolding item parameters that require substantially smaller sample sizes (Dalal et al., 2010; Foster et al., 2017; Carter et al., 2014). Dalal et al. (2010) explicitly ask how applied researchers can score ideal point measures with a sample size of 100, for example. They provide clarification to the reader, that researchers will rarely have access to large enough sample sizes to effectively utilize unfolding models. The intent of the questions posed in this response article seem to be to clarify certain aspects of ideal point models. Additionally, they try to generate potential future research questions to continue to enhance the utility of ideal point models. They state, "…practitioners would benefit from any research or advice on creating and evaluating ideal point scales when only a limited sample size is available" (p. 499). Other

researchers have heeded the clarifications posed by Dalal and colleagues (Carter et al., 2014). In their research described above, Carter et al. (2014) include sample size as limitation to their study findings. Interestingly, they quickly discuss the potential utility of Thurstonian scoring to reduce sample size for parameter estimation but note that sample sizes of around 300 were suggested for this framework decades ago (Guildford, 1954). Finally, Foster and colleagues (Foster et al., 2017) end their discussion of the GGUM with an appeal to researchers to continue developing different ways to estimate these models with fewer people. They state this will lead to a broader use of these methods.

Psychometricians have attempted to develop better estimation procedures and programs. The MML estimation procedure originally developed by Roberts et al. (2000) seems to be the most widely implemented item estimation technique. Other statistical approaches to item estimation have been developed since then, however. A more advanced approach to item estimation utilized Markov Chain Monte Carlo estimation procedures (MCMC; de la Torre, Stark, & Chernyshenko, 2006; Wang, 2013; Wang, de la Torre, & Drasgow, 2015). The ultimate intent was not specifically focused on reducing sample size requirements, but it was an aspect of the simulation studies that were conducted (de la Torre et al., 2006; Wang, 2013). The MCMC estimation outperformed MML using all of the metrics that were calculated in both studies. Regarding sample size specifically, estimation accuracy for both techniques improved as sample size increased (de la Torre et al., 2006; Wang, 2013). Wang (2013) found that samples of around 500 were adequate to achieve stable estimates. Though this trend was revealed to be more dramatic for MML estimation. In addition to MCMC estimation, another statistical

estimation approach for the GGUM has been developed, marginal maximum a posteriori (MMAP; Roberts & Thompson, 2011). Similar to MCMC estimation procedures, MMAP estimation accuracy was enhanced as sample size increased. These attempts to develop new statistical estimation procedures have been fruitful. But given the large majority of the estimation programs available still utilize the MML procedure, and the one that does not - MCMC GGUM (Wang et al., 2015) - has not resulted in required sample sizes that make the GGUM more feasible to psychological researchers.

It appears that there are no statistical approaches to item estimation that lend themselves to the sample sizes used by most psychological researchers. The original MML approach, along with the newer MCMC and MMAP approaches, all require sample sizes well over 500 to effectively achieve stable item estimates. On average, psychological research yields sample sizes around 200, and in 2006, the Journal of Applied Psychology had a median sample size of around 150 (Marzalek et al., 2011). Thus, it is unsurprising that IRT and the GGUM have not been utilized more in psychological research. Perhaps however, there are novel item parameter estimation techniques that enhances the statistical aspect that necessitates the need for a large sample size. If that were the case, it may provide an alternative approach to item estimation that allows a broader range of psychological researchers to enjoy the benefits of its advantages.

One of the first established techniques that might be used in lieu of statistical estimation is the process proposed by Thurstone in the 1920's. He established his method prior to the development of IRT or its parameters. But the general notion of his technique could lend itself nicely to certain item parameter estimation. Very generally, Thurstone

scaling requires statements to be presented to a group of judges. These judges rank order the statements from the most favorable / extreme to the least favorable / extreme. Judges are not being asked about their opinions of the statements, rather their task is to objectively order them from high to low. After sorting, tabulations of the judges' orders are calculated based on the number the degree of overlap in their orders. The tabulations and resulting statistics, at the time, were considered a very laborious process, which resulted in Likert's (1932) scaling technique taking over not long after Thurstone's publications (1928; 1929). Thurstone's scaling is ultimately aimed at obtaining the location estimates for each item. Thus, a similar process could be used to estimate GGUM location estimates, for use in θ estimation or to aid the MML process. As already noted, this was suggested for future research Carter and colleagues (Carter et al., 2014). Extending Thurstone's technique, perhaps using subject matter experts' (SMEs) opinions of item location could yield relatively accurate location estimates when compared to MML estimation.

The utilization of SMEs in the scaling process is quite common. Smith and Kendall (1963) used SMEs to rate behavioral statements that then resulted in a performance appraisal form. Similarly, other behavioral rating scales used in the performance assessment domain have ubiquitously utilized SMEs to calibrate stimuli or statements using average SME ratings (Campbell, Dunnette, Arvey, & Hellervik, 1973; Griffin, Neal, & Neale, 2000; Hedge Borman, Bruskiewicz, & Bourne, 2004). Judgments from SMEs are also commonly used in the development of situational judgment tests (Lievens & Sacket, 2006). Finally, Borman and colleagues (Borman et al., 2001; Schneider, Goff, Anderson, & Borman, 2003; Darr, Borman, St-Pierre, Kubisiak, &

Grossman, 2017) utilized SMEs to rate the effectiveness of each statement used in the development of a computerized adaptive rating scale (CARS). These CARS were developed to assess employee performance and utilized an unfolding model and pairwise preference items. The average rating of each item's effectiveness was used as the item's location parameter for scoring the unfolding model that was used. Despite using an unfolding model other than the GGUM, Borman and colleagues' utilization of SME ratings of location estimates yields particular promise for applications to other IRT models.

SME Ratings of Location

Chernyshenko and colleagues (Chernyshenko et al., 2007) used SME ratings of location estimates within their investigation of constructing ideal point scales. Because one of the main contentions of dominance modeling is that moderate items are not useful, they had SMEs rate the location (δ) of the items used in their analyses to understand where items fell on the trait continuum. While not the focal investigation, they found that the correlation between SME estimates and MML estimates of δ were .89. Other researchers have suggested that future research be aimed at understanding how SME judgments of item locations align with locations derived from empirically estimated methods (Oswald & Schell, 2010). Additionally, they questioned how potential discrepancies in the alignment might affect the test development process or in the scoring of persons (Oswald & Schell, 2010).

Stark et al. (2011) set out to understand just how viable SME estimates of δ could be in lieu of statistical estimation. Their intent was to provide evidence that Borman and colleagues' (Borman et al., 2001; Schneider, Goff, Anderson, & Borman, 2003; Darr,

Borman, St-Pierre, Kubisiak, & Grossman, 2017) technique for estimating location parameters was sound. Stark et al. (2011) focused on parameters for an ideal point model used by Borman, known as the Zinnes-Griggs (Zinnes & Griggs, 1974). The Zinnes-Griggs is an unfolding model designed to utilize pairwise-preference items and only requires δ item parameters. Their second study, a simulation, intended to understand how this affected an adaptive, pairwise-preference test, but their first study provides a good amount of information about SME estimates of item location more generally. The first important takeaway from Stark et al.'s (2011) findings was that SME estimates of δ, for an order scale and a self-control scale, were correlated with MML estimates .83 and .62, respectively. While the authors note that the correlation for self-control was a bit lower, these were still rather high correlations for the social sciences. This lower correlation was attributed to sizeable differences in estimates for a few items (Stark et al., 2011). To understand how these discrepancies affected person-level estimation, Stark and colleagues estimated person parameters using MML estimates and then with the SME estimates of δ. Remarkably, they found that despite the discrepancies, *both* scales yielded θ estimates using SME ratings that were highly correlated with θ estimates using MML estimates, .97 for order and .93 for self-control (Stark et al., 2011). Additionally, they found that validity correlations between person-level estimates of order and self-control with two outcome variables were extremely similar, and not statistically significantly different, when using SME ratings versus MML estimation (Stark et al., 2011). At this point, they argue that this is sufficient evidence to support the use of SME estimation in replace of MML estimation of δ estimates. Furthermore, they note that doing so would

likely not have an adverse effect on decisions made in applied settings like employee selection (Stark et al, 2011).

Chernyshenko, Stark and colleagues' (Chernyshenko et al, 2007; Stark et al., 2011) findings are important for multiple reasons. First, this presents an overall framework for non-statistical estimation of unfolding IRT item parameters. Second, they used what they considered the minimum number of SMEs possible, two. Coupled with their impressive results, it is possible that if more SMEs were used, individual discrepancies in ratings would have less of an impact. Similar δ estimates would likely lead to more highly correlated θ estimates across SME and MML estimation. On a related note, Stark et al.'s (2011) investigation focused solely on the Zinnes-Griggs unfolding model and pairwise-preference items. Despite this, a third reason for these studies' importance is the ability to generalize to other unfolding models and other item types. Because the Zinnes-Griggs only requires item location estimates, this technique and the findings can be tested within other frameworks. Interestingly, Stark et al. (2011) suggest as much in their discussion of future research.

Finally, as previously mentioned, Carter and his colleagues (Carter et al., 2014) suggest that future research is needed to investigate the extent to which Thurstonian scoring (Thurstone, 1928; Thurstone, 1929) could be used to calibrate item locations. The previous research examined in this section does not refer to the process of using SME ratings of location as Thurstonian scaling. Despite this, there is a number of different aspects adopted from Thurstone in the previous research. So much so, that one could argue that these earlier investigations provided initial evidence to support Carter et al.'s (2014) suggestion. An important next step is to take the results found from the previous

studies (Chernyshenko et al., 2007; Stark et al., 2011) and extend it to answer Carter et al.'s (2014) question about its effect on sample size. This is the aim of the current study.

The Current Study

Previous literature has argued that the majority of psychological research fails to utilize more modern approaches to psychological measurement and could benefit from such applications. Modern measurement models, such as the GGUM, hold great promise for researchers who require the measurement of constructs such as personality, attitudes, or interests. These could be argued to be a large majority of psychological researchers. A plausible explanation for the lack of utilization of these modern methods is the commonly unfeasible sample sizes required to obtain stable item parameter estimates. And finally, applied psychological researchers have shown that SME ratings of location parameters are highly correlated to estimates obtained from MML. Merging these two streams of literature together may provide evidence that psychological researchers can use SME estimates of item locations to reduce the burden of sample size when developing psychological measures using modern measurement methods, like the GGUM.

Previous research has shown that SMEs are fairly effective at rating where items are located on the underlying trait continuum (Chernyshenko et al., 2007; Stark et al., 2011). More importantly this research was conducted using only two SMEs. Thus, collecting responses from more than two SMEs should yield at least similar results. To confirm this, average SME ratings of δ estimates for individual items were compared to δ estimates obtained when using the GGUM's MML estimation algorithm.

H1a: SME estimates of δ are strongly aligned with δ estimates obtained from MML.

An important aim of this research was to provide psychological researchers more feasible modern, measurement approaches. With this in mind, one potential concern was the availability of SMEs. Neither of the investigations previously discussed provided insight into how SMEs were selected or any criteria that were used to do so. If one were to label only those who had expertise or a strong background in unfolding models as an SME, relatively few people would be available to serve in that role. If the criteria were lessened somewhat to only those who had a strong background in IRT, the available SMEs would increase. However, not all psychological researchers are required to learn IRT and fewer still use it on a day-to-day basis (Foster et al., 2017). Thus, this criterion may still be unfeasible to the everyday psychological researcher. To circumvent this issue and provide researchers with a more accessible process to unfolding scale development, training novices was considered a potential alternative. Applying best-practice techniques from the performance assessment domain could provide unfolding scale developers an empirically tested approach to training novice raters. This aim of this training was to ensure raters (1) understand the concepts, constructs, and models and (2) make just as accurate ratings as ideal point experts (see the Methods section for a more detailed description of the development of the training and Appendix B for the training materials).

H1b: Trainee estimates of δ are strongly aligned with δ estimates obtained from MML.

Stark et al. (2011) found that despite moderate differences between SME and MML δ estimates, using either in the estimation of person parameters yielded extremely similar person estimates. Thus, it was expected that a similar pattern emerges for the GGUM. Similarly, because it was expected that training novices will lead to similar

estimates as SMEs, a similar pattern is expected when using the δ estimates obtained from trainees.

One important point requiring discussion here, is the fact the GGUM requires the estimation of a number of item parameters other than δ. Recall from Equation 1, it requires the estimation of both a discrimination parameter (α) and at least one threshold parameter (τ) when using a dichotomous response format (for Likert-type scales, the number of threshold parameters increases to *C*-1, with *C* being the number of response options used). If this is the case, multiple questions arise. First, why are the location parameters the only thing being estimated by the SMEs or trainees? Second, how would the θ estimation process, EAP, be carried out if the other parameters are not being rated by SMEs or trainees? Both questions are important and require a detailed discussion. The first question has multiple answers. From a general ratings perspective, most if not all people would find rating the discrimination and threshold parameters of unfolding items extremely difficult. This would be especially true for IRT and psychometric novices. Even for the most experienced ideal point researchers, accurately estimating an item's threshold parameter would likely be considered a fool's errand. Thus, it seems important to first understand how accurate these two groups are at estimating item locations. The remaining answers focus on the GGUM itself. First, the location parameter is, without a doubt, the most important parameter for person estimation. As you begin to whittle down the parameters used in the model, the only one that is absolutely necessary for θ estimation is the δ parameter. The key to the unfolding process is the relative locations of θ and δ. The α and τ parameters are incidental in that they describe the degree of agreement (τ) and the distinction between the degrees of disagreement (α). Finally,

another important answer to the first question is that the α and τ parameters are relatively difficult to estimate compared to δ even when using MML with adequate sample sizes (*N* > 750; Roberts et al., 2001). Taken together, these answers make the case that SME estimation of one or both parameters may be not be fruitful.

To answer the second question – how one would carry out the person estimation process without SME or trainee estimates of other parameters – it is important to first discuss the MML estimation process more generally. The MML process requires a large calibration sample to effectively "zero-in" on item parameters to use in person estimation. The item parameter estimation process essentially starts with arbitrary values and uses what are referred to as "burn-in" iterations to obtain a useful foothold for the proper estimation to begin. However, the MML process is also flexible in that it allows for specific parameters to be used as the starting points for the MML process. If this approach, using rater estimates as δ parameter starting points, is effective, it should result in relatively similar parameter estimates as those obtained when no starting parameters are included. Thus, using the SME and trainee estimates of δ parameters as the starting parameters in the MML process should yield relatively similar θ estimates, especially if hypotheses 1a and 1b are supported.

H2a: Person estimates (θ) obtained using SME estimates of δ as starting parameters are strongly aligned with θ estimates obtained using no starting parameters.

H2b: Person estimates (θ) obtained using trainee estimates of δ as starting parameters are strongly aligned with θ estimates obtained using no starting parameters.

Previous research has shown that SME estimates of item locations can be used in person location estimation in other unfolding models (Chernyshenko et al., 2007; Stark et al., 2011). Finding support for the previous two series of hypotheses suggests that these findings can be generalized to the GGUM. The important extension of these findings is the effect that SME or trainee estimates of δ have on parameter estimation as sample size decreases. As previously discussed, researchers have argued that sample size is one of the main hinderances to the utilization of IRT (Borsboom, 2006) and the GGUM, specifically (Carter et al., 2014; Foster et al., 2017). The large sample size requirements are necessary to achieve stable item estimates (Embretson & Reise, 2000). Based on previous hypotheses, if SMEs or trainees can be used to provide useful δ starting parameters, it should allow the MML process to effectively estimate item parameters with a smaller sample. This would make sample size requirements less intensive to researchers. Thus, a number of decreasing sample sizes were used to estimate both item parameters (δ, α, τ) and person parameters (θ) using both SME and trainee δ estimates as starting parameters in the MML estimation process. These estimates will then be compared to the item and person estimates obtained from the total sample using the standard MML estimation process (i.e., no starting parameters).

RQ1a: To what extent do parameter estimates (δ, θ) obtained using SME and trainee estimates of δ as starting parameters and a sample size of 750 align with parameter estimates obtained using no starting parameters and the total sample?

RQ1b: To what extent do parameter estimates (δ, θ) obtained using SME and trainee estimates of δ as starting parameters and a sample size of 500 align with parameter estimates obtained using no starting parameters and the total sample?

RQ1c: To what extent do parameter estimates (δ, θ) obtained using SME and trainee estimates of δ as starting parameters and a sample size of 200 align with parameter estimates obtained using no starting parameters and the total sample?

RQ1d: To what extent do parameter estimates (δ, θ) obtained using SME and trainee estimates of δ as starting parameters and a sample size of 100 align with parameter estimates obtained using no starting parameters and the total sample?

Finally, while location estimates are the most important parameter in the GGUM estimation, it was previously stated that there are other parameters that are estimated by MML. To this day, no research has investigated the extent to which it is feasible to have SMEs make ratings about the discriminability, from an IRT perspective, of an item. It stands to reason that accurately rating the discrimination of an item likely requires much greater IRT and GGUM expertise, but if SMEs can provide estimates that are more precise starting parameters than random ones generated by the MML process, it may yield even better results. This was considered a first attempt at SME estimations of IRT discrimination parameters, thus no direct prediction was made about effectiveness. This line of investigation served as a first attempt to understand whether discrimination estimates are worthy of similar research seen for IRT location parameters (Stark et al., 2011).

RQ2: To what extent do person estimates (θ) obtained using SME estimates of δ and α as starting parameters align with θ estimates obtained using no starting parameters.

Method

The overarching goal of the present study was to empirically investigate the extent to which SME and trainee ratings of item locations affect person parameter estimation across various sample sizes. To do this, data was collected from three different sources. The SMEs and trainees completed similar rating tasks. The only difference between the two was the training – provided to trainees directly before the rating task – and the attention check and quality assurance items. The SMEs received no training or attention check items. Both groups were provided with trait definitions and descriptions of both high and low statements aligned with -3 and +3 on the rating scale, as they made their ratings. The third data collection collected actual responses to the items that were rated by the SMEs and trainees.

Participants

Subject Matter Experts. A total of 14 SMEs were contacted and asked to participate in the rating task. A total of nine completed the rating task, for a response rate of 64%, and were used in subsequent analyses. Those nine SMEs self-reported an average of five years of experience with unfolding IRT models. Only four of the nine SMEs considered themselves to be an expert in unfolding IRT models, but all participants had at least two years of experience working directly with these types of models. Thus, all participants were considered sufficient experts by this criterion.

Trainees. A total of 45 students were recruited from a university subject pool across four different class sessions. Each participant was compensated with extra credit points allotted by the class instructor. College students are likely the most convenient and accessible population for psychological researchers. However, students were considered

novices with no statistics or research methods background. Thus, it made sense to investigate the extent to which this sample can be trained and used to make item location ratings. Of the 45 that were recruited, they were predominantly female (78%) and white (76%). No respondents felt that they were not at all successful at the rating task and most felt that they were at least moderately successful or better (96%). Trainees were also asked to complete a series of attention check and quality assurance items. Four trainees failed to successfully complete the attention check items and / or appropriately responded to the quality assurance items. These four were removed and excluded from the analyses.

Item Responses. Participants that provided item responses were recruited via Amazon's Mechanical Turk (Mturk) service (https://www.mturk.com/mturk/). Because of the large sample size required for GGUM estimation, a total of 1,211 responses to the 40 items were recruited. While only 40 items were administered, two attention check and two quality assurance items were also included to ensure adequate and attentive participation. The initial paragraphs of the Results section provide a breakdown of the number of participants that were excluded using these criteria. Each participant was compensated $1.25 for completing the survey. The recruited sample had an average age of 34 and 57% were male. The majority of the sample was white (60.8%), and the remainder of the sample was Asian (28.4%), Black or African American (9.2%), American Indian or Alaskan Native (3.6%), Native Hawaiian or Pacific Islander (0.7%).

Measures

All three data collection processes either rated or responded to the same 40 statements, which can be found in Appendix A. The 40 statements measured one of two subdimensions of personality. The two dimensions were based on one of two

subdimensions from the Big Five Aspects Scales (BFAS) developed by DeYoung, Quilty, and Peterson (2007). The BFAS divides each of the big five personality constructs into two lower-order traits or facets. The two dimensions that were used were lower-order traits associated with conscientiousness and emotional stability. These two big five factors were selected because of their importance in Industrial-Organizational Psychology. The items were developed as a part of a propriety CAT used in high-stakes employment settings. The CAT was developed and is administered using the GGUM and is administered in a forced-choice pairwise preference format. Tests using a pairwise preference format with the GGUM require that item parameters be estimated using a single-statement, unidimensional approach (Stark, 2002) which can be done with the GGUM2004 program. Thus, items were developed and parameters estimated in the same format as used in this study. Although each dimension had an itembank of hundreds of items, only 20 items were selected for each. The items were selected to represent the full range of the trait continuum. The δ estimates used to ensure adequate representation were based on the original development of the CAT, which used Likert-type response options to derive initial item parameters. For the item responses, a 7-point Likert-type response scale was used ranging from *Strongly Disagree* to *Strongly Agree*. Other than the inclusion of the training materials described below, the rating process for the SMEs and the trainees was identical. To make the rating process as easy as possible, a rating scale of -3 to +3 was used. This was done because it provides a clear middle point (i.e., zero) and is symmetrical on both sides. Respondents were only able to select from -3 to +3 in 0.5 increments.

Drive. The personality dimension Drive is the lower-order factor of the big five dimension conscientiousness. It reflects the extent to which someone is proactive and persistent. Those who score high tend to be reliable, hardworking, and accountable. Those who score low tend to be reactive and less deadline-oriented. Based on the original development of the CAT, the location estimates sufficiently span the trait continuum (range$_{\delta}$ = -2.92 to 2.87; M_{δ} = .40). An example item is "*Even though it can be exhausting, I always deal with issues as soon as they come up*."

Positivity. The personality dimension Positivity is the lower-order factor of the big five dimension emotional stability (or neuroticism). It reflects the extent to which someone is happy, optimistic, and resilient. High scorers tend to be hopeful and positive. Lower scorers tend to be pessimistic and overwhelmed with obstacles. The location estimates sufficiently span the trait continuum (range$_{\delta}$ = -2.93 to 3.18; M_{δ} = 0.29). An example item is "*It is difficult to be cheerful when there are many problems to take care of*."

Training Development

The development of the training materials was based on empirical research of rater training within the performance rating domain. This is one of the typical strategies used to enhance rater effectiveness (Woehr & Huffcutt, 1994). While the focal point of the training differs from that of the performance appraisal, the empirical findings suggest a similar level of effectiveness might be uncovered for making location ratings. The training was designed to align with the relevant procedures outlined by Pulakos (1986) for Frame-of-Reference Training (FORT). The overarching aim of FORT is to match ratee behaviors to their appropriate performance dimensions and correctly judging the

effectiveness of those behaviors (Sulsky & Day, 1992). The extension of FORT to the present training has a similar aim to ensure trainees correctly understood the traits and appropriately identified item locations for those traits.

The training involved a series of presentation slides that were developed as well as an oral discussion of the same material. The slides (shown in Appendix B) introduced the rating task and an explanation of personality testing. After this, an explanation of the trait continuum on which they made their ratings was provided. This was followed by definitions of the two personality traits and explanations of typical behaviors of high and low scorers for each of each these traits. And finally, two example items and their associated item locations as well as a complete explanation as to why each item has this specific location was provided.

The presentations slides were inserted into the survey platform and were available to participants during the oral walkthrough of the slides as well as during the rating task. The oral presentation was a word-for-word vocalization of the slides. This was done to ensure consistency across data collection efforts. However, participants were invited to ask questions throughout the training and again after the completion of the training. Once the training slides had been completed, participants completed two practice ratings. These practice ratings allowed participants to make ratings. Depending on the ratings made, participants were either progressed to the next statement or they were provided further context to understand the statement's location and asked to re-rate the statement. Once the participant successfully rated where the statement was located, they moved on to two training-related attention check items.

Respondent Sampling & Parameter Recovery

Sampling of Respondents. One of the main investigations of the proposed research requires response data of varying sample sizes to be compared with "true theta" levels. This process resembles simulation methodologies known as parameter recovery which are common in IRT algorithm or model development (Feinberg & Rubright, 2016). Conventional IRT simulations are conducted by first randomly generating "true" theta levels, and then creating response data based on those true thetas to run through the model or algorithm a number of times. The theta levels that were estimated based on these algorithms are then compared to the true, generated thetas to evaluate how well the algorithm is functioning. Rather than conduct a simulation study, where the theta levels and response options are completely fabricated, a simulation-like study was carried out where the responses and "true" theta levels were real, but subsamples of varying sample sizes will be drawn from the total sample. In other words, samples of 750, 500, 200, and 100 respondents of the item response data collection were randomly drawn (without replacement) from the total sample collected within the item response data collection process. This process is similar to a study conducted by Sahin and Anul (2016) to investigate the effects of test length and sample size on item parameters in IRT. There were a total of 8 research conditions that were evaluated for both personality dimensions (sample size [750, 500, 200, 100] x location rating [SME, trainee]).

Parameter Recovery Criteria. To understand the extent to which SME and trainee ratings used as starting parameters in the MML process lead to a smaller sample size requirement, a comparison of the θ estimates with true θ parameters was required. The description of "true θ parameters" is really only true in simulation studies where they are randomly generated and fictional. Since true parameters are never really known

otherwise, θ estimates will be obtained using the total sample and no starting parameters. Based on research during the development of the GGUM (Roberts et al., 1999), samples of 750 to 1,000 should yield relatively stable item estimates without the requirement of any kind of starting parameters, which should lead to relatively accurate θ estimates. Thus, the θ parameters obtained using a samples of 818 and 821 (Positivity and Drive, respectively) and no starting parameters will be referred to as true θ estimates for the remainder of the manuscript. These true θ parameters were used to evaluate the estimated θ parameters. The estimated parameters refer to any θ parameter that is estimated using either SME or trainee δ estimates as starting parameters in the MML process or with any of the subsamples (i.e., 750, 500, 200, 100).

To estimate the accuracy of parameter estimation using the two rating sources, multiple evaluation metrics were calculated. It is important to note, again, that these metrics were calculated between the estimated θ parameters and the true θ parameters. As an example, say a person's responses to the 20 positivity items resulted in a true θ of 2.5. Then say this same person was also randomly selected to be included in the subsample of 750. Using the entire subsample of 750 participants, the item parameter estimation process, MML, would be carried out again with the SME and trainee δ estimates as starting parameters. The resulting item parameters would be used in the EAP process to estimate θ for each person in the 750 subsample. From this, say the same person's θ estimate when using the SME estimates as starting parameters resulted in a θ of 2.4 and when using trainee estimates resulted in a 2.8. These two estimated θ parameters will be compared against their true θ parameter which was estimated with the total sample and no starting parameters. This process was conducted for all subsamples and is the typical

process known as theta (or parameter) recovery in simulation studies. Pearson product-moment correlations will be one metric used to evaluate parameter recovery. The following equations describe the other two metrics, bias (Equation 2) and RMSE (Equation 3), that will be used.

$$\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_{True})}{n}, (2)$$

$$\sqrt{\frac{\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_{True}\right)^2}{n-1}}, (3)$$

Bias provides a measure of the average distance between the estimated and true parameter (Feinberg & Rubright, 2016). Positive bias occurs when the estimated parameters are, on average, greater than their corresponding true parameters and negative bias occurs when the estimated parameters are, on average, less than the corresponding true parameters. Equation 3, RMSE, is a measure of dispersion and the square root of the mean squared error which measures the squares of the deviations between the estimated and true parameters (Feinberg & Rubright, 2016). These are three of the most popular metrics to evaluate parameter recovery in simulation studies (Feinberg & Rubright, 2016). While the most pivotal investigations pertain to how these metrics change as sample size decreases, two of the criteria have specific rules-of-thumb for interpretation. Correlations between estimated and true parameters of $r \geq .70$ (Yoes, 1995; Field, 2013) and RMSE $\leq .33$ (Rudner, 1993) have been argued to be acceptable metrics. An estimate of bias $< |.20|$ will be used, in addition to rules-of-thumb for correlations and RMSE, as cut-off points to achieve support for the hypotheses stated above.

Results

Prior to any analytical investigations, data cleaning was performed on all three datasets that resulted from the three data collection efforts. Regarding the item response data collection, data was removed for all participants who completed the entire survey in less than two minutes. To complete just the 40 items – not including the informed consent or demographic items – in less than two minutes would require responding to each item in under 3 seconds on average. Since it is unlikely that one could provide quality data at that speed, these participants were removed. This resulted in the removal of 261 participants.

Due to potential issues related to "bots" completing surveys and measures that have been identified for crowdsourced based survey response collection like Mturk, the bot.detector R package (Prims & Motyl, 2018) was employed. This function creates a score for each response in a Qualtrics dataset that is intended to count the number of features that are typically associated with bots and / or survey farmers (e.g., longitude and latitude, timing, type of comments). Using this function, no bots were detected thus no cases were removed due to this investigation.

Next, each participant's responses were divided based on the two traits that were included in the survey (Drive and Positivity) and additional data cleaning was carried out. Because the parameter estimation of each scale is independent of the other, each scale was cleaned independently. Rather than removing anyone who had any missing data or failed attention checks for either trait, removing based on the individual traits meant that in cases where, for example, someone only had missing data for one trait and complete data for the other, the data for the complete trait was able to be retained. For the

Positivity data, 59 cases were removed due to failing the attention checks. Thirty-six cases were removed due to missing data. Finally, multiple indices were calculated to assess and identify possible insufficient effort responding (i.e., lack of variance, lack of responses used, longstring, and intra-individual response variance). Using these metrics, an additional 37 cases were removed which resulted in a total, cleaned sample of 818 for the Positivity data. Taking the same process with the Drive data resulted in the removal of 129 cases for a total, cleaned sample of 821 cases.

Data cleaning was completed for the trainee data collection, which focused on whether or not participants successfully responded to the attention checks. Only four cases were removed for a total, cleaned sample of 41 trainee cases. Finally, the SME data were reviewed and no participants were removed due to failing attention checks or incomplete data.

Hypotheses 1a and 1b

For all GGUM estimations, the mirt package (Chalmers, 2019) from the R environment (R Core Team, 2016) was utilized. To test the hypotheses that SME (H1a) and trainee (H1b) estimates of item location (δ) will strongly align with δ estimates obtained from MML, the correlation, bias, and RMSE of the SME and trainee estimates of δ compared to the true δ parameters were carried out for both personality traits, Positivity and Drive. Table 1 shows the average location ratings made by SMEs and trainees for each of the 20 items per personality dimension and the "true" δ estimates obtained from MML estimation. The results of the comparisons between the estimates shown in Table 1 are provided in Table 2. The results for neither the trainee nor SME estimates met the thresholds set that would suggest support for either Hypothesis 1a or

1b. However, in line with expectations, the SME δ estimates were more closely aligned with true δ estimates across most parameter recovery metrics for both personality traits. These results suggest that effectively estimating actual IRT parameters may be more difficult than previously thought (Stark et al., 2011). Additionally, Table 2 reveals that the results for Positivity yielded higher correlations than Drive, but also higher bias and RMSE. This is in contrast to expectation in that the higher the correlation observed, the lower the bias and RMSE one would typically expect. Bias and RMSE are measures of the difference between point estimates whereas correlation represents the relationship between the estimates. The finding suggests that while both SMEs and trainees were able to effectively order the items' locations, they had difficulty pinpointing where the items actually fell on the trait continuum. Spearman's rho correlation was also included in Table 2 to better understand the extent to which SMEs and trainees were effectively able to rank-order the item locations. Taking all of these results together, it appears that it was easier to rank-order the items in terms of their location / extremity but was harder to pinpoint the actual location estimate individually.

An additional point is worth mentioning here. The obtained estimates from the item parameter estimation resulted in a few items for each scale that had overly extreme item locations. Six items in the Positivity scale and one in the Drive scale were estimated to have item locations greater than |3|. A potential explanation for this is based on the traits measured by these scales, specifically the Positivity trait. The MML method used for estimating GGUM parameters has been shown to have difficulty estimating δ for items with extreme item locations (i.e., larger than |2.5|) when a small sample is available at that end of the trait continuum (de la Torre, Stark, & Chernyshenko, 2006). In other

words, since only a few people are likely to have extreme-low θ for the trait Positivity, it was difficult to estimate accurate item parameters for items with low item locations. These observed item locations for these items fell outside the range of possible item location ratings available to SMEs and trainees. Thus, it was not unexpected to find that both sets of raters were relatively unsuccessful in their alignment with the estimated locations. Post hoc analyses were conducted where only the items with MML estimated locations within the range of the possible ratings (-3 to +3) were used in analyses carried out in the exact same manner as above. The results for the Positivity scale revealed that the bias (SME = .11, trainee = .75) and RMSE (SME = .70, trainee = 1.12) were much lower when these items were removed (see Table 2b). Interestingly however, the correlation was also lower when these items were removed. This suggests that despite the SME and trainee location ratings being closer in proximity to the actual estimated locations, the removal of those extreme items negatively affected the rank-order relationship between the MML estimates and the SME and trainee ratings. A similar trend was found for the Drive scale except correlations for both SME and trainee slightly improved with the removal of the one item. This suggests that while the MML procedure did result in highly extreme location estimates, those estimates were indeed rated as extreme by SMEs and trainees and the removal of those items negatively affected the correlations between SMEs/trainees ratings and the MML estimates. In any case, the results of these post-hoc analyses still failed to meet the criteria established for support of Hypothesis 1a and 1b.

Hypotheses 2a and 2b

Despite the lack of support for either SMEs' or trainees' ability to accurately identify item location estimates, the use of the SME- and trainee-based estimates of δ may still result in fairly well-aligning person parameters (θ). To better understand how these estimated starting parameters affected θ estimation, the next series of hypotheses were tested. Table 3 shows the descriptive statistics and correlations of all θ estimates obtained using the total samples. The bold correlations indicate intra-trait inter-estimation relationships. This reveals that for Positivity, both SME- and trainee-based θ estimates were highly correlated with each other and with the true Positivity θ estimates. For Drive, the SME-based θ estimates were highly correlated with the true estimates and this correlation was similar to those between the Positivity estimates. The trainee-based Drive estimates, however, seem to be fairly different from the other two Drive estimates, with the correlations in the .70's rather than the high .90's as is seen with the other correlations.

In line with previous research (Stark et al., 2011), recovery metrics for SME-based θ estimates, shown in Table 4, reveal Hypothesis 2a, which posited that SME ratings of location parameters used as starting parameters in the MML process would yield relatively similar θ estimates, was supported for both Positivity and Drive traits. The results for Hypothesis 2b were not as clear. While all of the metrics for the Positivity trait suggest support for Hypothesis 2b, the RMSE metric for trainee-Drive was outside of the acceptable threshold (and the correlation was also at the lower end of the acceptable range). The RMSE metric is the square root of the average of squared errors. Thus, larger differences have a disproportional effect on RMSE. This likely explains why despite relatively little bias, there was significant RSME. In other words, trainee-based

Drive θ estimates were well aligned with true θ estimates for a lot of the cases, but certain exaggerated outliers disproportionately affected the RMSE estimate.

Research Questions 1a through 1d

The main investigation of this research was to understand to what extent SME and trainee ratings of item locations used as starting parameters in the MML process affected both δ and θ estimation as sample size decreased. Specifically, Research Questions 1a through 1d questioned how aligned recovered parameters would be with true parameters, when using SME- and trainee-based starting location parameters at various sample sizes (i.e., 750, 500, 200, 100). Because of the novelty of this particular investigation, no specific hypotheses were provided. However, to assess if the results suggest a worthwhile finding, the thresholds used to evaluate Hypotheses 1 and 2 also served as indicators here. See Table 5 and Table 6 for results. The ultimate hope was that as sample size decreases, the alignment between the true parameters (δ, θ) and the estimates obtained using SME and trainee locations as starting parameters remained somewhat consistent. Given this, and in line with simulation research, graphical representation of the evaluation metrics across sample sizes was also reviewed (see Figures 4 and 5). One note is that when using the trainee-Drive ratings as starting parameters for the sample of 100, the GGUM model failed to successfully run. Therefore, no analyses could be conducted, and no results could be reviewed for this series. It is worth noting here that the SME-Drive ratings were successfully ran with the 100 subsample. This provides further evidence that expertise with unfolding models is helpful requirement for making item location ratings, especially with very low item calibration sample sizes.

As shown in Table 5, bias and RMSE estimates across subsamples for both Positivity and Drive failed to meet the appropriate thresholds for SMEs and trainee estimates of location. This is further evidenced by Figure 4 which not only shows that the aforementioned metrics failed to meet appropriate thresholds, but they also dramatically shift for subsamples 200 and 100. This aligns with the research and findings from earlier simulation studies aimed at understanding sample size requirements for GGUM estimation (Roberts et al., 2002).

Somewhat aligned with the results examined for Hypothesis 1a and 1b, for Positivity, correlation metrics for both SME and trainee location estimates were above the acceptable threshold (Figure 4, Row A). This suggests that when using either SME or trainee location ratings as starting parameters in the MML process, the MML process returns δ estimates that are similarly aligned as true δ estimates. This result was found despite the fact the location ratings are fairly different across rated versus true sources – as evidenced by large bias and RMSE metrics in Table 2. Interestingly, this pattern was not observed for Drive (Figure 4). This provides additional support for idea that the constructs being measured and rated by either SMEs or trainees, plays an important part in how effective either group could be when making ratings about item location.

The next investigation was to what extent person parameter estimates obtained using SME and trainee location estimates as starting parameters aligned with true person parameters (θ). As shown in Table 6, for both Positivity and Drive across subsamples 750 and 500, only one scenario failed to meet acceptable thresholds across metrics. This was the Positivity-Trainee-750 sample (RMSE > .33). However, the 200 and 100 subsamples failed to meet the RMSE threshold for all scenarios, despite reaching appropriate

thresholds for both bias and correlations (see Figure 5). This is similar to the findings uncovered for Hypothesis 2b. Like those findings, this can likely be explained by substantial differences in a small number of cases that has a larger impact on RMSE estimates compared to bias and correlation. Further, it appears that as sample size decreased, the more impact the few significant theta estimation errors had on RMSE. Importantly, these findings suggest that despite somewhat substantial errors observed in the SME and trainee location estimates, it did not seem to have a huge impact on the person estimates obtained from the EAP process. Stark and colleagues (2011) found a similar pattern using the Zinnes-Griggs unfolding model and only SME estimates of location.

Research Question 2

To investigate Research Question 2, which questioned to what extent would recovered parameters align with true parameters using SME-based location *and* discrimination parameters as starting parameters, the same exact procedure was carried out as for Hypothesis 2a (see Table 7). The only difference with the analyses for Hypothesis 2a was that average discrimination ratings estimated by SMEs were introduced in the model as starting parameters for the alpha item parameter. This, again, was conducted for both Drive and Positivity scales. Results suggest that θ estimates obtained when using both δ and α as starting parameters are essentially identical to those obtained with just SME-based δ parameters. Results are shown in Table 8 and when rounded yield the same recovery metrics as those shown in Table 4 for SME-based estimates. However, the model that included starting parameters for α from SMEs required more than 80 additional iterations (257) to successfully converge compared to

the estimation that only included δ SME starting parameters (175). This suggests that while similar θ estimates were obtained, the SME-based α parameters could be argued to have hindered the estimation process because the same results were obtained with significantly more iterations required to reach the same estimates.

Post-Hoc Investigation

Because of the data gathered and tested to investigate Research Question 2, data were available to also perform a similar investigation as Research Questions 1a through 1d using both location ratings and discrimination ratings made by SMEs. These analyses essentially combine Research Questions 1a through 1d with Research Question 2 to investigate the extent to which SME ratings of item locations *and* item discrimination used as starting parameters in the MML process affected both δ and θ estimation as sample size decreased.

To empirically test this, the same procedure was carried out as was used to test Research Questions 1a through 1d. The only difference was that discrimination ratings made by SMEs were used for each item within the GGUM as α starting parameters in addition to location ratings.

Results from the subsample analyses regarding the alignment of δ parameter estimates when using SME ratings of both item location and item discrimination as starting parameters with true δ parameter estimates are provided in Table 9. In general, the results across both traits and subsamples fail to provide support for the utility of the collection and implementation of discrimination parameter ratings. This is clearly evident when comparing results presented in Table 5 with those in Table 9. Including discrimination starting parameters in during model estimation negatively affected the

recovery metrics obtained in some cases. For example, when only location starting parameters were used, bias for the Drive-500 subsample increased from 0.75 to 1.60 when discrimination starting parameters were also included. While in some cases the results were negatively impacted, a good deal of the obtained results were either essentially the same or exactly the same across the two series of estimations. This provides additional support for the lack of utility of item discrimination parameters as starting parameters from an item parameter perspective. Because of the relative similarity across the two series of item parameter estimations, little differences were expected at the person parameter estimation level. But for consistency, these results were also reviewed.

Table 10 provides the recovery metrics regarding the alignment between θ estimates obtained when using the item parameters that resulted from using SME ratings of both item location and discrimination as starting parameters and the true θ estimates. As expected given the results of the location recovery metrics, the results were either identical or essentially identical across both traits and all subsamples. The average difference between the recovery metrics for the two series of estimations was zero or very close to zero (bias = 0.000, RMSE = 0.035, r = -0.019) and provides further support for the notion that SME estimated discrimination starting parameters provide little to no value in θ estimation compared to just location starting parameters.

Discussion

Psychometricians continue to research and refine our understanding of ideal point models and the methods by which we implement them. Unfortunately, this literature has yet to have a noticeable impact on research methods conducted by more general psychological researchers. This study set out to investigate a potential avenue to enhance

the feasibility of these models and the sample sizes required for effective parameter estimation. While the results of the SME and trainee ratings as starting parameters yielded mixed results, this study yields new insights around the feasibility of these techniques. Additionally, the results of the study suggest that despite findings that in some ways run counter to previous research (Stark et al., 2011), the process outlined by Stark and his colleagues is both practical and effective in certain circumstances.

The most interesting findings relate to the comparisons of the SMEs and trainees ratings and of item locations as well as resulting θ estimates when those ratings were used as starting parameters. Considering the location ratings specifically, SMEs ratings tended to yield smaller bias and RMSE estimates, yet smaller correlations with MML-based location estimates. Additionally, for the 100 sample of trainees, the Drive data was unable to successfully run, where the SMEs had no problem running. While this could potentially be due to the random subset of responses drawn for this particular subsample, this finding aligns with a majority of the results in providing evidence that expertise or, at the very least, experience with IRT and unfolding models has a positive effect on parameter ratings. Despite larger bias and RMSE findings, trainees' ratings as starting parameters actually led to larger correlations with MML-based locations. One potential explanation for this finding can be described in the way the two groups cognitively worked through the problem of rating an item or the entire item set. Because SMEs understand trait continuums from an IRT perspective, they likely focused on where each item was located irrespective of the other items. Conversely, trainees may have focused more on ensuring that the items, as a set, were in the correct rank-order. While no theory-based research has looked into this type of cognitive interpretation, it may be a

worthwhile investigation for future research. Despite all of this, these two sets of criteria were quite close across SME and trainee ratings. So while the differences are interesting, the finding that they were, for the most part, aligned suggests trainees may be a sufficient substitute to SMEs to gather item location ratings.

The previously noted findings suggest that with the population used as novices in this study – undergraduate university students – similar ratings of item location can be obtained to ratings made by more experienced IRT or unfolding researchers. These findings indicate that with little training, undergraduate students can provide well-aligned ratings of item location that can, in some cases, suffice as starting parameters. Thus, this particular population provides psychological researchers with a more feasible alternative to obtaining item location ratings for GGUM starting parameters or other more creative uses.

A second, and more positive finding is that despite not finding support for SMEs ability to effectively locate item locations, using their ratings as starting parameters in the MML process lead to relatively well-aligned person-level θ estimates. This aligns with findings from Stark et al. (2011) that suggested that accurate person estimates could be obtained from an unfolding model despite differences in item location when compared to MML estimates. This is important because it aligns with their results using the Zinnes-Griggs unfolding model, but it also extends these findings to the GGUM. Stark and colleagues (2011) argued that, “The most important finding is that despite some obvious differences between the respective location values, the resulting trait scores correlated highly.” A similar conclusion can be drawn for SMEs here. Concerning the trainee’s ratings as starting parameters, θ estimates were well aligned according two of the three

criteria that was evaluated (bias and correlation) for Drive and all three for Positivity. This suggests that, similar to SMEs, trainees' estimates can be considered reasonable approximations for starting parameters in the MML process.

Another interesting finding pertains to the sample size investigation. In their development of the GGUM, Roberts and colleagues (2002) found that 750 to 1,000 participants were needed to effectively estimate stable item parameters to use for person parameter estimation. Unfortunately, after completing the data cleaning procedures described above, the total sample size for both scales were less than 1,000. Thus, analyses of the total sample were conducted on the remaining sample. While both exceeded the lower end of the sample size suggested by Roberts et al. (2002), it is worth noting that had the sample size exceeded the 1,000 threshold, different results may have been obtained.

The findings of the sample size investigation described above suggest that both SME and trainee estimates as starting parameters for samples of 750 and 500 yield relatively well-aligned theta estimates as the MML process with no starting parameters. Additionally, the analyses on samples of 750 and 500 resulted in little change in alignment with true estimates, particularly for person estimates. This provides initial evidence that ratings of item locations used as starting parameters can be used with smaller sample sizes than Roberts et al. (2011) originally found. While this is positive, because this trend failed to continue when using sample sizes of 200 and 100, it does not increase the feasibility of the GGUM for psychological researchers. Additionally, the MCMC GGUM process found that it only required 500 participants to effectively estimate item parameters. The MML program was utilized over the MCMC because the

MML was the first unfolding estimation available and still remains the most popular GGUM estimation technique today. Additionally, the wealth of research and investigation of unfolding IRT models using the MML far outweighs any other GGUM estimation program. For these reasons, the MML seemed to be the most appropriate starting place for this line of research. Future research should consider a similar study using a Bayesian approach like MCMC (MCMC GGUM).

Finally, it was found that including SME rated item discrimination parameters provided little to no value in the estimation of either item locations or person trait levels. As explained earlier, in the GGUM, the most important parameter is the item location, δ, and α and τ are essentially by-products of the distances between a person's location and δ. Thus, it is unsurprising that SME discrimination ratings had little effect on estimation. While this investigation failed to find that discrimination ratings affected results, an important takeaway is that psychologists developing measures using unfolding models like the GGUM would be well-advised to stick to collecting ratings of item location only. This directly impacts the feasibility for most psychological researchers because if training efficacy could be enhanced, novices could be used for rating item locations without researchers having to worry about training them on item discrimination as well.

Limitations

As with all research, this study had several limitations that should be noted. The greatest of these was the sample sizes of the total sample. While precautions were attempted to ensure at least 1,000 respondents for each scale, data cleaning led to the removal of a greater number of participants than was expected. This resulted in a total sample size lower than the noted threshold required for GGUM MML item parameter

estimation. While the total sample size was still within the lower bound threshold typically stated ($N = 750$), multiple research studies have found that samples of at least 1,000 lead to highly stable item parameter estimates.

A second limitation, which is somewhat related to the first, is the fact that a simulation-like study was conducted, rather than an actual simulation study. The study described in this manuscript only compares a single draw of differing sample sizes and compares the recovered item and person parameters against estimated parameters rather than actual, known parameters. While there are advantages and disadvantages to either procedure, additional techniques might provide more interpretable results (e.g., bootstrap).

Another limitation pertains to the scales or constructs used for the study. Conscientiousness and Emotional Stability have been shown to be important constructs in the area of Industrial-Organizational Psychology. However, additional scales measuring other Big 5 aspects or other types of personality constructs might have a provided a more complete understanding of how SMEs and trainees differ in their rating of different types of constructs or traits.

Future Research

The majority of this research attempted to lay the groundwork and set the direction for potential future research. As already stated, future research should consider a similar study using either simulation or related procedures, like bootstrapping. Taking the average parameter estimates observed over hundreds or thousands of replications may lead to more consistent results.

An additional direction for future research should focus on the constructs that allow for accurate SME and / or trainee ratings of item locations. The findings of this research suggest that Positivity was easier to rate for both SMEs and trainees compared to Drive items. This begs the question: how does the construct being rated affect the rater's ability to effectively rate the items. Using research like the general population's understanding of different constructs combined with decision-making or cognitive theory, future research may be able to better determine what constructs are easier or more difficult to rate from an item location perspective. One particularly interesting research paradigm might be to develop hypotheses about which constructs are easier or more difficult to rate across a wide variety. This could include personality domains at both higher- (i.e., Five Factor Model) and lower-order levels, as well as non-personality type constructs like attitudes (e.g., job satisfaction, militarism). Incorporating aspects of the constructs as well research from areas like cognitive decision-making to predict which constructs are easier to rate should provide the field with clear directions for constructs for which item location ratings are appropriate and those that should be avoided.

A related area of future research that would benefit this area would be to investigate to what extent the rater's level of expertise affect their ability to make accurate ratings of item characteristics. A simple study might focus on gathering data similar to the SME data collection utilized in this study but perhaps expanding the population to include anyone who has taken an IRT as a part of graduate studies. Then comparing results across various expertise demographics, for example: number of years working with IRT, level of interaction / use of unfolding models, perceived expertise levels and so on. This would aid future researchers in the exact utility of highly

experienced SMEs and balancing the trade-off that likely comes with only collecting data from such a small population.

In a similar vein, future research should look to investigate what leads to more proficient trainees. There is plenty of research opportunities available in investigating the type of information provided when training novices. This study provided high level information with the hopes that it would provide just enough information needed without over complicating the task. Future research would be well advised to consider how novices might best understand IRT and unfolding aspects without the need for advanced research methods or statistical training. A research design wherein participants are randomly assigned to one of various training designs would undoubtedly be helpful in guiding psychological researchers as they look to develop training that is effective, yet practical. A more specific, and potentially interesting investigation could focus on having novices first consider the rank order of items within a construct or scale. Considering the items as a whole scale, rather than individually should lead to stronger correlations with true locations and potentially even lower bias and RMSE.

Finally, a similar study using the MCMC GGUM program may yield more useful results. The MCMC GGUM program (Wang, 2014; Wang, de la Torre, & Drasgow, 2015) was found to have achieve stable item parameter estimates with a sample size of only 500 participants. Additionally, to estimate parameters, the MCMC GGUM requires users to provide the order of the items on the latent trait continuum. A non-perfect rank ordering of the items was revealed to lead to much more accurate estimates during simulation studies (Wang, 2014). This information would be gather based on SME or trainee ratings and used both as the starting parameters and in deciding on the rank order

of the items. This seems like a particularly fruitful body of potential research given the user requirements of the MCMC GGUM and its suggested sample size requirements.

Conclusion

The results of this study provide interesting results for the field of psychometrics and unfolding IRT models. While a majority of the results failed to align with expected hypotheses, they did show patterns that could be interpreted and would likely lead to important future research. The research showed that starting parameters could affect the sample size required for effective item parameter estimation, however only at sample sizes larger than typically used in psychological research. Despite this, it should serve as a useful first step towards innovative techniques in which unfolding IRT models are more feasible to broader psychological research.

References

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, *17*(3), 253–276. https://doi-org.ezproxy.umsl.edu/10.1177/014662169301700307

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology*, *86*(5), 965–973. https://doi-org.ezproxy.umsl.edu/10.1037/0021-9010.86.5.965

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440. https://doi-org.ezproxy.umsl.edu/10.1007/s11336-006-1447-6

Boyce, A. S., Conway, J. S., & Caputo, P. M. (2016). Development and validation of Aon Hewitt's personality model and Adaptive Employee Personality Test (ADEPT-15). *Unpublished technical report*.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, *57*(1), 15–22. https://doi-org.ezproxy.umsl.edu/10.1037/h0034185

Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M.-C., & Delgado, K. M. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: How theoretically appropriate measurement makes an empirical

difference. *Journal of Applied Psychology*, *99*(4), 564–586. https://doi-org.ezproxy.umsl.edu/10.1037/a0034688

Chalmers, P.R. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1-29.

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523–562. https://doi-org.ezproxy.umsl.edu/10.1207/S15327906MBR3604_03

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*(1), 88–106. https://doi-org.ezproxy.umsl.edu/10.1037/1040-3590.19.1.88

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance*, *22*(2), 105–127. https://doi-org.ezproxy.umsl.edu/10.1080/08959280902743303

Clark, L. A. (2006). When a psychometric advance falls in the forest. *Psychometrika*, *71*(3), 447–450. https://doi-org.ezproxy.umsl.edu/10.1007/s11336-006-1500-5

Conn, S. R., & Rieke, M. L. (1994). Construct validation of the 16PF. *The 16PF*.

Coombs, C. H. (1960). A theory of data. *Psychological Review*, *67*(3), 143–159. https://doi-org.ezproxy.umsl.edu/10.1037/h0047773

Dalal, D. K., Gibby, R. E., & Zickar, M. (2010). Six questions that practitioners (might) have about ideal point response process items. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*(4), 498–501. https://doi-org.ezproxy.umsl.edu/10.1111/j.1754-9434.2010.01279.x

Darr, W., Borman, W. C., St, P. L., Kubisiak, C., & Grossman, M. (2017). An applied examination of the computerized adaptive rating scale for assessing performance. *International Journal of Selection and Assessment*, *25*(2), 149–153. https://doi-org.ezproxy.umsl.edu/10.1111/ijsa.12167

de la Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov Chain Monte Carlo Estimation of Item Parameters for the Generalized Graded Unfolding Model. *Applied Psychological Measurement*, *30*(3), 216–232. https://doi-org.ezproxy.umsl.edu/10.1177/0146621605282772

de la Torre, J., & Yuan Hong. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement*, *34*(4), 267–285. https://doi-org.ezproxy.umsl.edu/10.1177/0146621608329501

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. https://doi-org.ezproxy.umsl.edu/10.1037/0022-3514.93.5.880

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). Development of the Tailored Adaptive Personality Assessment System

(TAPAS) to support Army selection and classification decisions. *Fort Belvoir, VA: US Army Research Institute for the Behavioral and Social Sciences*.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*(4), 465–476. https://doi-org.ezproxy.umsl.edu/10.1111/j.1754-9434.2010.01273.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers.

Embretson, S. E. (2004). The Second Century of Ability Testing: Some Predictions and Speculations. *Measurement: Interdisciplinary Research and Perspectives*, *2*(1), 1–32. https://doi-org.ezproxy.umsl.edu/10.1207/s15366359mea0201_1

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36–49. https://doi-org.ezproxy.umsl.edu/10.1111/emip.12111

Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll* (4th ed.). London, UK: Sage.

Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods*, *20*(3), 465–486. https://doi-org.ezproxy.umsl.edu/10.1177/1094428116689708

Griffin, M. A., Neal, A., & Neale, M. (2000). The contribution of task performance and contextual performance to effectiveness: Investigating the role of situational

constraints. *Applied Psychology: An International Review*, *49*(3), 517–533. https://doi-org.ezproxy.umsl.edu/10.1111/1464-0597.00029

Guilford, J. P. (1954). *Psychometric methods, 2nd ed*. McGraw-Hill.

Guion, R. M., & Ironson, G. H. (1983). Latent trait theory for organizational research. *Organizational Behavior & Human Performance*, *31*(1), 54–87. https://doi-org.ezproxy.umsl.edu/10.1016/0030-5073(83)90113-7

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement*, *30*(2), 143–155. https://doi-org.ezproxy.umsl.edu/10.1111/j.1745-3984.1993.tb01071.x

Hedge, J. W., Borman, W. C., Bruskiewicz, K. T., & Bourne, M. J. (2004). The development of an integrated performance category system for supervisory jobs in the US Navy. *Military Psychology*, *16*(4), 231–243. https://doi-org.ezproxy.umsl.edu/10.1207/s15327876mp1604_2

Heiser, W. J. (2006). Measurement without copper instruments and experiment without complete control. *Psychometrika*, *71*(3), 457–461. https://doi-org.ezproxy.umsl.edu/10.1007/s11336-006-1501-4

Houston, J. S., Borman, W. C., Farmer, W. F., & Bearden, R. M. (2006). *Development of the navy computer adaptive personality scales (NCAPS)* (No. NPRST-TR-06-2). NAVY PERSONNEL RESEARCH STUDIES AND TECHNOLOGY MILLINGTON TN.

King, D. R., & Roberts, J. S. (2015). ScoreGGUM: An R package for estimating GGUM person parameters using pre-calibrated item parameters and disagree–agree

response data. *Applied Psychological Measurement*, *39*(6), 494–495. https://doi-org.ezproxy.umsl.edu/10.1177/0146621615570469

Kolakowski, D., & Bock, R. D. (1973). NORMOG: Maximum likelihood item analysis and test scoring: Normal ogive model. *Ann Arbor: National Educational Resources*.

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, *91*(5), 1181–1188. https://doi-org.ezproxy.umsl.edu/10.1037/0021-9010.91.5.1181

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22 140*, 55.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Liu, C. W., & Chalmers, R. P. (2018). Fitting item response unfolding models to Likert-scale data using mirt in R. *PloS one*, *13*(5), e0196292.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–348. https://doi-org.ezproxy.umsl.edu/10.2466/03.11.PMS.112.2.331-348

Merenda, P. F. (2007). Update on the decline in the education and training in psychological measurement and assessment. *Psychological Reports*, *101*(1), 153–155. https://doi-org.ezproxy.umsl.edu/10.2466/PR0.101.5.153-155

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Scientific Software International.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi-org.ezproxy.umsl.edu/10.1177/014662169201600206

Muraki, E., & Bock, R. D. (1991). PARSCALE: Parameter scaling of rating data. *Computer program]. Chicago: Scientific Software, Inc*.

Naemi, B., Seybert, J., Robbins, S., & Kyllonen, P. (2014). *Examining the WorkFORCE™ assessment for Job Fit and core capabilities of the FACETS™ Engine*. ETS Research Report No. (RR-14-32). Princeton, NJ: Educational Testing Service.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi-org.ezproxy.umsl.edu/10.1177/01466216000241003

Oswald, F. L., & Schell, K. L. (2010). Developing and scaling personality measures: Thurstone was right—But so far, Likert was not wrong. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*(4), 481–484. https://doi-org.ezproxy.umsl.edu/10.1111/j.1754-9434.2010.01275.x

Prims, J., Motyl, M. (2018). A tool for detecting low quality data in internet research. GitHub: https://github.com/SICLab/detecting-bots

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision*

Processes, *38*(1), 76–91. https://doi-org.ezproxy.umsl.edu/10.1016/0749-5978(86)90027-0

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reise, S. R. (2010). Thurstone might have been right about attitudes, but Drasgow, Chernyshenko, and Stark fail to make the case for personality. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*(4), 485–488. https://doi-org.ezproxy.umsl.edu/10.1111/j.1754-9434.2010.01276.x

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1998). The generalized graded unfolding model: A general parametric item response model for unfolding graded responses. *ETS Research Report Series*, *1998*(2), i-53.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1999). Estimating Parameters in the Generalized Graded Unfolding Model: Sensitivity to the Prior Distribution Assumption and the Number of Quadrature Points Used.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*(1), 3-32.

Roberts, J. S., & Thompson, V. M. (2011). Marginal maximum a posteriori item parameter estimation for the generalized graded unfolding model. *Applied Psychological Measurement*, *35*(4), 259–279. https://doi-org.ezproxy.umsl.edu/10.1177/0146621610392565

Rudner, L. M. (1998). *An on-line, interactive, computer adaptive testing tutorial*. Retrieved from http://echo.edres.org:8080/scripts/cat/catdemo.htm

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Kuram ve Uygulamada Eğitim Bilimleri/Educational Sciences: Theory & Practice*, *17*(1), 321–335.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100.

Schneider, R. J., Goff, M., Anderson, S., & Borman, W. C. (2003). Computerized adaptive rating scales for measuring managerial performance. *International Journal of Selection and Assessment*, *11*(2–3), 237–246. https://doi-org.ezproxy.umsl.edu/10.1111/1468-2389.00247

Spector, P. E., & Brannick, M. T. (2010). If Thurstone was right, what happens when we factor analyze Likert scales? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*(4), 502–503. https://doi-org.ezproxy.umsl.edu/10.1111/j.1754-9434.2010.01280.x

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT Approach to Constructing and Scoring Pairwise Preference Items Involving Stimuli on Different Dimensions: The Multi-Unidimensional Pairwise-Preference Model. *Applied Psychological Measurement*, *29*(3), 184–203. https://doi-org.ezproxy.umsl.edu/10.1177/0146621604273988

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied*

Psychology, *91*(1), 25–39. https://doi-org.ezproxy.umsl.edu/10.1037/0021-9010.91.1.25

Stark, S., Chernyshenko, O. S., & Guenole, N. (2011). Can subject matter experts' ratings of statement extremity be used to streamline the development of unidimensional pairwise preference scales? *Organizational Research Methods*, *14*(2), 256–278. https://doi-org.ezproxy.umsl.edu/10.1177/1094428109356712

Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology*, *77*(4), 501–510. https://doi-org.ezproxy.umsl.edu/10.1037/0021-9010.77.4.501

Tendeiro, J. N., & Castro-Alvarez, S. (2019). GGUM: An r package for fitting the generalized graded unfolding model. *Applied Psychological Measurement*, *43*(2), 172–173. https://doi-org.ezproxy.umsl.edu/10.1177/0146621618772290

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Scientific Software International.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, *34*(4), 273.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554. https://doi-org.ezproxy.umsl.edu/10.1086/214483

Thurstone, L. L., & Chave, E. J. (1929). The measurement of attitude.

Wang, W. (2015). A bayesian markov chain monte carlo approach to the generalized graded unfolding model estimation: The future of non-cognitive measurement

[ProQuest Information & Learning]. In *Dissertation Abstracts International: Section B: The Sciences and Engineering* (Vol. 75, Issue 10–B(E)).

Wang, W., de la Torre, J., & Drasgow, F. (2015). MCMC GGUM: A new computer program for estimating unfolding IRT models. *Applied Psychological Measurement*, *39*(2), 160–161. https://doi-org.ezproxy.umsl.edu/10.1177/0146621614540514

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *Logist user's guide: Logist 5, version 1.0*. Educational Testing Service.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*(3), 189–205. https://doi-org.ezproxy.umsl.edu/10.1111/j.2044-8325.1994.tb00562.x

Yoes, M. (1995). An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model. *St. Paul, MN: Assessment Systems Corporation*.

Zinnes, J. L., & Griggs, R. A. (1974). Probabilistic, multidimensional unfolding analysis. *Psychometrika*, *39*(3), 327–350. https://doi-org.ezproxy.umsl.edu/10.1007/BF02291707

Figures

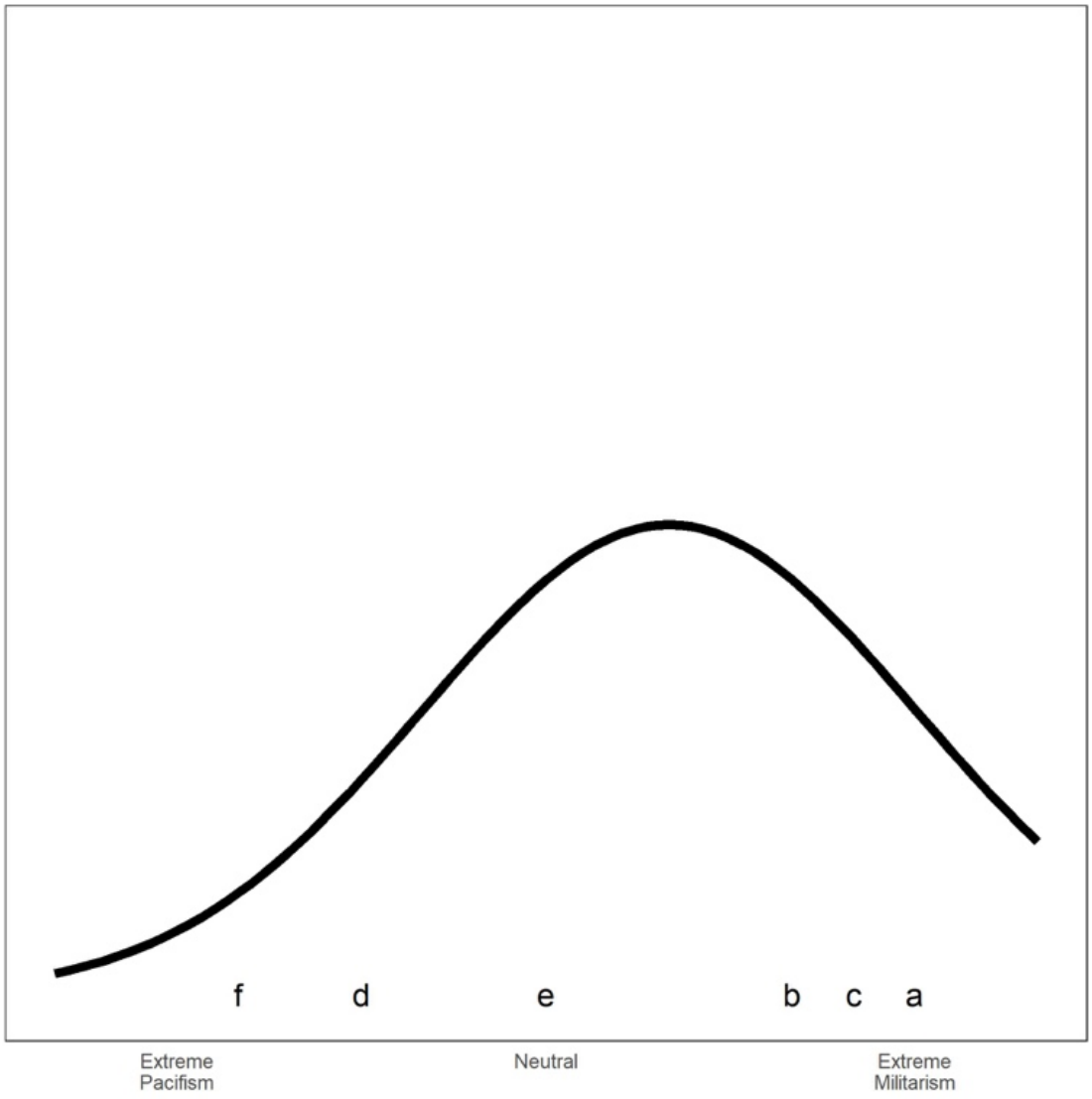


Figure 1. Locations of six militarism-pacifism attitude statements. Reprinted from Thurstone (1928, p. 537).

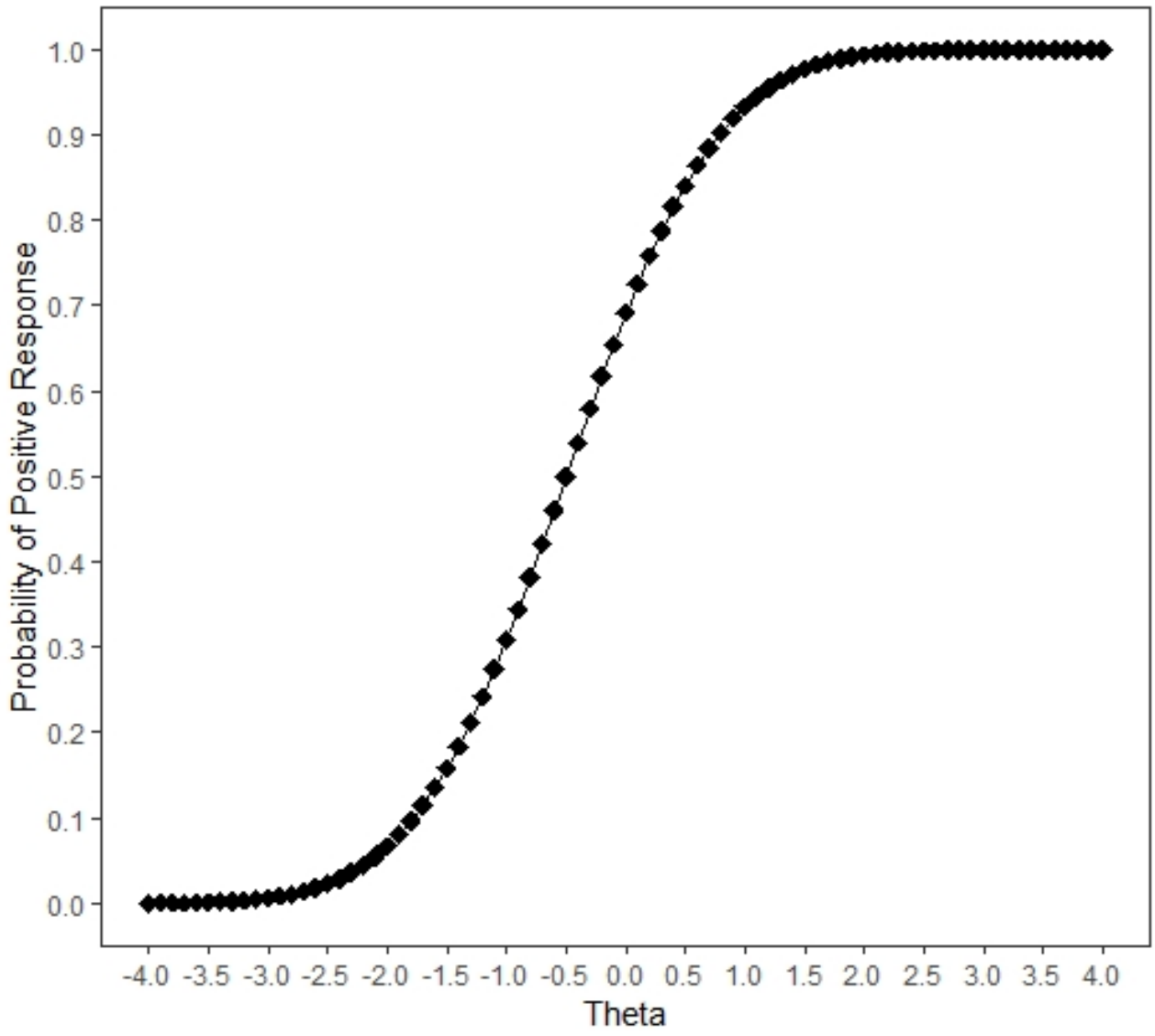


Figure 2. Example of a dominance response process model.

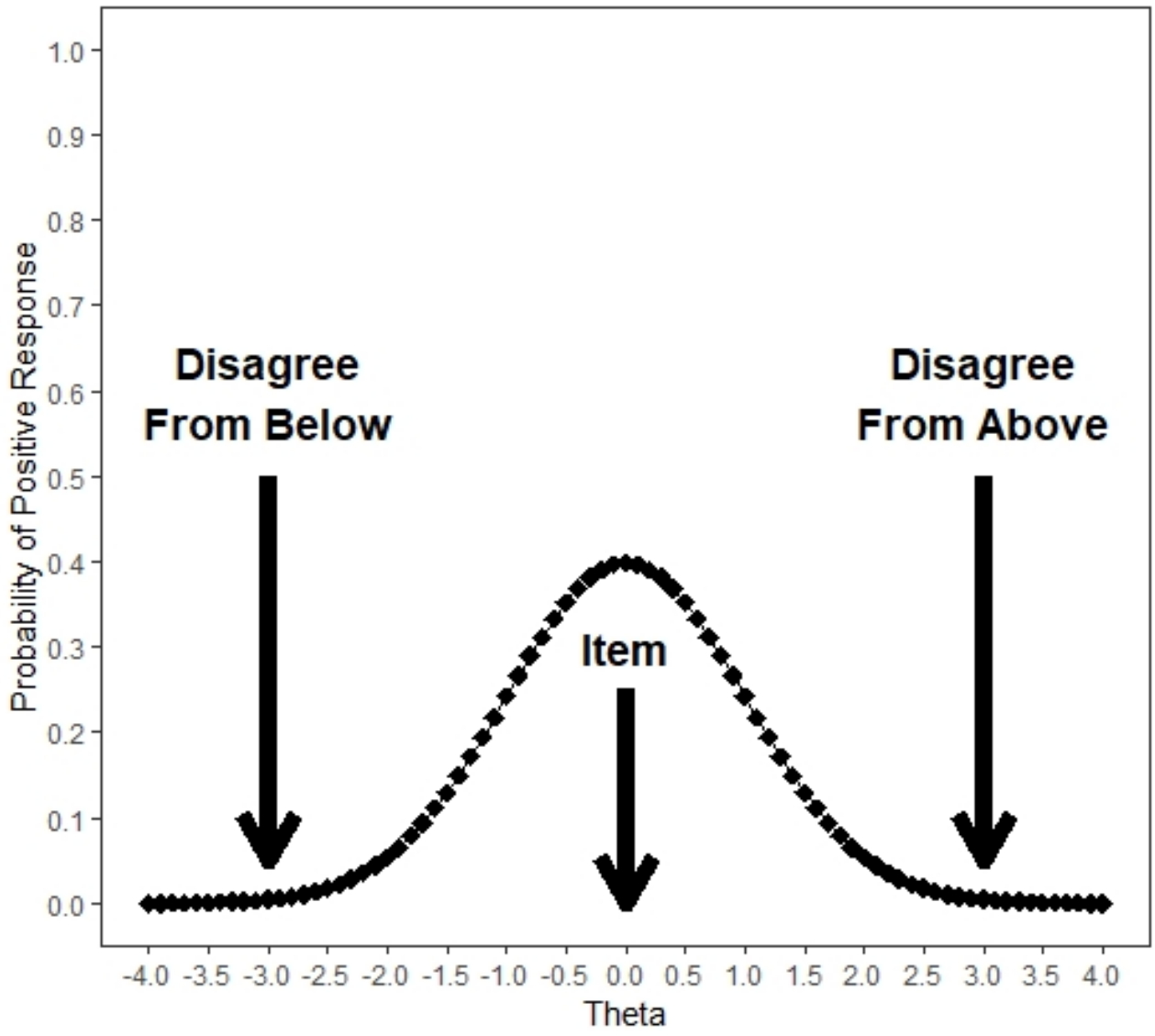


Figure 3. Example of an ideal point response process model.

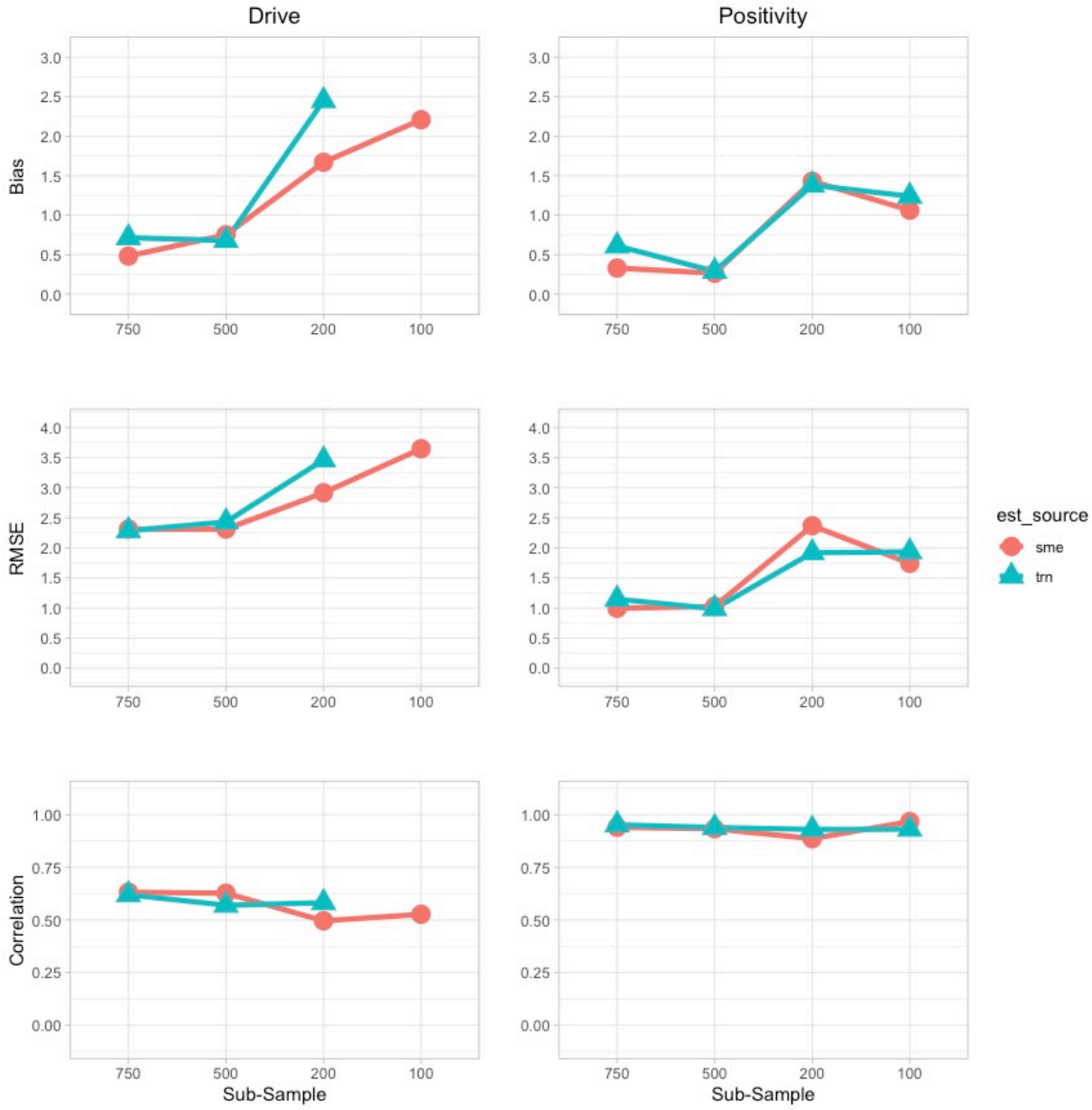


Figure 4. Plots of Parameter Recovery Metrics of Item Location Estimates Obtained Using SMEs and Trainees Estimates of Item Location as Starting Parameters Across Subsamples.

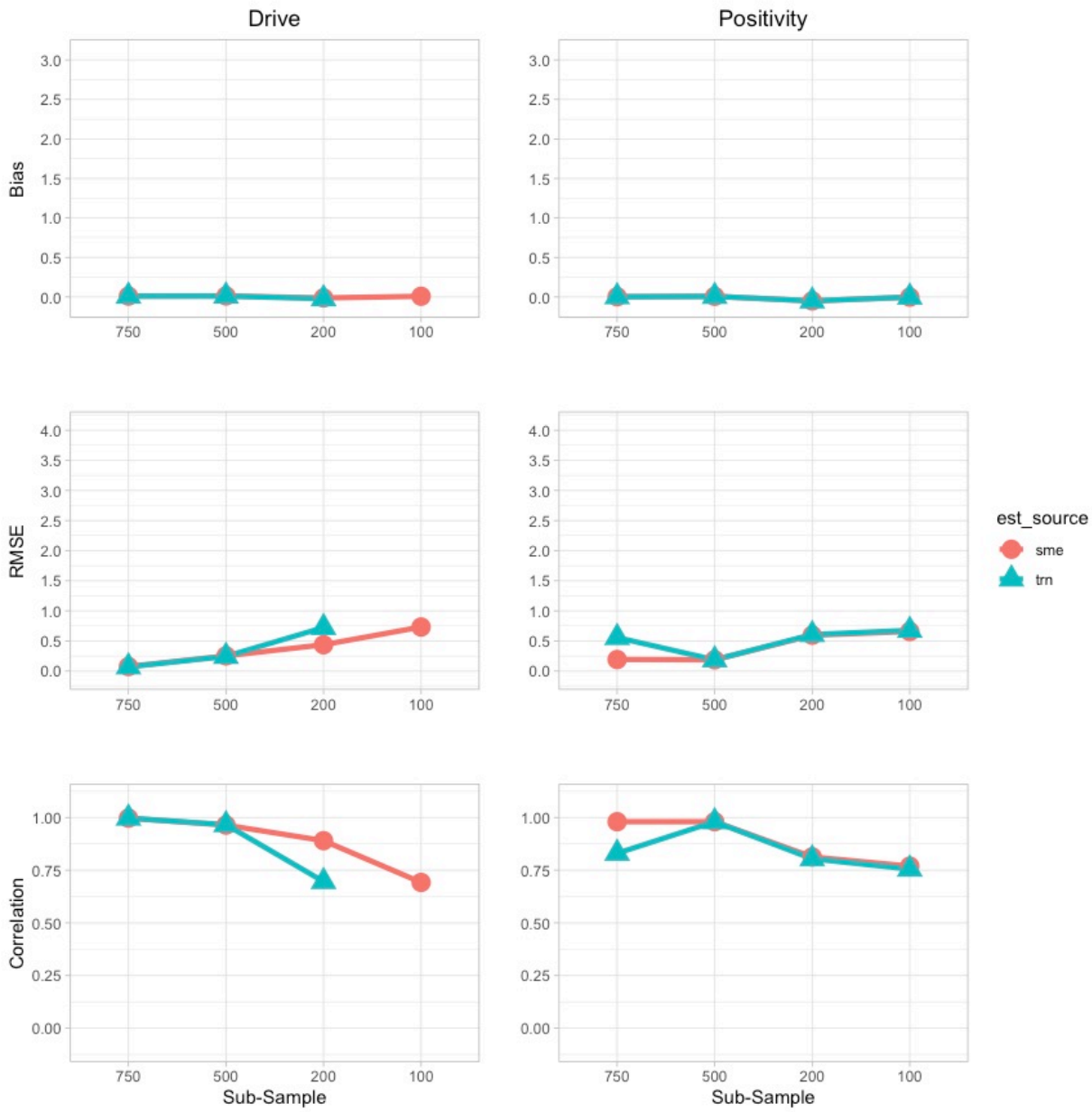


Figure 5. Plots of Parameter Recovery Metrics of Theta Estimates Obtained Using SMEs and Trainees Estimates of Item Location as Starting Parameters Across Subsamples.

Table 1. Comparison of Location Parameter Estimates of SMEs and Trainees with MML Estimates

	Positivity			Drive		
Item	SME	Trainee	True	SME	Trainee	True
1	-2.11	-2.21	-5.21	-0.94	-1.28	-5.36
2	-0.78	-0.89	-5.74	-1.28	-1.63	-1.25
3	-0.83	-1.15	-3.94	-0.89	-1.00	-1.10
4	-0.11	-0.09	0.12	-1.06	-1.00	-1.03
5	-0.83	-0.88	0.00	-1.44	-1.57	-0.24
6	-1.83	-2.16	-5.48	-0.17	0.15	0.06
7	-1.50	-1.10	-4.60	1.44	2.34	1.04
8	-1.50	-1.63	-5.28	1.28	1.96	0.68
9	0.11	0.93	0.92	-0.17	0.30	0.92
10	-0.28	0.49	1.15	2.78	2.67	0.53
11	1.72	2.11	0.93	1.61	2.40	0.61
12	1.56	2.49	1.12	2.22	2.43	0.81
13	1.50	2.50	0.91	1.67	2.07	0.53
14	2.11	2.62	0.95	1.94	2.49	0.67
15	1.22	1.93	1.13	2.44	2.70	0.90
16	1.17	1.94	1.08	0.67	1.54	0.99
17	1.39	2.57	1.02	1.56	2.54	0.91
18	1.78	2.41	1.05	2.33	2.59	0.79
19	1.39	2.35	1.07	1.17	2.00	0.64
20	1.22	1.55	0.95	0.39	1.22	0.98

Table 2a. Parameter Recovery of Item Location Estimates for SMEs and Trainees

		Bias	RMSE	r	ρ
Drive	SME	0.673	1.408	0.606	0.449
	Trainee	1.041	1.577	0.694	0.524
Positivity	SME	1.162	2.096	0.855	0.657
	Trainee	1.582	2.172	0.877	0.692

ρ = Spearman's rho for rank-order correlation.

Table 2b. Parameter Recovery of Item Location Estimates for SMEs and Trainees – Extreme Item Locations Removed

		Bias	RMSE	r	ρ
Drive	SME	0.476	1.030	0.742	0.371
	Trainee	0.881	1.320	0.840	0.454
Positivity	SME	0.111	0.699	0.646	0.108
	Trainee	0.753	1.120	0.777	0.182

ρ = Spearman's rho for rank-order correlation.

Table 3. Descriptive Statistics and Correlations for Positivity and Drive Theta Estimates Obtained Using Different Starting Parameter Sources

Trait SOURCE	*M*	*SD*	*N*	1	2	3	4	5
1. Drive SME	0.0012	0.92	821					
2. Drive TRN	-0.0080	0.94	821	**.79**				
3. Drive TRUE	-0.0009	0.92	821	**.99**	**.75**			
4. Positivity SME	0.0013	0.94	818	.63	.59	.60		
5. Positivity TRN	0.0011	0.94	818	.63	.59	.60	**.99**	
6. Positivity TRUE	-0.0015	0.94	818	.62	.56	.60	**.98**	**.98**

NOTE: SME = subject matter expert. TRN = trainee.

Table 4. Parameter Recovery of Person Estimates Obtained Using SMEs' and Trainees' Estimates of Item Location as Starting Parameters

		Bias	RMSE	*r*
Drive	SME	0.002	0.143	0.988
	Trainee	-0.007	0.652	0.753
Positivity	SME	0.003	0.18	0.982
	Trainee	0.003	0.183	0.981

Table 5. Parameter Recovery of Item Location Estimates Obtained Using SMEs' and Trainees' Estimates of Item Location as Starting Parameters Across Subsamples

		SME			Trainee		
Trait	Subsample	Bias	RMSE	*r*	Bias	RMSE	*r*
Drive	750	0.483	2.311	0.632	0.715	2.284	0.620
	500	0.752	2.312	0.627	0.680	2.431	0.570
	200	1.670	2.915	0.496	2.450	3.461	0.582
	100	2.209	3.646	0.527	--	--	--
Positivity	750	0.331	0.996	0.940	0.611	1.147	0.952
	500	0.266	1.028	0.933	0.290	0.989	0.939
	200	1.430	2.368	0.886	1.381	1.919	0.930
	100	1.063	1.743	0.967	1.240	1.934	0.931

Table 6. Parameter Recovery of Person Estimates Obtained Using SMEs' and Trainees' Estimates of Item Location as Starting Parameters Across Subsamples

		SME			Trainee		
Trait	Subsample	Bias	RMSE	*r*	Bias	RMSE	*r*
Drive	750	0.011	0.075	0.997	0.011	0.070	0.997
	500	0.013	0.252	0.964	0.012	0.244	0.966
	200	-0.012	0.437	0.890	-0.022	0.723	0.696
	100	0.009	0.732	0.692	--	--	--
Positivity	750	0.003	0.192	0.980	0.003	0.557	0.830
	500	0.008	0.187	0.980	0.008	0.187	0.980
	200	-0.050	0.595	0.813	-0.050	0.607	0.804
	100	-0.002	0.657	0.769	-0.002	0.677	0.755

Table 7. Comparison of Discrimination Parameter Estimates of SMEs with MML Estimates

	Positivity		Drive	
Item	SME	True	SME	True
1	1.89	0.36	0.83	0.46
2	1.06	0.41	0.89	0.37
3	0.56	0.45	1.06	0.47
4	0.78	0.58	1.00	0.32
5	1.00	0.32	1.33	0.27
6	1.33	0.45	0.83	0.48
7	1.11	0.53	0.89	1.50
8	1.56	0.42	0.89	0.72
9	0.78	0.87	0.56	1.19
10	0.67	1.17	1.78	1.13
11	1.50	1.33	1.33	1.31
12	1.44	1.41	1.67	1.63
13	1.33	0.83	1.33	1.26
14	1.61	1.06	1.50	1.15
15	1.11	2.03	1.61	1.55
16	1.50	1.99	1.17	2.00
17	1.39	2.05	1.06	2.51
18	1.83	1.45	1.50	1.75
19	1.39	1.27	1.22	1.22
20	1.44	1.64	0.83	1.71

Table 8. Parameter Recovery of Person Estimates Obtained Using SMEs' Estimates of Item Location and Discrimination as Starting Parameters

	Bias	RMSE	*r*
Drive	0.002	0.143	0.988
Positivity	0.003	0.18	0.982

Table 9. Parameter Recovery of Item Location Estimates Obtained Using SMEs' Estimates of Item Location and Discrimination as Starting Parameters Across Subsamples

		SME		
Trait	Subsample	Bias	RMSE	*r*
Drive	750	0.479	2.314	0.632
	500	1.602	3.197	0.488
	200	1.670	2.915	0.496
	100	2.194	3.590	0.528
Positivity	750	0.394	1.277	0.898
	500	0.296	1.247	0.899
	200	1.475	2.499	0.869
	100	1.066	1.751	0.967

Table 10. Parameter Recovery of Person Estimates Obtained Using SMEs' Estimates of Item Location and Discrimination as Starting Parameters Across Subsamples

		SME		
Trait	Subsample	Bias	RMSE	*r*
Drive	750	0.011	0.075	0.997
	500	0.011	0.653	0.760
	200	-0.013	0.317	0.942
	100	0.008	0.730	0.693
Positivity	750	0.003	0.190	0.980
	500	0.008	0.187	0.980
	200	-0.049	0.595	0.813
	100	-0.002	0.657	0.769

APPENDIX A: Personality Items

Dimension	#	Item Stem
Drive	1	Sometimes, I find it difficult to get down to work.
Drive	2	When a task is difficult, I will sometimes give up.
Drive	3	Sometimes, I find it difficult to get down to work.
Drive	4	I am not always able to successfully carry out my plans.
Drive	5	There are more important things to focus on in life than working hard at a job.
Drive	6	I tend to feel defeated when I experience a setback, but I am usually able to complete the task.
Drive	7	Others know that they can depend on me.
Drive	8	I like to get things done right away so that I do not have to think about them anymore.
Drive	9	I put in as much time and effort into my work as others.
Drive	10	I would do anything to get a task done.
Drive	11	People say they are surprised about how much I can accomplish in a short period of time.
Drive	12	My persistence has led me to finish tasks others may not have thought possible.
Drive	13	I am proud that I get tasks done faster than others.
Drive	14	I enjoy difficult tasks that require hard work and dedication, even if it means giving up my free time.
Drive	15	No matter what, I always finish what I start.
Drive	16	When I say I am going to do something, I usually follow through.
Drive	17	People who know me best would say that I am a hard worker.
Drive	18	I will spend as much time as it takes for me to be great at my job.
Drive	19	I am willing to work longer hours than the average person.
Drive	20	I usually complete my work on time.
Positivity	1	I often think about all the things that can go wrong in a situation.
Positivity	2	After a failure it may take me some time to try again.
Positivity	3	I am pessimistic sometimes.
Positivity	4	Overall, I spend about as much time being happy as I do being sad.
Positivity	5	I find life challenging most of the time, but I think a lot of people feel that way.
Positivity	6	Some people think I have a dark outlook on the future.
Positivity	7	If something very bad happens, it takes some time before I feel happy again.
Positivity	8	It is hard to let negative feelings go after an argument.
Positivity	9	After some time has passed, I will usually try again after failing at something the first time.
Positivity	10	When obstacles occur in my life, I sometimes believe that I will get through them.
Positivity	11	Although some people think I am not realistic, I always think things will turn out well in the end.

Dimension	#	Item Stem
Positivity	12	When I have many deadlines to meet, I am optimistic that I will be able to meet them.
Positivity	13	I like to cheer up those around me.
Positivity	14	I can hardly wait to see what life has in store for me in the years ahead.
Positivity	15	Even when things do not go as planned, I am still a generally happy person.
Positivity	16	I am able to see the good in most situations, even ones that at first seem bad.
Positivity	17	There is always something in my day that makes me feel happy.
Positivity	18	Being positive will always help ease a tense situation.
Positivity	19	I am happy even when things are not perfect.
Positivity	20	Most of the time, I agree with people that say you can make your dreams come true.

APPENDIX B: Training Materials and Instructions

What follows is a short informational training regarding personality testing.

In the training you will be provided with information that describes and explains two personality traits to help you complete a rating task.

After the training, you will be asked to make ratings about characteristics of personality questions which you will learn about in the training.

What follows is a short informational training regarding personality testing.

In the training you will be provided with information that describes and explains two personality traits to help you complete a rating task.

After the training, you will be asked to make ratings about characteristics of personality questions which you will learn about in the training.

What follows is a short informational training regarding personality testing.

In the training you will be provided with information that describes and explains two personality traits to help you complete a rating task.

After the training, you will be asked to make ratings about characteristics of personality questions which you will learn about in the training.

A typical personality test requires a person to respond to a number of statements indicating behaviors, tendencies, or preferences…

…Usually by selecting from various responses from disagree to agree.

Below are two examples of statements used in personality tests.

When I meet someone I like, I expect that we will become friends.

1. Strongly Disagree
2. Slightly Disagree
3. Neutral
4. Slightly Agree
5. Strongly Agree

I am confident when I communicate, but only in small groups.

1. Strongly Disagree
2. Slightly Disagree
3. Neutral
4. Slightly Agree
5. Strongly Agree

The two statements measure the personality trait **Extraversion**.

When I meet someone I like, I expect that we will become friends.

1. Strongly Disagree
2. Slightly Disagree
3. Neutral
4. Slightly Agree
5. Strongly Agree

I am confident when I communicate, but only in small groups.

1. Strongly Disagree
2. Slightly Disagree
3. Neutral
4. Slightly Agree
5. Strongly Agree

The numbered responses are typically provided for endorsement by the test-taker.

When I meet someone I like, I expect that we will become friends.

1. Strongly Disagree
2. Slightly Disagree
3. Neutral
4. Slightly Agree
5. Strongly Agree

I am confident when I communicate, but only in small groups.

1. Strongly Disagree
2. Slightly Disagree
3. Neutral
4. Slightly Agree
5. Strongly Agree

When taking a personality test, a person normally responds to multiple statements that are designed to measure the same trait (for example, Extraversion).

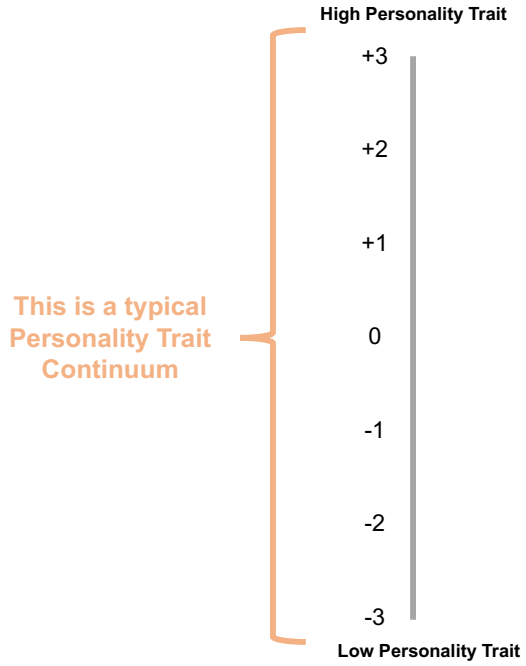
You do this because the statements represent various levels of the trait.

(in other words, some represent very low Extraversion, and others very high, some in the middle)

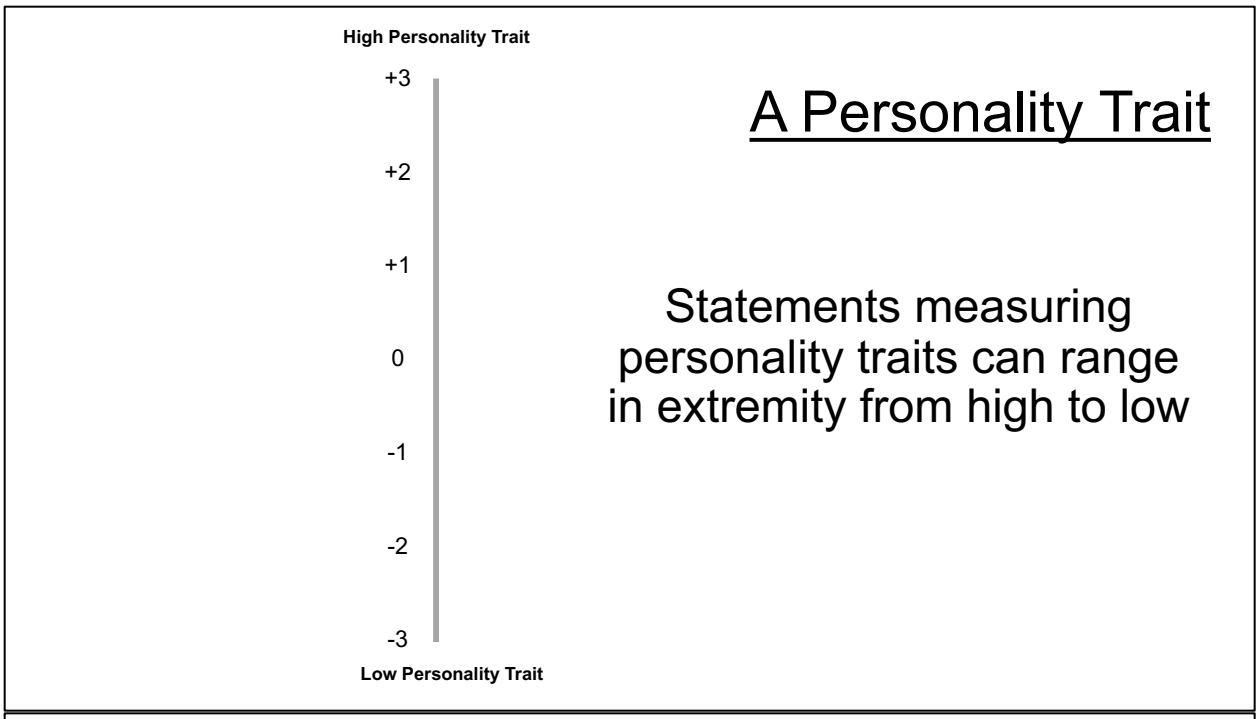
When taking a personality test, you normally respond to multiple statements that are designed to measure the same trait (for example, Extraversion).

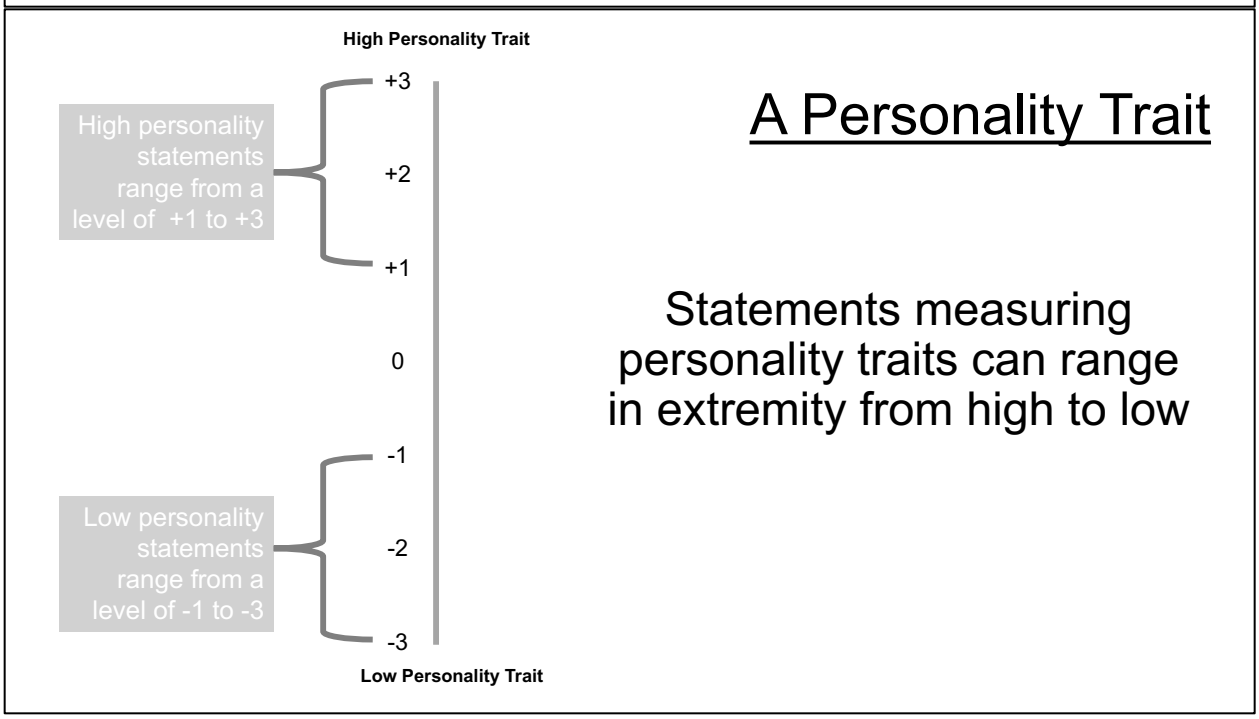
You do this because the statements represent various levels of the trait.

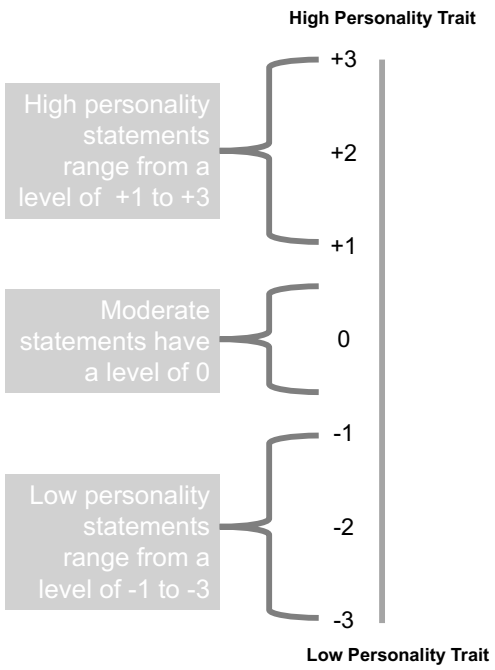
(in other words, some represent very low Extraversion, and others very high, some in the middle)



A Personality Trait


High Personality Trait
+3
+2
+1
0
-1
-2
-3
Low Personality Trait
A Personality Trait
Statements measuring personality traits can range in extremity from high to low


High Personality Trait
High personality statements range from a level of +1 to +3
+3
+2
+1
0
-1
-2
-3
Low personality statements range from a level of -1 to -3
Low Personality Trait
A Personality Trait
Statements measuring personality traits can range in extremity from high to low



A Personality Trait

Statements measuring personality traits can range in extremity from high to low

Having an understanding of the level of each statement helps estimate where the person actually falls on that personality trait.

(For example, if a person says they disagree with low level statement and strongly agree with high level statement we can infer that they are at least somewhat high on the trait of Extraversion)

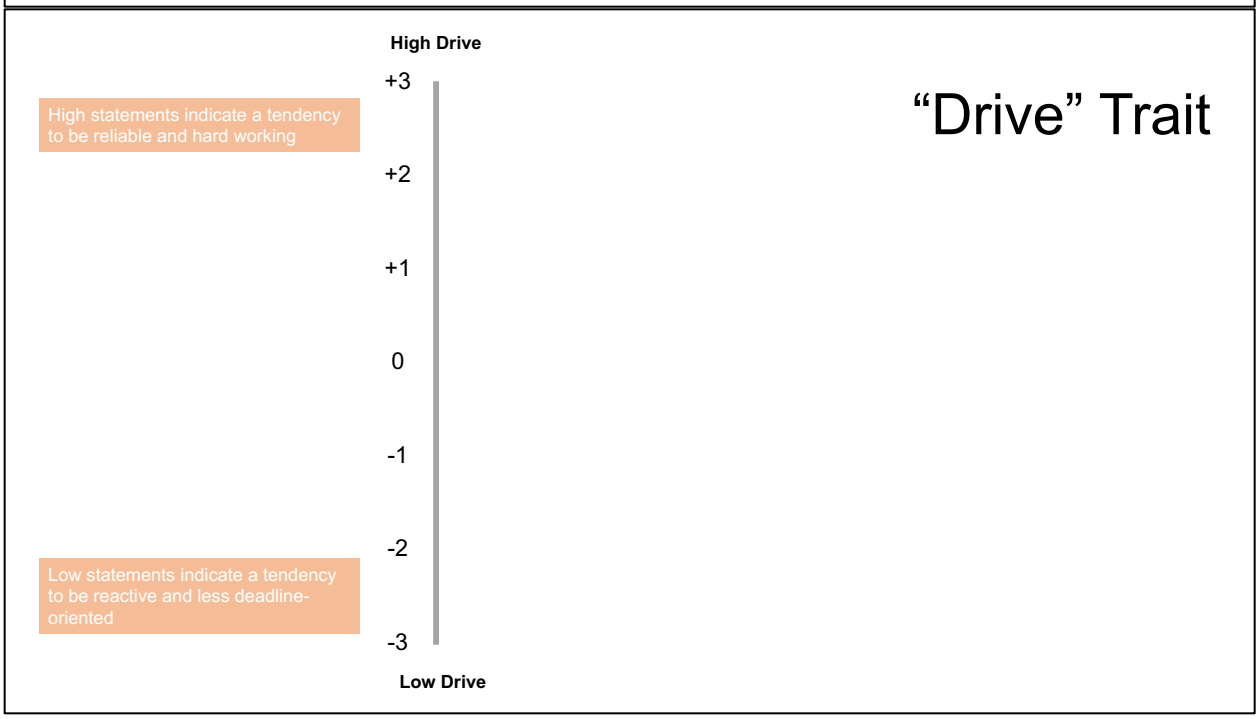
The goal of the task at the end of this study is to "rate" the appropriate level of statements for each of the traits discussed next

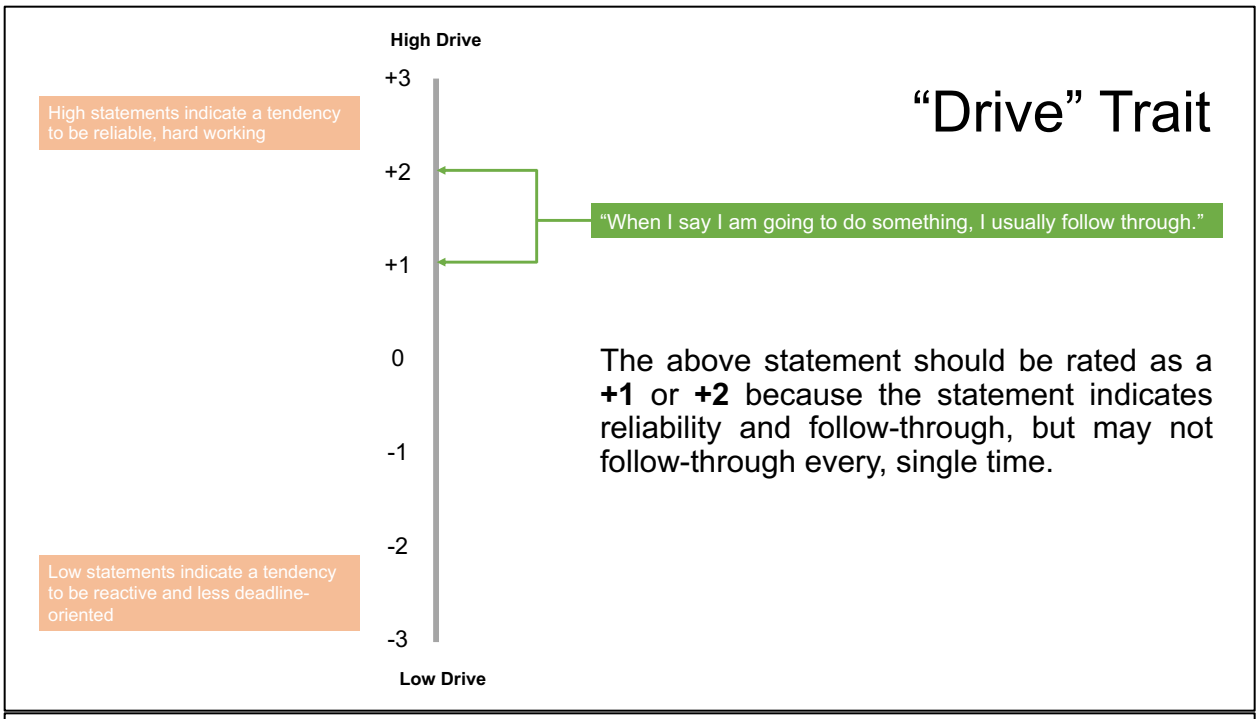
The remaining slides serve as a training regarding the two specific personality traits that on which you will be making ratings.

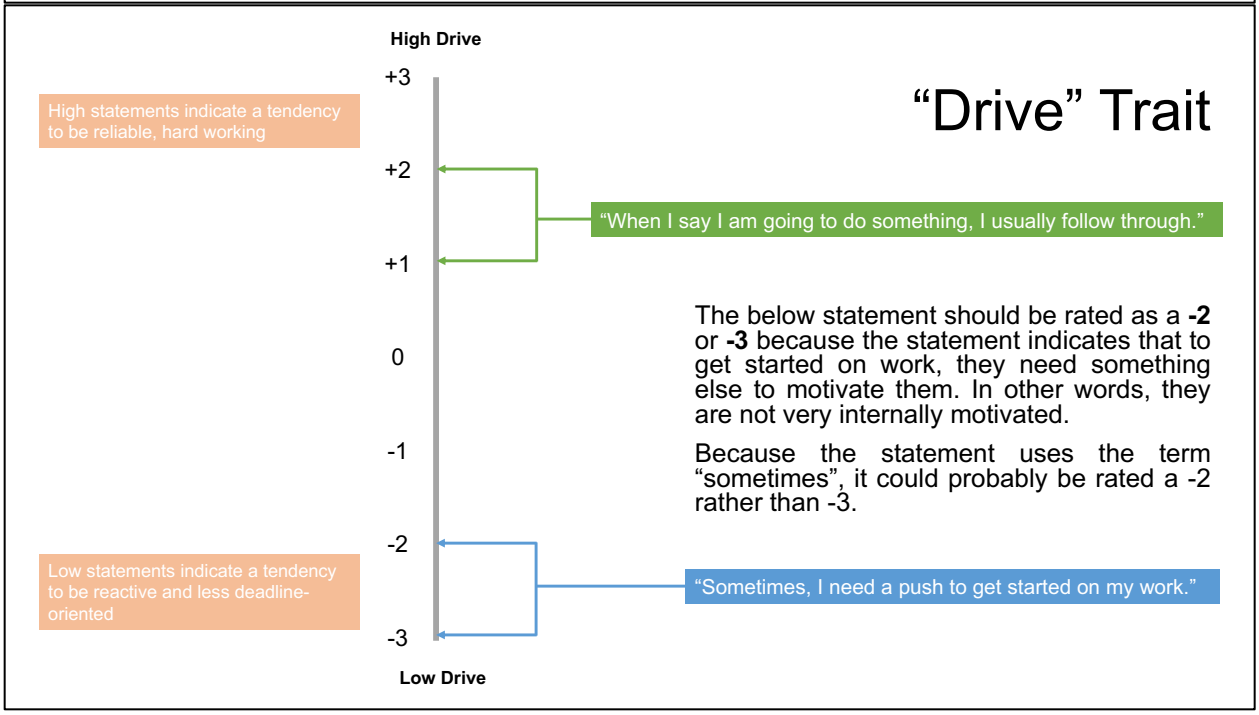
The goal of the task at the end of this study is to "rate" the appropriate level of statements for each of the traits discussed next

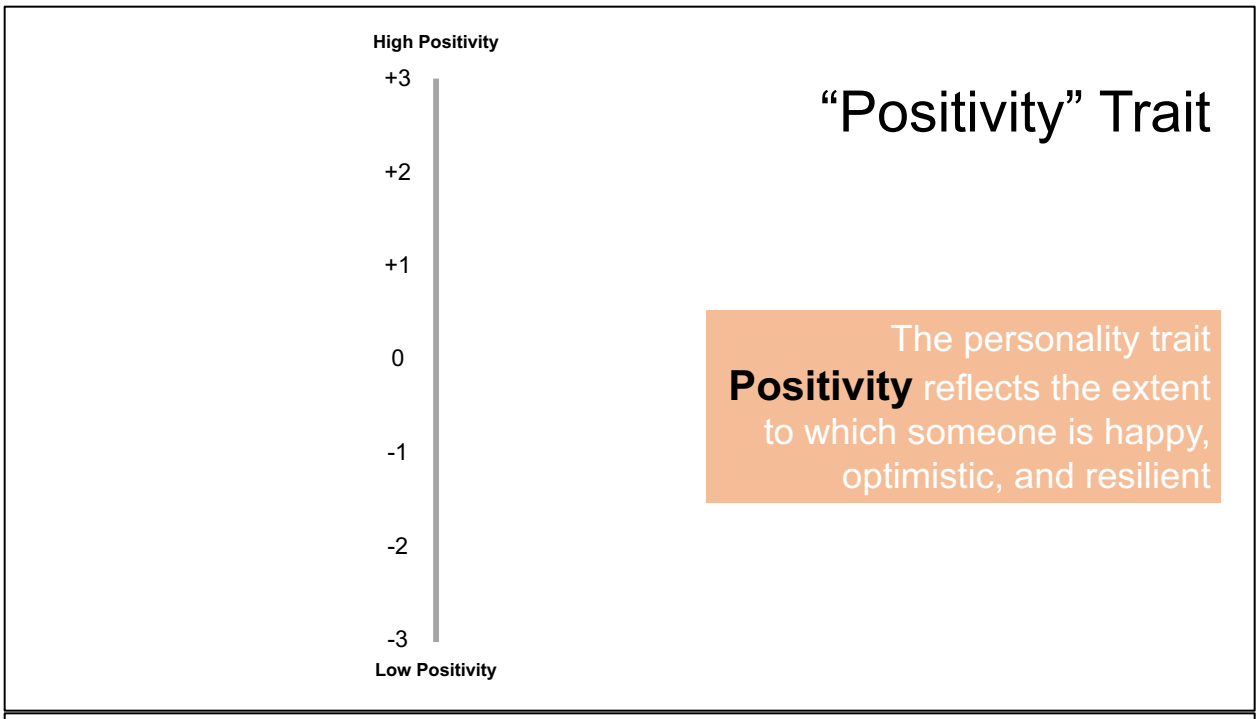
The remaining slides serve as a training regarding the two specific personality traits on which you will be making ratings

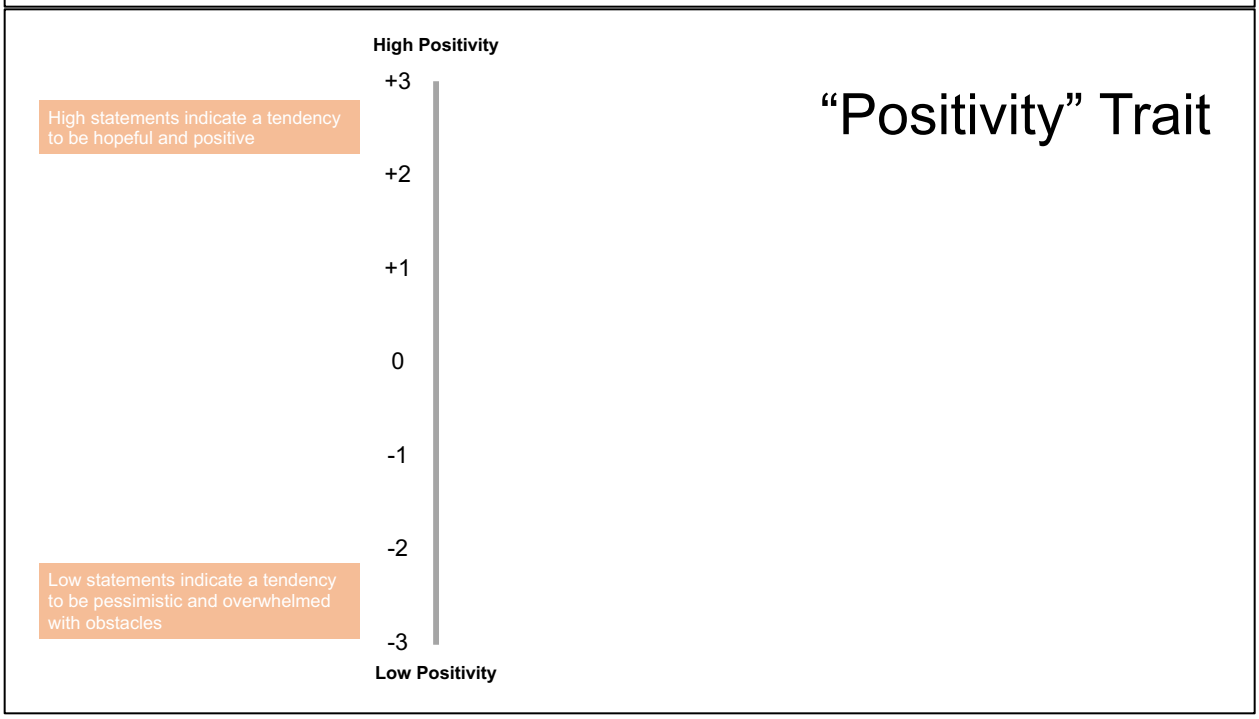

High Drive
+3
+2
+1
0
-1
-2
-3
Low Drive
"Drive" Trait
The personality trait Drive describes the extent to which someone is productive, proactive, and dependable

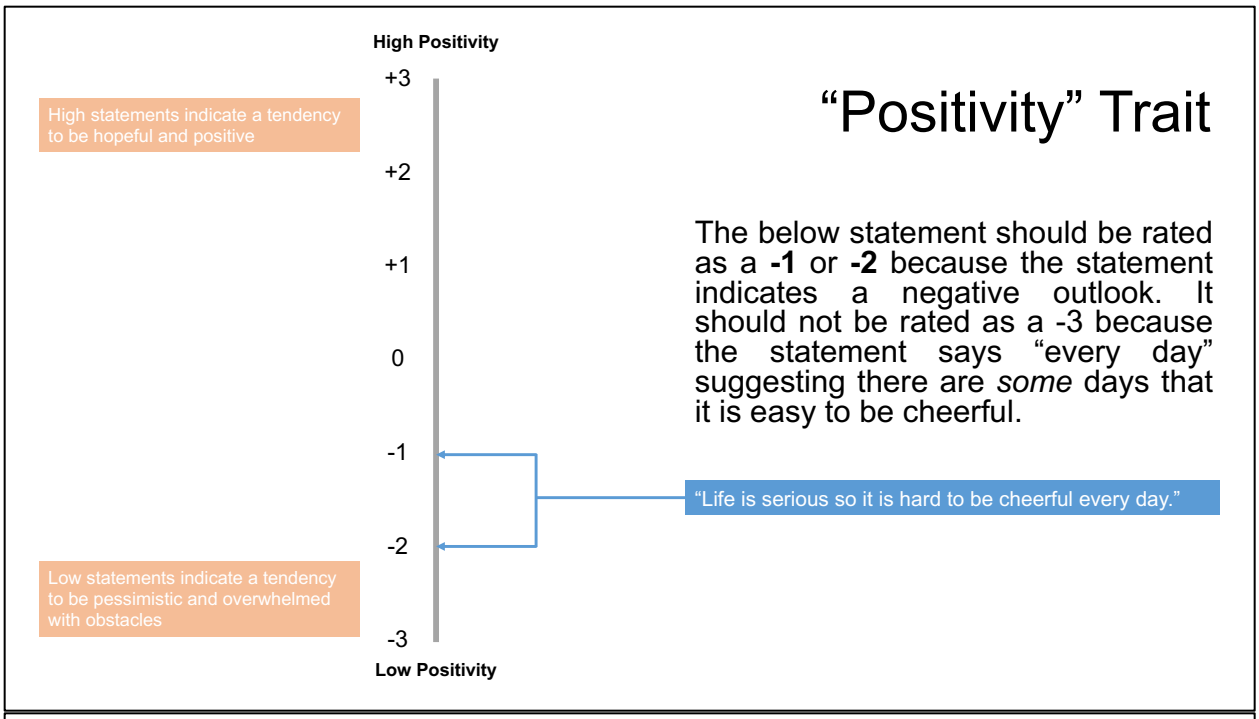

High Drive
+3
+2
+1
0
-1
-2
-3
Low Drive
"Drive" Trait
High statements indicate a tendency to be reliable and hard working
Low statements indicate a tendency to be reactive and less deadline-oriented

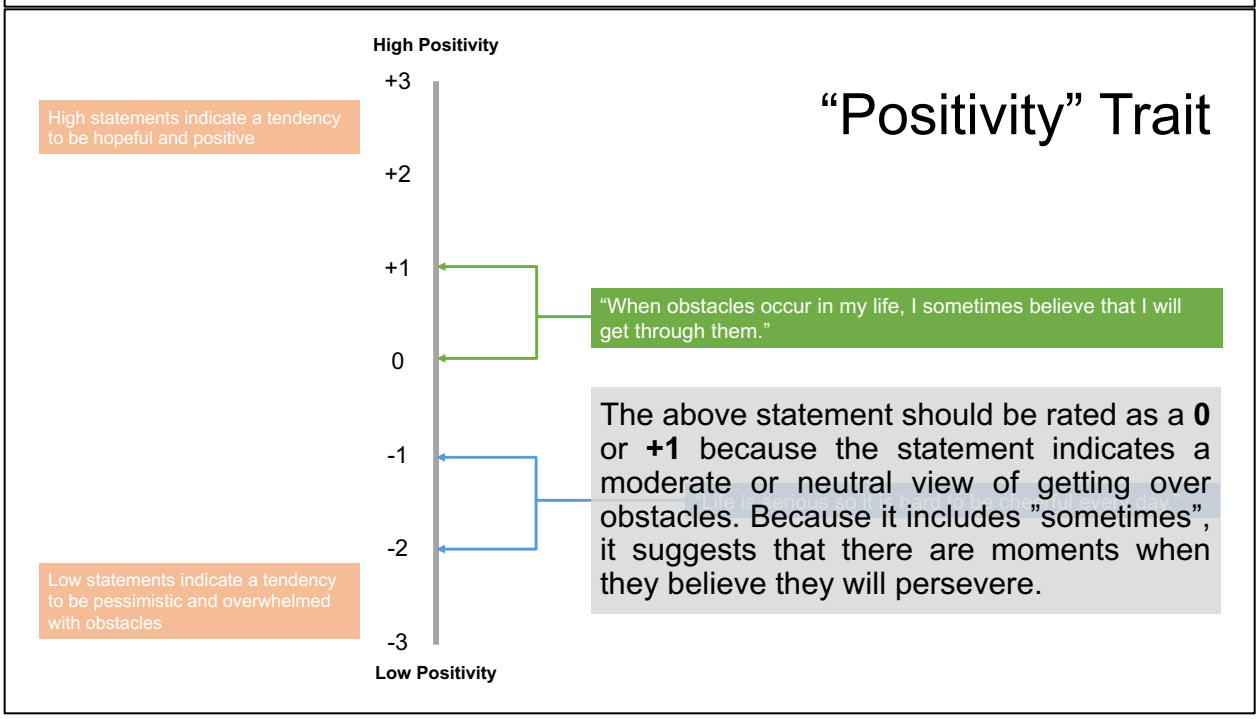

High Drive
+3
+2
+1
0
-1
-2
-3
Low Drive
High statements indicate a tendency to be reliable, hard working
Low statements indicate a tendency to be reactive and less deadline-oriented
"Drive" Trait
"When I say I am going to do something, I usually follow through."
The above statement should be rated as a +1 or +2 because the statement indicates reliability and follow-through, but may not follow-through every, single time.


High Drive
+3
+2
+1
0
-1
-2
-3
Low Drive
High statements indicate a tendency to be reliable, hard working
Low statements indicate a tendency to be reactive and less deadline-oriented
"Drive" Trait
"When I say I am going to do something, I usually follow through."
The below statement should be rated as a -2 or -3 because the statement indicates that to get started on work, they need something else to motivate them. In other words, they are not very internally motivated.
Because the statement uses the term "sometimes", it could probably be rated a -2 rather than -3.
"Sometimes, I need a push to get started on my work."


High Positivity
+3
+2
+1
0
-1
-2
-3
Low Positivity
"Positivity" Trait
The personality trait **Positivity** reflects the extent to which someone is happy, optimistic, and resilient


High Positivity
+3
+2
+1
0
-1
-2
-3
Low Positivity
"Positivity" Trait
High statements indicate a tendency to be hopeful and positive
Low statements indicate a tendency to be pessimistic and overwhelmed with obstacles


High Positivity
+3
+2
+1
0
-1
-2
-3
Low Positivity
High statements indicate a tendency to be hopeful and positive
Low statements indicate a tendency to be pessimistic and overwhelmed with obstacles
"Positivity" Trait
The below statement should be rated as a -1 or -2 because the statement indicates a negative outlook. It should not be rated as a -3 because the statement says "every day" suggesting there are some days that it is easy to be cheerful.
"Life is serious so it is hard to be cheerful every day."


High Positivity
+3
+2
+1
0
-1
-2
-3
Low Positivity
High statements indicate a tendency to be hopeful and positive
Low statements indicate a tendency to be pessimistic and overwhelmed with obstacles
"Positivity" Trait
"When obstacles occur in my life, I sometimes believe that I will get through them."
The above statement should be rated as a 0 or +1 because the statement indicates a moderate or neutral view of getting over obstacles. Because it includes "sometimes", it suggests that there are moments when they believe they will persevere.

That completes the training.

Next you will answer 2 questions and then receive specific instructions for the rating task.