EFFECTS OF RESPONDENT TRAINING ON SELF-REPORT PERSONALITY ASSESSMENT: AN ITEM RESPONSE THEORY APPROACH

BY

LUYAO ZHANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Fritz Drasgow, Chair
Professor Hua-Hua Chang
Professor R. Chris Fraley
Professor Daniel A. Newman
Professor Brent W. Roberts
Professor James Rounds

ABSTRACT

Within the item response theory (IRT) framework and inspired by the rater training literature, this study explored the effects of short online respondent training on personality item interpretation and responding and the number of response categories (i.e. polytomous vs. dichotomous) on item performance, model-data fit, and criterion-related validity. Participants recruited from MTurk (n = 1977) were randomly assigned to 1 of the 4 groups differing in training (i.e. training vs. no training) and response scale (i.e. 4-point Likert scale vs. dichotomous), and their responses to dominance and ideal-point personality measures were analyzed with GGUM, SGR, and 2PL. Results indicated that training was associated with more well-performing and more discriminating and informative intermediate items on the ideal-point scales when a dichotomous response scale was used. The dichotomous scale in general was related to better fit, while criterion-related validity stayed unaffected by both training and the response scale. Participants reported that they had been confused about personality items before, and were positive about the online training, which was consistent with the finding that trained participants on average spent 32 seconds less finishing the ideal-point surveys. Implications for future research and practice are discussed.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Fritz Drasgow, for his guidance throughout the dissertation process and throughout my time here at UIUC. This work would not have been possible without him. In addition, I would like to thank Professor Brent Roberts for his direction and support. I'm also utterly grateful for the other members of my dissertation committee: Professors Hua-Hua Chang, Chris Fraley, Dan Newman, and Jim Rounds. Each of them has contributed uniquely to this work with their valuable and insightful suggestions and comments. I would also like to thank my great friend Liwen for always being there for me.

献给爸爸妈妈。

To my parents.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

In 1990, a graduate student in Stanford University named Elizabeth Newton did an experiment where participants were assigned to the roles of either “tapper” or “listener”. Each of the tappers was asked to tap out the rhythm of a well-known song he or she picked, and the listener’s job was to guess the song. While the tappers thought they did an amazing job describing the song, and predicted that half of the listeners would get it right, the listeners ended up correctly guessing only 3 of the 120 songs tapped out. Tappers’ huge overestimation of the correct rate occurred because, in the tapper’s head, the tapping was accompanied with the melody, and it was impossible for them to imagine how clueless the listeners were when hearing only the tapping (Heath & Heath, 2006). The point is, once we’ve gained the knowledge of something, we forget what it was like when we didn’t know it, and thus when we try to communicate it to other people, we are likely to leave out some important information that we don’t realize is unbeknown to them. And this, the curse of knowledge, can make communication ineffective.

The curse of knowledge has been widely studied in the business world, where effective communication is highly critical for marketers to customers, managers to employees, and corporate headquarters to the front line (Heath & Heath, 2006).

On the other hand, the curse of knowledge hasn’t attracted much attention in the field of self-reported personality testing, where communication between researchers and participants is mainly dependent on survey items. When an item is created, we researchers have very specific and clear expectation for how it should be interpreted and processed because of the psychometrics training we’ve received in graduate school. But are the participants as knowledgeable as us about this? Perhaps yes for some simple and straightforward items describing extreme trait levels such as “I’m outgoing”, or “I’m always sad”. Such extreme items can be found in most personality measures and are usually analyzed, if within the item response theory framework, with dominance models. Recently, ideal-point models have started to get attention as they are believed to be more appropriate for describing the underlying response process used by respondents when a noncognitive construct such as an attitude or personality trait is being measured, as items assessing these types of constructs ask about typical rather than maximum behaviors (Drasgow, Chernyshenko, & Stark, 2010). The use of ideal-point models has enabled researchers to write intermediate items measuring people of moderate trait levels, so that the scale will cover a wider range of the latent trait continuum, compared with scales consisting of extreme items only. Some intermediate items are longer and more complicated than extreme items, often containing two clauses so as to describe behaviors reflecting a medium trait level. For example, “Although I have a daily organizer, I have a hard time keeping it up to date” is a typical intermediate item. According to Brown and Maydeu-Olivares (2010), such an item could be confusing to respondents as many of them will find that both clauses don’t apply, which may be frustrating and lead to random responses. However, in the eye of a researcher trained in psychometrics, especially supporters of the ideal-point response process, this item couldn’t be more clear: you agree with the item only when both parts apply, and you disagree either because you are not organized enough to even have a daily organizer (i.e. disagreeing from below), or because you are so organized that you have an organizer and are able to keep it up to date (i.e. disagreeing from above). What we as researchers often forget is that our participants do not have such systematic knowledge, and by assuming that they can process items as readily and painlessly as we do, we may have overlooked the information imbalance between us and them, and thus are afflicted with the curse of knowledge.

In fact, slightly more than 60% of all intermediate items carefully written by Huang and Mead (2014) and Cao, Cho, and Drasgow (2015) turned out to be good intermediate items with nonmonotonic item response functions (IRFs), which is impressive given the prevalent pessimism about the possibility of writing good intermediate items (Brown & Maydeu-Olivares, 2010; Credé, 2010; Oswald & Schell, 2010). But what about the other 30 percent and more? The researchers expected them to work but they didn’t. One can blame it on the researchers/writers, and that’s what the researchers did, although they had little idea what went wrong. Researchers simply assumed that if an item was good enough, the respondents would have no problem interpreting it as expected, so all intermediate items would work and show unfolding as intended. The problem is that researchers are so familiar with various types of intermediate items they write that they can’t imagine how baffled innocent participants might be when reading them. The researchers are like those tappers who couldn’t get the melody out their head when tapping and thus underestimated the difficulty of the task for listeners ignorant of the tune of the song. Blaming failed items solely on the writers is comparable to saying that bad tapping is completely responsible for the low correct rate. You can have John Bonham as the tapper and it’s still not going to solve the problem, not just because he's dead, but also because tapping skill is at least not the only problem here. The absence of melody is responsible as well.

Admittedly all those items were not perfect, and more effort should be put into sharpening them up, but this shouldn't exempt the respondents from being considered as a potential solution. The thing is, every time an item turns out to be an unpleasant surprise during data analysis, the communication between researchers and participants fails, and the two sides should both be closely examined. Several studies have been conducted trying to write good intermediate items (e.g. Chernyshenko, Stark, Drasgow, & Roberts, 2007; Huang & Mead, 2014; Cao et al., 2015), but little effort has been devoted to figuring out what is going on with the respondents. One study, by LaPalme, Tay, and Wang (2017), found that high verbal ability, an individual characteristic, was related to responding via an ideal-point response process, rather than a dominance process, for affect and attitude items. Their explanation was that higher verbal ability leads to better understanding of the item and hence more precise introspection, a key feature of the ideal-point process.

The debate has long been going on over whether the dominance or the ideal-point IRT model should be the "go-to" model for self-reported personality data, and the critics of the ideal-point IRT model such as the Generalized Graded Unfolding Model (i.e. the GGUM; Roberts, Donoghue, & Laughlin, 2000) insist that it's not worth the time because (1) good intermediate items that can only be analyzed by models like the GGUM are hard to develop, and (2) the GGUM has repeatedly been found to yield comparable criterion-related validity as dominance models such as the 2PLM or the SGRM. Studies have already shown that intermediate items are possible to write (Huang & Mead, 2014; Cao et al., 2015), but it seems that there's not much that can be done to improve the criterion-related validity of GGUM trait estimates, at least on the scale development side (Chernyshenko et al., 2007; Huang & Mead, 2014; Cao et al., 2015), although GGUM is believed to better model the response process than a dominance model (Broadfoot, 2008).

Therefore, in the current thesis, I focused on the people who responded to personality surveys, as I hypothesized that GGUM's problems were partly due to the respondents' lack of knowledge of the ideal-point response process and intermediate item interpretation, a point that had been overlooked by researchers due to the curse of knowledge. I was curious about whether eliminating the information imbalances by educating respondents about how personality items were expected to be processed and interpreted would improve the precision of GGUM estimates and lead to better model-data fit and validity. Based on past studies, the method used for scale development (i.e. dominance approach vs. ideal-point approach; Carter, Dalal, Guan, LoPilato, & Withrow, 2017; Tay, Ali, Drasgow, & Williams, 2011) and the number of response categories (i.e. four-point Likert vs. dichotomous scale; Chernyshenko et al., 2007; Broadfoot, 2008) were also taken into consideration.

CHAPTER 2: LITERATURE REVIEW

Item response theory (IRT)

Within the IRT framework, rather than the entire test, the unit of analysis is the item. They are used to estimate a respondent's standing on the latent trait continuum (Wainer & Mislevy, 2000). IRT is a test theory that focuses on describing the nonlinear relationship between the latent trait level (i.e. theta), item parameters, and item response patterns. Unlike classical testing theory (CTT), IRT is test and sample independent, meaning that a respondent should have the same theta estimated no matter which set of items he or she answers, and a given item should have the same difficulty and discrimination no matter who responds.

When the IRT method is adopted, the first question is which IRT model to use. It is always important to choose the correct psychometric model, which helps researchers have deeper insight into the nature of people's responses, and avoid mistakes in results and conclusions (Drasgow et al., 2010). Today, two types of approaches are widely used for developing and analyzing self-report personality measures: the dominance approach and the ideal-point approach. The ideal-point approach has received great attention over the past two decades, but the dominance approach has been the more accepted and used approach for analyzing personality survey responses.

In the current study, the dominance IRT models used were Samejima's Graded Response Model (SGRM; Samejima, 1969), and the two-parameter logistic model (2PLM). The former is able to handle polytomous data, while the latter models dichotomous data. As to the ideal-point model, the GGUM, now the most popular ideal-point IRT model, was applied to both polytomous and dichotomous self-report personality data.

The Dominance Perspective

The dominance approach derived from Likert's (1932) approach to analyzing rating scales, and was later named by Coombs (1964). It assumes that the higher a participant's trait level, the more likely she will answer positively. Therefore, the relationship between the probability of endorsement and the trait level can be represented by a monotonically increasing function (Hambleton, Swaminathan, & Rogers, 1991). Common techniques used in the dominance approach for personality scale development or analyses include item-total correlations, discriminant analysis, principal component analysis (PCA), and factor analysis (FA; Roberts, Laughlin, & Wedell, 1999).

With the dominance approach, an item is considered good if it has a strong correlation with the other items and loads highly onto the same factor as most of the other items. Relatively neutral items are more likely to fail to meet such criteria (van Schuur & Kiers, 1994), and eventually will be excluded from the scale. This usually leads the scale to end up containing only the extreme items.

When an IRT perspective is taken, the monotonic nature of the dominance approach will produce the monotonically increasing ogive item characteristic function (IRF; Figure 28). The y-axis stands for the probability of a positive response ranging from 0 to 1. As shown in the IRF, as the trait level goes up along the x-axis, the probability of endorsement goes up. This property is the essential feature of the dominance IRT approach.

In this study, two types of widely used dominance IRT models, the 2PLM and the SGRM were used for dichotomous and polytomous data, respectively.

Two-parameter Logistic Model (2PLM). The item response function (IRF) for the 2PLM is:

$$P_i\,(\theta) = \frac{\exp[Da_i(\theta-b_i)]}{1+\exp[Da_i(\theta-b_i)]},$$

where $P_i\,(\theta)$ is the probability of a random respondent correctly with trait level θ answering Item i correctly or positively.

There are two item parameters in a 2PLM. The difficulty parameter, b_i, is the point on the latent trait (θ) scale where the probability of a correct response is equal to 0.5. As suggested by the names, the larger the difficulty parameter, the higher the trait level is required for a positive response. a_i is the discrimination parameter, and it reflects the degree to which an item is able to discriminate between respondents with different latent trait levels. The value of a_i is proportional to the slope of the probability function at the location of b_i on the trait continuum (Hambleton et al., 1991). Thus the larger a_i is, the steeper the item characteristic curve (ICC) will be. D is the scaling factor that lets the logistic function resemble as closely as possible the normal ogive curve, and is usually set equal to 1.702 (Valbuena, 2004).

Samejima's Graded Response Model (SGRM). SGRM (Samejima, 1969) is an extension of the 2PLM (Kosinski, 2009) and is one of the most popular polytomous models in personality research. Under SGRM, a polytomous response is broken down to a series of binary response sets by boundary response functions (BRF), which are obtained by successively merging response options (Kosinski, 1999). The probability of a respondent with a trait level equal to θ selecting response option k equals the probability of endorsing response option k and higher minus that of endorsing response option k+1 and higher. The probability of selecting option k on item i is given by:

$$P_{i,k}\,(\theta) = \frac{1}{1+\exp\left[-Da_i\left(\theta - b_{i,k}\,\right)\right]} - \frac{1}{1+\exp\left[-Da_i\left(\theta - b_{i,k+1}\,\right)\right]}$$

In SGRM, separate difficulty parameters $b_{i,k}$ are estimated for each step of an item, while every item has just one discrimination parameter a_i for all steps. The scaling factor D has the same meaning as in 2PLM.

The Ideal-Point Perspective

The ideal-point approach was first introduced in Thurstone's 1928 paper to measure attitudes, and the term was coined later by Coombs (1964). One of the greatest differences between the dominance approach and the ideal-point approach lies in one of their assumptions. According to the ideal-point approach, a respondent will endorse a statement only if the statement is reflecting his or her level of the latent trait. The closer the item location is to the respondent's standing on the trait continuum, the higher the probability of endorsement, regardless of whether the trait level is high, medium, or low. When a respondent disagrees with an item, he or she could disagree either from above the item or below the item.

Drasgow and colleagues (2010) argued that compared to the dominance approach such as CTT, FA, and the dominance IRT models, the ideal-point approach should be more appropriate for self-report personality testing. This is because just like attitude items, personality items also require introspection. The dominance approach, on the other hand, should be more suitable for cognitive ability tests where one's maximum capacity is measured.

Unfolding IRT Models. IRT models developed based on the ideal-point assumption allows the IRF to bend down after the peak (Figure 29), which is called "unfolding". Unfolding is often observed when an ideal-point model is applied to intermediate or neutral items endorsed by respondents with a moderate trait level. When an item has an extreme location, meaning it takes an extreme trait level to endorse the item, unfolding still happens, but the resulting IRFs are approximately monotonic (Roberts et al., 2000) and very similar to IRFs produced by dominance IRT models, as the unfolding usually happens beyond the range of observed values of the latent trait. Since unfolding models are able to handle both extreme items and neutral items, they may be considered as a more general form of the dominance model.

The most widely used unfolding model is the GGUM (Roberts et al., 2000), which was used in the current study.

Generalized Graded Unfolding Model (GGUM). GGUM is applicable to both dichotomous and polytomous response data. As discussed above, ideal point models assume a response process different from dominance models. According to Roberts et al. (2000), the GGUM was developed based on four basic premises about the response process. The first premise is that an individual tends to agree with an item with a standing on the trait continuum that's close to her own trait level. The second premise is that a respondent disagrees with an item because the item trait level is either higher or lower than her own trait level. Similarly, a person closer to an item on the latent trait continuum can also agree with this item from either above or below. The third premise is that subjective responses (not observed responses) to attitude statements follow a cumulative item response model. The last premise is that an individual is equally likely to agree with an item located either *h* units above or below her position on the attitude continuum. Developed from these four premises, the formal definition of the GGUM is:

$$\mathrm{P}\,[Z_i = z \mid \theta_j\,]$$

$$= \frac{\exp\{\alpha_i[z(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}]\} + \exp\{\alpha_i[(M-z)(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}]\}}{\sum_{w=0}^{C}\{\exp\{\alpha_i[w(\theta_j - \delta_i) - \sum_{k=0}^{w}\tau_{ik}]\} + \exp\{\alpha_i[(M-w)(\theta_j - \delta_i) - \sum_{k=0}^{w}\tau_{ik}]\}\}}$$

This function gives the probability associated with the *j*th respondent's observable response to the *i*th item. Z_i is the observable response to item *i*, and z ranges from 0 to C, with 0 standing for the strongest level of disagreement, and C standing for the strongest level of agreement. C equals the number of response options minus 1. M equals 2*C+1, representing the number of subjective response categories minus 1. α_i is the discrimination parameter, and δ_i is the location parameter of item i on the latent trait continuum. τ_{ik} is the location of the *k*th subjective response category threshold on the theta continuum relative to the location of the *i*th item. The τ_{ik}s are symmetric about the point $(\theta_j - \delta_i) = 0$.

Dominance IRT models vs. Ideal-point IRT models

Both the dominance and the ideal-point models are often used in today's self-report personality research, and the debate over which one should be preferred is ongoing. Quite a few studies have compared the performance of these two types of IRT models by examining the model fit, criterion-related validity, easiness of creating intermediate/neutral items, and the number of respondents following either the dominance or the ideal-point response process.

The response process and model-data fit. When choosing the IRT model for analyzing and developing self-report personality measures, researchers aim to pick the one that best describes the underlying response process of the respondents, or in other words, the model that fits the data generated by participants using a specific response process. This is essential, as model-data misfit is likely to lead to inaccurate estimation of test scores, cross-cultural comparisons, and the study of context effects (Stark, Chernyshenko, Drasgow, & Williams, 2006). For one type of IRT model to be recognized as more appropriate than others, better model-data fit should be observed first. However, by the year of 2001, despite the great number of studies applying various dominance IRT models, including the one-, two-, and three-parameter logistic models (1PLM, 2PLM, and 3PLM, respectively) and the SGRM to data collected using different personality measures (Ellis, Becker, & Kimmel, 1993; Schmit & Ryan, 1997; Rouse, Finger, & Butcher, 1999; Stark et al., 2006), very few studies had carefully examined their model-data fit. Chernyshenko, Stark, Chan, Drasgow and Williams (2001) first fitted the 2PLM, 3PLM, and SGRM to data from Goldberg's Big Five Factor Markers (Goldberg, 1997, 1998) and the fifth edition of the Sixteen Personality Factor Questionnaire (16PF; Cattell & Cattell, 1995; Conn & Rieke, 1994). To their surprise, these measures developed with the dominance approach all showed fit that was less than satisfactory. 2PLM and 3PLM fitted some of the scales reasonably well, and SGRM generally didn't fit well. It seemed that the dominance IRT models were unable to capture the characteristic of the personality data. Inspired by Levine's nonparametric maximum likelihood formula scoring model (MFSM; Levine, 1984; Levine & Williams, 1991, 1993), Stark et al. (2006) used the same data from 16PF as in Chernyshenko et al. (2001) and applied 2 dominance models (2PLM and MFSM with dominance constraints) and 2 ideal-point models (GGUM and MFSM with ideal-point constraints) to it. Looking at the IRFs, the authors found that that nine of the 16PF scales actually had items that showed nonmonotonicity even though the measure was invented based on the dominance assumption. Also, by examining the adjusted χ^2 /df ratios for item singles, doubles, and triple and the IRFs, it was found that for the seven scales with no unfolding items, dominance and ideal-point models had similar fit, while for seven of the nine scales with unfolding items, MFSM with ideal point constraints showed the best fit, and GGUM did a great job as well. Dominance models showed better fit only for two of the nine scales containing unfolding items. This is the first study that compared the model-data fit of dominance and ideal-point models for personality data, and the results suggested that the ideal-point IRT models based upon the ideal-point response process might be the more appropriate ones for self-report personality testing.

Ever since Stark and colleagues' 2006 groundbreaking study, more studies have compared the fit of GGUM and dominance IRT models, and yielded similar results that GGUM had better or equivalent fit compared to dominance model. For example, the Chernyshenko Conscientiousness Scale (CCS) scale was developed based on the ideal-point assumption, and it was found that 2PLM had terrible fit for the Order scale due to the existence of unfolding items, while GGUM had no problem fitting the data well (Chernyshenko et al., 2007). Weekers and Meijer (2008) fitted 1PLM and GGUM to both dominance-based (Dutch personality inventory, the NPV-J; Luteijn, van Dijk, & Barelds, 2005) and ideal-point-based personality scales (the Dutch translation of the Order scale of the CCS; Chernyshenko et al., 2007). They found that GGUM managed to fit both measures well, while 1PLM was only able to fit the NPV-J, but not the CCS Order scale because it failed to model the intermediate items, again. GGUM was also found by Carter and Dalal (2010) to have better fit than the SGRM for the Work Scale of the Job Descriptive Index (JDI). In a simulation study, Tay et al. (2011) fitted 2PLM, SGRM, and GGUM to dichotomous and polytomous data generated from each of these models, and found that these models generally worked the best when they are applied to the data they were used to generate, plus GGUM was able to fit 2PLM data well when the test was short. Ling, Zhang, Locke, Li, and Li (2016) fitted GGUM to the Circumplex Scales of Interpersonal Values (CSIV; Locke, 2000), and, as expected, GGUM fitted better than the generalized partial credit model, a dominance model (GPCM; Muraki, 1992).

The studies on IRT model fit have provided indirect evidence that the ideal-point response process is likely to be the process used by respondents for answering personality items. More direct evidence was reported in LaPalme et al. (2017). The authors adopted a pairwise comparison method other than model fit examination to explicit study the noncognitive within-person item response process. Rooted in Thurstone's comparative judgment model (Thurstone, 1927), this method generates ipsative data that result in ordered preferences (e.g. A is preferable to B). During the online study, the participants were first randomly given 15 adjective items from the Minimarker scale of the Big 5 personality traits (Saucier, 1994), and they were asked to indicate their levels of endorsement to these adjectives (e.g., 1 = I am not creative, 2 = I am slightly creative, 3 = I am moderately creative, 4 = I am very creative, 5 = I am extremely creative). This first set of items served as respondents' self-reported latent standing on the personality trait continuum. After the regular personality survey was the pairwise preferences survey, using the same 15 adjectives. However, this time the participants were presented with every adjective 10 times, and each time only two of the five endorsement levels were given. For example, participants were provided with a paired comparison of "I am not creative" and "I am slightly creative", and asked which was more like them. There were 10 unique combinations of all endorsement levels in total, and participants completed all 10 paired comparisons for all 15 adjectives, with both the order of paired-comparison sets and the items within each set randomized.

Judging from the different assumptions underlying the dominance and ideal-point response processes, LaPalme et al. (2017) believed that there would be a difference in their paired preferences between respondents using the ideal-point and the dominance response processes. For example, suppose items A, B, C, D, and E, from least difficult to most difficult, are evenly spaced along a unipolar latent continuum. Respondents who followed the ideal-point response process tend to endorse items that are closer to their latent trait level, and therefore, if an individual's trait level is B based on the self-reported personality survey, then the ordered preferences for the paired comparison survey should be B>A=C>D>E, as B is a perfect match with the trait standing, and the farther away an item is from B, the less likely it will be endorsed. On the other hand, if the respondents answered items following a dominance response process assuming that the higher the trait level, the higher the endorsement probability, they would be more likely to endorse items that are easier (i.e. farther below the individual's trait standing), and therefore, the order of preference of an individual with the trait level at B should be A>B>C>D>E. Note that these predictions were made based on the weak stochastic transitivity (WST; Tversky, 1969) of the pairwise preferences between ordered response options, which assumes that given a set of preferences (e.g., B > A) for decisions lying on an ordinal scale, the preference for option A must be independent of the other presented options (e.g., B, C, D, and E), and the order of preference must not change (e.g., preference A>B>C>D>E).

With such expectation in mind, the authors compared the regular self-reported responses with the paired comparison responses, and decided that only individuals who consistently followed one response process for all 150 paired comparisons would be considered as adopting that response process. They found that on average 37.69% of all individuals strictly used the ideal-point response process, while only 12.20% used the dominance response process.

Although the model fit studies and pairwise comparisons suggest that in general the ideal-point response process is preferred for self-report personality tests, they've also showed that the dominance models are not completely unreasonable either. Actually, Brown and Maydeu-Olivares (2010) believed that although respondents engaged in introspection and a comparison process, what they compared themselves to might not always be the standing of the item. Whether respondents compare themselves to the item location (i.e. ideal-point process) or to a certain threshold (i.e. dominance process) should depend on the targeted construct and the items measuring it. For example, many times an individual will endorse an item with a binary (endorse/not endorse) choice if it's utility is larger than a threshold (Brown & Maydeu-Olivares, 2010), which is consistent with a dominance response process, and thus calls for a dominance model. As a result, in order to further investigate whether a threshold or an ideal-point mechanism best describes the response process, researchers should turn to intermediate items, which, compared to extreme items, are much more effective in differentiating an ideal-point and a dominance process (Brown & Maydeu-Olivares, 2010). This suggestion is consistent with previous studies mentioned before reporting that GGUM outperformed dominance models in model-data fit mostly when there were items with nonmonotonic IRFs (i.e. good intermediate items).

Ironically, GGUM's unique ability to model intermediate items has been used against its utility, as some researchers have found intermediate items difficult and time-consuming to write, and there haven't been clear guidelines (Brown & Maydeu-Olivares, 2010; Dalal, Withrow, Gibby, & Zickar, 2010). Fortunately, in recent years, effort has been made to develop and evaluate intermediate items, and impressive progress have been made.

Intermediate Items. Thurstone (1928), the supporter of the later coined phrase "the ideal-point response process" (Coombs, 1964) was one of the very first to include intermediate items in a measure, although it was attitude instead of personality that was being measured. Thurstone used six statements that represented the low, medium, and high levels of attitudes toward militarism-pacifism, and an individual's attitude was estimated by using the mean of the levels of the statements that the individual endorsed. This is a critical scaling method, as it allows respondents who endorse the same number of statements to be differentiated (Drasgow et al., 2010).

Likert, on the other hand, was against intermediate items, because he believed that intermediate items were double-barreled and incapable of differentiating people's attitudes, and that their low item-total correlations suggested that they failed to measure whatever the other items were measuring (Likert, 1932). Therefore, intermediate items are nowhere to be found in Likert's measures. When measuring internationalism, Likert (1932) proposed an alternative to the Thurstone scaling that was later called "the dominance response process" (Coombs, 1964), and it was the well-known 5-point response scale with integer scoring (i.e. "Strongly Disapprove = 1," "Disapprove = 2," "Undecided = 3," "Approve = 4," and "Strongly Approve =5,"). Likert also introduced reverse scoring for negative items, a step after which an individual's total score could be computed using the sum or mean of item scores (Likert, 1932). This scaling and scoring approach was found to have as high or higher reliability as Thurstone's method (Likert, 1932).

As mentioned earlier, intermediate items are the key for GGUM having better fit than dominance models. When there are no nonmonotonic items on a scale, it's likely that the GGUM will end up having fit similar to the dominance models, which makes the GGUM a lot less attractive due to its complexity. Unlike Likert from almost nine decades ago, today, even researchers who are not big fans of the ideal-point response process have admitted that intermediate items are important and useful (e.g. Oswald & Schell, 2010; Brown & Maydeu-Olivares, 2010), but they worry that good intermediate items are hard to create and can be too confusing for respondents. Therefore, their suggestion was that we stick to extreme items and dominance models for now (Oswald & Schell, 2010; Brown & Maydeu-Olivares, 2010), but "now" was 2010.

Since 2010, effort has been devoted to developing intermediate personality items. If good intermediate items can be written, then the GGUM will be one step closer to being the more versatile and appropriate model for self-report personality research. Two studies that focused on intermediate items are Huang and Mead (2014) and Cao et al. (2015). Both papers explored the possibility of writing unfolding items with various tactics, and evaluated their performance at both the item and the scale level.

Huang and Mead (2014) wrote ideal-point items using three tactics: average, double-barreled, and neutral. Items written with the average tactic describe typical behaviors of individuals with an average level of the trait (e.g. “My ability to plan is at about average”, or “I am about as careful as most others” for conscientiousness). The authors believed that such items will show unfolding because respondents with low and high trait levels will disagree with them from below and above, respectively, and only those score in the middle will agree. Cao et al. (2015) also wrote this type of intermediate items and named it “**A**verage”.

The second tactic used by Huang and Mead was called “double-barreled”, and items falling into this category would be made up of two parts with opposing stimuli in order to describe mixed behaviors. Example items are “Sometimes I’m industrious and other times I’m lazy” and “Although I have a daily organizer, I have a hard time keeping it up to date (Chernyshenko et al., 2007). Respondents were expected to disagree with such items until both parts apply. Cao et al. (2015) didn’t have the same label, but came up with the “**F**requency” (e.g. “Sometimes I can tolerate the messiness of my room.”) and the “**T**ransition” (e.g. “I can ignore a mess for a long time, but eventually I clean it up.”) categories, which overlap mostly with the “double-barreled” category of Huang and Mead (2014).

The third tactic of Huang and Mead (2014) was “neutral”, and items in this category would be carefully worded to consist of stimuli that were between the two extremes of the trait continuum, as least in the writers’ opinion. An example item would be “I trust what people say until they prove me wrong” for agreeableness. The “neutral” category is similar to the fourth and also the last category of intermediate items in Cao et al. (2015) labeled “**C**ondition” (e.g. “I will lead a group only when I’m interested in getting the task done” for the dominance facet of extraversion).

The overlap is not complete between the three tactics of Huang and Mead (2014) and the four categories (i.e. “FACT”) of Cao et al. (2015), and I won’t elaborate on this as it’s beyond the scope of the current study. What matters most is that both groups of researchers had very similar views about what intermediate items should look like, and they appeared to have covered the complete range of possible kinds of ideal-point items.

In terms of the performance of these intermediate items, Huang and Mead (2014) wrote 76 ideal-point items in total for the five Big Five personality dimensions, and 60.5% of them were considered successful, as their item parameters were able to be estimated, and the IRFs were nonmonotonic as expected. This supported the argument that good intermediate items can be developed. However, this rate was significantly lower than what the authors found with the dominance items they wrote, whose successful rate was 76.52%, suggesting that it was harder to write good ideal-point items than dominance items. Among the three tactics, the “neutral” one performed the worst, with only 15 out of the 35 items in this category turning out to be successful, giving a success rate of 43%. The other two tactics, “average” and “double-barreled” both had over 70% of items that worked as intended, with the former being 79% and the latter being 73%. In fact, with the “neutral” items removed, the ideal-point items would have comparable success rate as the dominance items. In terms of dimensions, openness, with a success rate of a mere 21%, was found to be the hardest to write good intermediate items for, while success rates for the other four dimensions were all above 60%.

Patterns discovered by Cao et al. (2015) were generally consistent with Huang and Mead (2014). In Study 1, for each of the three lower-order personality facets (i.e. Order under Conscientiousness, Dominance under Extraversion, and Curiosity under Openness) they studied, twelve intermediate items were written, three for each of the “FACT” categories. About 64% of the 36 new intermediate items turned out to have nonmonotonic IRFs, and similar to Huang and Mead (2014), the “A”, or the “Average” category had the best performance, with 8 of all 9 items considered good, followed by the “F” and “T” categories with 6 and 5 successful items, respectively. The “C” type had the weakest performance with only 4 unfolding item working. As mentioned earlier, in general, “A” is the same as the “average” tactic, “F” and “T” correspond to the “double-barreled” tactic, and “C” is similar to the “neutral” tactic. Therefore, in both papers, ideal-point items concerning average stimuli consistently worked the best, with the “double-barreled” items combining opposing stimuli as the runner up, and the “neutral” items involving stimuli of medium levels being the least effective. Moreover, Cao and colleagues also discovered that the Curiosity scale had the most failed intermediate items (i.e. 7 out of 12), adding to the conclusion of Huang and Mead (2014) that it was harder to write good intermediate items for openness. In study 2, Cao et al. (2015) kept only the Order and the Dominance scales, and collected more data. With 5 out of 6 items working, “T” joined “A” as the most effective category, leaving “F” and “C” at the bottom, with only 3 and 2 working of all 6 items, respectively. In summary, the “A” type, whose items are as simple as extreme items, was no doubt the best-performing category. “F” and “T”, or “double-barreled” items worked to some extent, but the absolute success rate varies from sample to sample. “C” or “neutral” is the least effective category with a fail rate over 50% or even 60%.

As to why a number of intermediate items didn't work out as expected, Huang and Mead (2014) frankly admitted it was "an unanswered question". So far, attempted explanations have focused mainly on the characteristics of the items and the scales, such as the word and topic choice (Huang & Mead, 2014), or the 4-point Likert scale being too familiar to respondents (Cao et al., 2015). Interestingly, Broadfoot (2008) fitted the GGUM to dominance personality measures, and found some extreme items showing unfolding IRFs, which is the characteristic of intermediate items. For example, an item as extreme as "I'm always prepared" actually had a nonmonotonic IRF. On the other hand, items that seemed neutral (i.e. items that contained "sometimes") were not flagged as unfolding at all. The author was just as uncertain as everyone else.

What I've learned from these studies is that item responses do not always function as expected. Importantly, little attention has been paid to the respondents. The varying success rates of intermediate items across samples and the puzzling inconsistency between item types (i.e. extreme or intermediate) and their IRFs (i.e. folding or unfolding; Huang & Mead, 2014), in my opinion, is pointing to the possibility that besides the writing of intermediate items, their interpretation may not be universal. In other words, the items as well as the respondents could be both responsible for the functioning of personality items. One untested assumption underlying failed intermediate items is that respondents interpret even complicated items as intended by item writers. However, LaPalme et al. (2017) told a different story: high verbal ability led to more consistent use of the ideal-point response process for responding to attitude and affect items. It seems that not everyone shares the item writers' understanding of noncognitive items, only those with high verbal skills.

In the current study, I provided some of the participants with explicit knowledge about how to answer personality items, especially the intermediate items. This is particularly important because intermediate items have been found to yield higher test information at the extreme trait levels (Huang & Mead, 2014; Cao et al., 2015), which has important implications for personnel selection and for clinical diagnosis. Thus, providing information to respondents about how to answer such items may result in improved measurement.

Criterion-Related Validity. Although some intermediate items performed well (e.g. the "average" type and some of the "double-barreled" type), it was found that the criterion-related validities of ideal point measures were just comparable to that of the dominance model measures (e.g. Ling et al., 2016; Stark et al., 2006; Chernyshenko et al., 2007). According to Stark et al. (2006), ideal-point models fitted the data as well as or even better than dominance models, suggesting higher precision of ideal-point trait estimates. If this is true, then higher criterion-related validity should be expected for ideal-point scores. However, Chernyshenko et al. (2007) collected data with the Order Scale of the CCS, to which they fitted both the dominance and the ideal-point models and found that both correlated very similarly with external criteria such as study behavior and health behaviors. Since the data for the ideal point analyses were dichotomous, they hypothesized that dichotomization would reduce the amount of information contained in the data, preventing the GGUM from having superior criterion-related validity. Therefore, Broadfoot (2008) in Study 1 applied the polytomous ideal-point model GGUM and the polytomous dominance model GPMC to data collected using the 50 Big Five Factor Markers from the International Personality Item Pool (IPIP; Goldberg, 1990). The author correlated the respondents' estimated scores with their academic performance and scores on a situational judgment test (SJT). In addition, impact analyses were carried out to compare the rank order of respondents obtained with the two models. Although the author found that GGUM and GPMC had similar criterion-related validities, the impact analyses showed that trait estimates differed substantially in rank order at the upper end of the score distribution. According to the author, such a difference could have important implications for personnel selection, as very different decisions would be made based on the IRT model adopted, if a top-down selection strategy is to be used (Broadfoot, 2008). GGUM also was found to have uncovered more non-linear relationships than GPMC (i.e. 7 vs. 1) between personality traits and the criteria. However, whether the criteria-related validity, rank order, or non-linear relationships were accurate couldn't be determined by the empirical study.

As a result, Broadfoot (2008) included a Study 2, which was a simulation study to further compare the performance of the two IRT models. In Study 2, empirical item parameters for the Conscientiousness and the Agreeableness scales were used to generate responses, and these two scales were chosen because they represented scales with a lot of unfolding (7 out of 10 items) and minimal unfolding (1 out of 10), respectively. Responses were generated using either GGUM or GPMC depending on the property of each of the **items, not scales**. Therefore, the responses to the 7 conscientiousness and the 1 agreeableness items that were unfolding were generated using GGUM, whereas responses to the others were generated with the GPCM.

It turned out that the number of true unfolding items did influence the performance of the models. For the Agreeableness scale with predominantly extreme items, the criterion-related validity was more accurately estimated by the GPCM than the GGUM, whereas the GGUM recovered the correlation more precisely between thetas and the external criteria for the Conscientiousness scale.

In terms of the rank orders, consistent with the empirical study, greater rank order differences were found at the upper end of the theta distribution. As to which model got the distribution right, again, for the Agreeableness scale, the GPCM thetas showed higher correlations with the true thetas, while GGUM did a better job recovering thetas for the Conscientiousness scale. The same pattern held that GPCM was favored by the more extreme scale (i.e. the Agreeableness scale) and the GGUM by the more unfolding one (i.e. the Conscientiousness scale) based on thetas at the upper end.

Broadfoot (2008) reported that the GGUM detected more curvilinear relationships than the GPMC. Carter et al. (2017) added to this conclusion with two simulation studies exploring the performance of the GGUM and SGRM in detecting curvilinear relationships. They generated responses by using both the dominance and the ideal-point approaches, which was equivalent to assuming that the underlying response process was either dominance or ideal-point. All data were analyzed by both models, and it was found that when the generation approach (i.e. the assumed response process) matched the model used for analysis, the detection performance was the best with appropriate power and Type I error rates.

In sum: (1) significantly more people appeared to utilize the ideal-point response process (e.g. LaPalme et al., 2017) consistently, which could be predicted by verbal ability, although many people used a mixture of both; (2) the GGUM had comparable fit to the dominance model and superior fit when the scale contains intermediate items with non-monotonic IRFs (Chernyshenko et al., 2001; Cao et al., 2015); (3) intermediate items expanded the range of the trait continuum that a scale could cover and provided useful information about an individual, but they could only be properly modeled by ideal-point models (Stark et al., 2006; Drasgow et al., 2010); (4) contrary to what many researchers thought, intermediate items could be written and many of them worked very well, but why some of them didn't work was unknown (Huang & Mead, 2014; Cao et al., 2015) and the person factor had never been considered; (5) the criterion-related validity of ideal point and dominance trait estimates were often very similar; GGUM worked better than the dominance models for predicting external criteria and detecting curvilinear relationships when the scale held a certain amount of nonmonotonicity and the underlying response process is ideal-point instead of dominance.

Importantly, in "self-report" personality testing for adults, the "self" is understudied. One can argue that respondents make no mistakes as any reaction is the "natural" reaction, but this assumes away any problems. Instead, researchers may have overlooked an important source of variance: Respondents' understanding of what researchers are asking them to do. LaPalme et al. (2017) showed that the inconsistency between participants' responses and researchers' expectation might be a result of the participants' difficulty in comprehending the items. Such confusion could be even worse when the sample has low motivation such as students participating for course credit and MTurk workers getting paid 50 cents/hour. All in all, while a great amount of time and energy have been devoted to improving items and scales, the possibility shouldn't be overlooked that the participants are underperforming when they take the survey due to the lack of knowledge about how to do it. LaPalme et al. (2017) examined verbal ability as a predictor of response process with attitude and affect items, so I believed it was reasonable to hypothesize that some clarification of the respondents' task would improve the quality of the data.

Therefore, in the current study, I taught some of the participants about the ideal-point response process and intermediate items in order to eliminate the potential information imbalances, and see how it would affect the psychometric properties of the ideal-point (i.e. the GGUM) and the dominance IRT models.

Personality and Industrial-Organizational Psychology (I-O Psychology)

Today, personality testing is enjoying its second heyday in I-O Psychology (Hough & Schneider, 1996). The first one lasted for about one decade and ended when personality was concluded to have correlations with major work outcomes that did not differ from zero (Locke & Hulin, 1962; Guion & Gottier, 1965). Then, the five-factor structure of personality (i.e. the Big Five) and meta-analytic studies on validity brought the second peak of personality, especially when researchers finally started to pay attention to artifacts such as restriction of range and unreliability of criterion measures. Extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience, are the five personality dimensions that are believed to cover most trait-related adjectives in the English language (Goldberg, 1981, 1990, 1992, 1993), as well as most existing personality inventories (Costa, Busch, Zonderman, & McCrae, 1986; McCrae & Costa, 1989; Costa & McCrae, 1988). This widely accepted framework then was used in a series of meta-analyses, with artifacts corrected, to demonstrate the usefulness of the Big Five for predicting important work-related outcomes. For example, the five personality factors were found to have corrected mean validities ranging from 0.16 for extraversion to 0.33 for agreeableness (Tett, Jackson, & Rothstein, 1991). Barick and Mount (1991) obtained the same results: Conscientiousness in particular was found to be a meaningful predictor of performance across all occupational groups. When multiple factor scales were combined to form a compound variable, even higher validity was observed. Ones, Viswesvaran, and Schmidt (1993) created a new measure of integrity by integrating the conscientiousness, agreeableness, and emotional stability measures, and found it correlated with overall job performance at 0.41 after artifacts were accounted for. Compared to cognitive ability tests, personality measures also had less adverse impact against minorities (e.g., Sackett, Burris, & Callahan, 1989; Ones et al., 1993; Feingold, 1994; Hough, 1996). Having meaningful and consistent correlations with job-related criteria while only being weakly associated with intelligence, personality has become a most interesting predictor in personnel selection.

Conscientiousness

Conscientiousness is a personality construct that reflects individual differences in characteristics such as being diligent, organized, rule abiding, self-controlled, and responsible to others (Roberts, Lejuez, Krueger, Richards, & Hill 2014; Roberts, Jackson, Fayard, Edmonds, & Meints, 2009), and is related to a variety of important outcomes.

Conscientiousness and Overall Job Performance. Among the Big Five, conscientiousness has emerged to be one of the strongest and the most stable predictors of job performance (Anderson & Viswesvaran, 1998; Barrick & Mount, 1991; Salgado, 1997). This is intuitive, in that people who are more conscientious tend to work harder and be more responsible and self-disciplined, all of which are likely to be beneficial to almost all types of work. According to Barrick and Mount's (1991) meta-analysis, conscientiousness was the most consistent predictor of three kinds of job performance (job proficiency, training proficiency, and personnel data) across five different types of occupations (professionals, police, managers, sales, and skilled/semi-skilled). The correlations between conscientiousness and job performance ranged from 0.20 to 0.23. These findings were later replicated by Salgado (1997) using a European community sample. Salgado found an overall validity coefficient of .25 for conscientiousness, which was the highest among the Big Five, and that the validity existed across occupations (police, managers, sales, and skilled labor), and the coefficients ranged from .16 to .39.

In 2000, Hurtz and Donovan revisited the relationship between the Big Five personality dimensions and job performance, and pointed out that previous meta-analyses (Barrick & Mount, 1991; Tett et al., 1991; Salgado, 1997) were flawed in construct validity, for a large number of the measures used in earlier studies were not designed to explicitly measure the Big Five personality dimensions. Therefore, the authors included only the scales that were explicitly designed to measure the Big Five, and concluded that the true validity of conscientiousness was 0.2 across occupations and performance criteria. One year later, overcoming the deficiency of having a small number of studies in some of the previous meta-analyses (e.g., Tett et al., 1991), Barrick, Mount, and Judge (2001) published a study where they quantitatively summarized 15 prior meta-analyses that studied the relationships between the Five Factor Model (FFM) and job performance. Conscientiousness again stood out as the most valid predictor across performance measures and occupational types. It had the highest average true score correlation estimate of the five personality dimensions, ranging from the mid .20s to low .30s, with the upper bound of the 90% credibility values of these validity estimates in the upper .30s. To avoid having to classify predictors by construct, Hogan and Holland (2003) in their meta-analytic study took in only the studies that used the Hogan Personality Inventory (HPI). In HPI, the prudence construct served as conscientiousness (mean correlation between the two was .51), which had a correlation with overall job performance at 0.24.

Conscientiousness and Contextual Performance or Citizenship Behavior. Being an important part of the job performance, contextual performance (Motowidlo & Van Scotter, 1994) has also been found to be associated with conscientiousness. Unlike task performance, which is usually task-oriented and required in job description, contextual performance is focused on meeting or exceeding what is prescribed by organizational roles, and spontaneously going beyond the roles to perform behaviors such as helping and cooperating with colleagues, protecting organizations from harm, defending organization's reputation, undertaking self-development, and so on (Katz & Kahn, 1966). Inspired by studies as such, Hurtz and Donovan (2000) explored the predictability of task and contextual performance by the Big Five personality traits, and discovered that the true validity coefficient of conscientiousness for task performance (ρ = .15) was lower than those for job dedication (ρ = .18) and interpersonal facilitation (ρ = .16), both of which according to Van Scotter and Motowidlo (1996) are facets of contextual performance. Hogan and Holland (2003) broke down overall job performance to getting ahead (i.e. task performance) and getting along (i.e. contextual performance), and observed that the estimated true validity of prudence (i.e. conscientiousness) for contextual performance was .31. This was higher compared to the validity for task performance (ρ = .20). Therefore, conscientiousness predicts task performance, and predicts contextual performance even better.

In I-O Psychology, sometimes organizational citizenship behavior (OCB) is used almost as an interchangeable concept of contextual performance, and similar to contextual performance, OCB, too, has been found to be predicted by conscientiousness. Organ and Ryan (1995) in their meta-analysis found that conscientiousness was the only variable significantly positively correlating with both the altruism component (directed towards individuals; corrected *r* = .22) and the generalized compliance component (directed towards organizations; corrected *r* = .30) of OCB. Miller, Griffin, and Hart (1999) noticed that conscientiousness was a valid predictor of OCB (*r* = .42) above and beyond neuroticism and extraversion. In comparisons to task performance, citizenship performance was more strongly associated with conscientiousness. For example, Motowidlo and Van Scotter (1994) obtained a correlation between the dependability facet of conscientiousness and citizenship performance at .31, and it was only .18 between dependability and task performance, though both were significant.

Conscientiousness and Other Important Criteria. Besides job-related outcomes, conscientiousness has been perceived as one of the most consequential trait for adaptive social functioning. It was found to be positively related to important life outcomes including marital stability (Tucker, Kressin, Spiro, & Ruscio, 1998), participation in healthy behaviors (Bogg & Roberts, 2004), and longevity (Friedman et al., 1993).

Also, according to Poropat (2009), conscientiousness is among the FFM dimensions the most closely related to academic performance (AP). Conscientiousness resembles the Webb's *w* factor (Webb, 1915), or as in Digman (1989), the willingness to achieve, both of which were found linked to AP (De Raad & Schouwenburg, 1996). Poropat (2009) pointed out that due to the link between Conscientiousness, sustain effort, and goal setting (Barrick, Mount, & Strauss, 1993), the dimension also contributed to a variety of AP-related behaviors, such as concentration on homework, following requirements (Trautwein, Ludtke, Schnyder, & Niggli, 2006), time management and effort regulation associated with learning (Bidjerano & Dali, 2007).

Given the importance of conscientiousness for predicting a variety of important academic, job, and life outcomes, the current study will focus on the conscientiousness dimension of the Big Five personality dimensions.

Narrow Facets of Conscientiousness. The studies mentioned above are all based on definitions and measurement models of conscientiousness that are somewhat different. Some of them used a measure of global conscientiousness while the others focused on narrow facets of the dimension, such as Achievement, Order, or Self Control. Paunonen (1998) reported that narrow trait measures of the Big Five predicted criteria better than the broad trait measures. Other studies comparing the broad and narrow measures drew similar conclusions (e.g. Ashton, 1998; Paunonen & Ashton, 2001). Roberts, Chernyshenko, Stark, and Goldberg (2005) found that all 6 underlying factors (Industriousness, Order, Self-control, Responsibility, Traditionalism, and Virtue) of conscientiousness had both differential predictive validity and incremental validity beyond the general factor of Conscientiousness when used to predict a variety of criteria, including work dedication, drug use, and health behaviors. Dudley, Orvis, Lebiecki, and Cortina (2006) found that narrow facets of conscientiousness (i.e., Achievement, Dependability, Order, and Cautiousness) had their unique strength for predicting various types of job performance (i.e., task performance, contextual performance, and counterproductive work behavior), even above and beyond global conscientiousness, but the magnitude depended on the particular type of criterion.

Avdic (2013) found that the broad dimension of conscientiousness predicted task performance, and in stepwise multiple regression analyses containing facets of conscientiousness as predictors of overall job performance, the competence facet emerged as the only meaningful predictor.

In 2013, Salgado, Moscoso, and Berges used a Schmid-Leiman transformation to partition the common variance in the facets of conscientiousness, and found that the narrow measures containing only specific variance in fact didn't predict job performance or have incremental validity above and beyond global conscientiousness. The findings are consistent with a couple of other studies such McManus and Kelly (1999), and Allen, Facteau, and Facteau (2004).

Whether the narrow facets or the broad dimension should be used for measuring conscientiousness is beyond the scope of the current study. However, the studies mentioned above have proved that conscientiousness, at both higher or lower-order, is a meaningful and consistent predictor of a number of crucial outcomes relating to school, work, family, and health, and thus the present study will focus on this construct. To be more specific, three of the most representative facets of this broad personality dimension will be measured, and they are industriousness, orderliness, and self-control. Industriousness is characterized by being hard-working and striving to achieve. Orderliness reflects the tendency to be organized, neat and tidy. Self-control, or impulse control, according to Peabody and De Raad (2002), reflects the propensity to be careful and controlled. These are the facets that keep emerging as stable facets of conscientiousness across studies (Roberts, Bogg, Walton, Chernyshenko, & Stark, 2004; Saucier & Ostendorf, 1999; Perugini & Gallucci, 1997; Peabody & De Raad, 2002). The facet of responsibility was thought to be another main facet of conscientiousness (Chernyshenko, 2003; Roberts et al., 2014), but factor analyses showed that it was a highly problematic facet on the CCS, especially when compared with its three fellows above. In both U.S. and U.K. samples, the Responsibility scale barely held up and quite a few items of this facet loaded onto industriousness and virtue (Green, O'Connor, Gartland & Roberts, 2016). The authors consequently concluded that responsibility failed to consistently emerge as a coherent factor.

The three facets included in the present study were not only stable and coherent factors of conscientiousness, but were also good predictors. For example, order and self-control had been found to be negatively correlated with a series of health and risk-related behaviors such as recent binge drinking, smoking, and overall risk behaviors (Green et al., 2016). Industriousness, on the other hand, had been constantly found to be a great predictor for performance-related criteria, including job performance, especially that of veteran employees (Stewart, 1999) and academic performance (MacCann, Duckworth, & Roberts, 2009).

Considering the fact that industriousness, orderliness, and self-control together represent the main characteristics of conscientiousness, consistently emerge as coherent factors, and have notable correlations with various outcomes, the current study focused on them as the most characteristic aspects of conscientiousness.

Curiosity

The most important reason why I included the curiosity facet in the current study was that it had been found by researchers to be the hardest facet/trait to write good intermediate items for

(Huang & Mead, 2014; Cao et al., 2015). Therefore, I believed curiosity measures would be useful for testing the strength of respondent training. Also, as a trait relating to knowledge acquisition, excitement to new experience, learning, and thinking (Mussel, Spengler, Litman, & Schuler, 2012), curiosity was found to be a predictor of overall job performance (e.g. Harrison, 2009; Reio & Callahan, 2004; Mussel, 2013), above and beyond 12 cognitive and noncognitive predictors (Mussel, 2013), though not across all positions (Mussel et al., 2012). In addition, people who were high in curiosity were found to be less sensitive to social rejection (Kawamoto, Ura, & Hiraki, 2017), suggesting that the trait could improve psychological and social functioning.

Considering the importance of curiosity for both item writing and job and life outcomes, it was included in this study.

Core Self-Evaluations

Judge, Locke, and Durham (1997) proposed an integrative personality trait termed Core self-evaluations (CSE), which was indicated by four well-established traits: (1) self-esteem: the overall value that one places on himself/herself (Harter, 1990); (2) generalized self-efficacy: the evaluation one has about oneself of how well he/she can perform across situations (Locke, McClear, & Knight, 1996); (3) neuroticism: the tendency to focus on negative aspects of the self and to have a negative cognitive or explanatory style (Watson, 2000); and (4) locus of control: beliefs that one can influence or control events and their outcomes (Rotter, 1966). Measured with the Core Self-Evaluations Scale (Judge, Erez, Bono, & Thoresen, 2003), the CSE was found to predict job satisfaction, job performance, and life satisfaction, above and beyond the five-factor model (Judge et al., 2003).

Being able to predict a variety of life and job outcomes, CSE was included in the current study.

Response Scale: Polytomous vs. Dichotomous

When a Likert-type response scale is used, and the sample size is relatively small (e.g. n = 300), some items may have a response category (e.g. “Disagree”) endorsed by too few participants to yield robust IRT estimates (Cao et al., 2015). Dichotomization, therefore, is sometimes considered a solution to such a problem, but the assumption that dichotomous scoring and polytomous scoring yield equivalent psychometric properties may not be valid. For example, Chernyshenko et al. (2007) suspected that the reason why they didn’t observe better criterion-related validity for the GGUM was that dichotomization of the responses may have reduced the amount of psychometric information, although Broadfoot (2008) later found that polytomous scoring didn’t improve validity. Cao et al. (2015) reported better model-fit for the dichotomized data (*n* = 375) than for the polytomous data (*n* = 811), while in a simulation study, Tay and colleagues (2011) found that polytomous IRT models (i.e. GRM and polytomous GGUM) in general yielded better model-data fit than the dichotomous models (i.e. 2PLM and dichotomous GGUM). Polytomous scoring was also found in simulated computerized adaptive test (CAT) on innovative items to have slightly better measurement precision (Jiao, Liu, Haynie, Woo, & Gorham, 2012), whereas in the same study with real data no difference was found between the two scoring methods. Vispoel and Kim (2014) reported that for the Balanced Inventory of Desirable Responding (Paulhus, 1991, 1999), a polytomous IRT model, GRM, provided consistently more precise estimates than dichotomous IRT models such as the 1PLM and the 2PLM, as well as its fellow polytomous model, the partial credit model (PCM). Given the mixed results in the literature, and the fact that both dichotomous and polytomous response scales were being used in research, I decided to include response scale format as an independent variable in the current study.

For the dichotomous condition, participants simply indicated if they agreed or disagreed with each of the items. For the polytomous condition, I used a 4-point Likert scale without a neutral option (i.e. “1= Strongly disagree”, “2 = Disagree”, “3 = Agree”, and “4 = Strongly agree”), as studies found that response scales having an odd-number of response options with a neutral option (e.g. Likert’s 1932 5-point rating scale) failed to work as intended for personality testing, regardless of the IRT approach being used. For example, when a dominance scale was used, due to the lack of intermediate items, the neutral option was included so that people with neutral attributes could endorse it (Likert, 1932; Kalton, Roberts, & Holt, 1980). However, in IRT analyses, researchers realized that the middle option was actually used by respondents as a “default” option when they didn’t want to select other options (Kulas, Stachowski, & Haynes, 2008). Hanisch (1992) reported that on the JDI, the “?” option was endorsed by respondents not to express neutrality but instead a negative sentiment. When mixed IRT models were used to analyze such data (e.g. Hernández, Drasgow, & González-Romá, 2004; Carter, Dalal, Lake, Lin, & Zickar, 2012), it was found that there was a class of respondents who appeared to use the middle to indicate confusion rather than the level of their job satisfaction. The middle option was also endorsed when respondents lacked strong, crystallized opinions (Presser & Schuman, 1980).

When an ideal-point approach is used, there simply is no need for a middle option reflecting neutrality, as ideal-point scales have intermediate items that will be endorsed by individuals with moderate trait levels. The use of a middle option could lead to poor fit of the GGUM, while four- and six-option response scales with no neutral option worked well with it (Dalal, Carter, & Lake, 2014).

Therefore, in the current study, I used a four-point Likert scale for the polytomous condition.

Respondent Training

There hadn't been a lot of studies where researchers trained or coached respondents to answer self-report items, but between the late 70s and the early 90s, a lot of studies on rater training were conducted where raters (e.g. supervisors, students etc.) were coached to evaluate the performance of others (e.g. employees, instructors etc.). An experimental design was usually adopted by this type of study, and observer ratings given by trained and untrained subjects were compared in terms of reliability, validity, accuracy, and error. In general, trained raters were found to have better performance with reduced errors such as halo, leniency, and contrast, compared with untrained raters (e.g. Borman, 1975; Latham, Wesley, & Pursell, 1975; Bernardin & Walter, 1977; Bernardin, 1978; Ivancevich, 1979; Bernardin & Pence, 1980; Pulakos, 1984), especially when the training was focused on avoiding errors (i.e. Rate Error Training, RET). On the other hand, rater training that aimed specifically to improve rating accuracy (i.e. Rater Accuracy Training, RAT) led to more accurate ratings. Interrater reliability was also found to be higher among the trained raters (e.g. Bernardin & Walter, 1977; Shohamy, Gordon, & Kraemer, 1992; Kramer, de Roten, & Drapeau, 2011). Although RET had been believed by some researchers to have harmed accuracy (e.g. Hedge & Kavanagh, 1988; Bernardin & Pence, 1980), the meta-analytic study conducted by Woehr and Huffcutt (1994) concluded that RET actually resulted in a modest increase in the measuring accuracy.

The finding that rater training normally led to better observer ratings got me wondering if training effects also existed on respondents of self-report personality tests. To train respondents to provide self-evaluations in a more consistent manner, I concentrated on these three aspects:

1. ***The Ideal-Point Response Process.*** I asked the participant in the training group that when answering to an item, they read the statement, think about themselves (i.e. engage in

introspection), and decide if the statement is closely describing them. I emphasized that they only agree with a statement if they believed that it was an accurate description of them. The instruction was written in layperson's language to avoid confusion.

2. ***The "Agree Only When Both Apply" Rule.*** I reminded participants that they might encounter items that partially applied to them, which they should disagree with. Brown and Maydeu-Olivares (2010) were concerned that an item that didn't completely apply might cause confusion for the participant, who might get frustrated and give random responses. Therefore, in this study, I told the participants what to do in a situation like this, so that frustration and random responses would be less likely to occur.
3. ***Disagree from Both Directions.*** The ideal way for an intermediate item to work is that only people of medium trait levels will endorse it, and the others will disagree either because their trait levels are too low (i.e. disagree from below) or because their trait levels are too high (i.e. disagree from above). Knowing when to disagree with an intermediate item is as important as knowing when to agree with it. Therefore, I, with the help of several examples, explained to participants what it meant to agree and disagree with each of them, so that the participants were able to confidently answer items later without having to scratch their heads.

To summarize, in the current study, respondent training and response format were the two between-group independent variables, leading to a 2 X 2 design. Within each of the four groups, three facets of conscientiousness, industriousness, order, and self-control, the curiosity facet of openness, as well as CSE were measured. Each of these 4 Big-Five personality narrow facets was measured with two types of personality measures. These instruments were developed with different approaches (i.e. dominance vs. ideal-point) and amounts of nonmonotonicity, considering the findings from previous studies that the approach (i.e. dominance vs. ideal-point) used for developing a scale and the amount of unfolding on a scale were both related to the performance of IRT models. Life satisfaction, counterproductive work behavior (CWB), health behavior, and academic performance were measured as external criteria.

Hypotheses and Research Questions

Since training should improve participants' understanding of the personality items, there should be more well-preforming intermediate items in the trained groups than in the untrained groups. Based primarily on this premise, this study tested and explored the hypotheses and research questions listed below.

Model-Data Fit

Hypothesis 1a: GGUM will have better fit in the trained group than in the untrained group, given that the other conditions are the same.

Hypothesis 1b: 2PL and SGR will have worse fit in the trained group than in the untrained group, given that the other conditions are the same.

Hypothesis 1c: For dominance scales, fit of GGUM will be no worse than that of the dominance models, given that the other conditions are the same.

Hypothesis 1d: For ideal-point scales, fit of GGUM will be better than that of the dominance models, given that the other conditions are the same.

Research Question 1: How will the number of response categories affect model fit?

Criterion-Related Validity

Hypothesis 2a: In the trained group, GGUM will have higher criterion-related validity than the dominance models for the ideal-point measures, given that the other conditions are the same.

Hypothesis 2b: There will be no difference in criterion-related validity in all the other conditions.

Research Question 2: Will the number of response categories have an effect on the criterion-related validity, given that the other conditions are the same?

Intermediate Items

Hypothesis 3a: More items will turn out to be intermediate in the trained group than in the untrained group when the measures are dominance, given that the other conditions are the same.

Hypothesis 3b: More items will turn out to be intermediate in the trained group than in the untrained group for the ideal-point measures, given that the other conditions are the same.

Research Question 3: How will the number of response categories affect the number of functioning intermediate items on measures?

CHAPTER 3: METHODOLOGY

Study Design

The design of the study is illustrated in Table 1. This 2 X 2 design generated four between-subject groups that differed from each other in the conditions of training (i.e. training vs. no training) and response scale (i.e. 4-point vs. 2-point).

Participants

A total of 2437 Amazon Mechanical Turk (MTurk) workers participated in this study, and they were randomly assigned to one of the 4 groups. Among these participants, 443 were dropped for failing to answer all 3 quality control questions right, and 17 were dropped for giving invariant responses on one or multiple scales. Therefore, I ended up with 1977 participants in total (valid response rate = 81.12%). Demographic information for the sample can be found in Table 2.

Measures

Personality

IPIP. The 300-item International Personality Item Pool (IPIP; Goldberg et al., 2006) based on Costa and McCrae (1992) NEO-PI-R facets has 300 items for 30 facets of the Big Five factors. In this study, industriousness, order, self-control, and curiosity were measured by the 10-item Achievement-Striving, Orderliness, Cautiousness, and Adventurousness scales, respectively. Information regarding measure reliability for each of the 4 groups can be found in Table 3.

CPS and More Intermediate Items. The Comprehensive Personality Scale (CPS) is a result of years of work in Dr. Fritz Drasgow's lab, and it was developed using the ideal-point approach (Wang, 2013). The CPS consists of 440 items that cover a full set of 22 personality facets derived from the Big-Five model. In terms of item extremity, each facet has approximately equal numbers of statements reflecting high, medium, and low trait levels (Wang, 2013). I also combined the CPS items with intermediate items written by Cao and colleagues (2015), hoping to maximize the training effect by including more intermediate items. Since Cao et al. (2015) didn't measure industriousness or self-control, these two facets were measured by the CPS alone. The number of items for each of the 4 CPS scales was 20, and Cao et al. (2015) developed 8 intermediate items for the Curiosity scale and 9 for the Order scale.

CSES. The Core Self-Evaluations Scale (CSES; Judge et al., 2003) was used to measure core self-evaluations. The CSES contains 12 items, and was developed with the dominance approach (Judge et al., 2003). Although the CSES was developed with a 5-point Likert-type scale ranging from "Strongly disagree" to "Strongly agree", with a middle option of "Neutral", in the current study, participants were presented with either a 4-point (1 = "Strongly disagree", 2 = "Disagree", 3 = "Agree", 4 = "Strongly agree") or a dichotomous "Disagree - Agree" response scale, depending on which group they were assigned to. Reliability of the measure can be found in Table 3.

Criteria

Information regarding the reliability of the criterion measures (except for Academic Performance) are listed in Table 3, and items on the criterion measures can be found in Appendix B.

SWLS. I measured participants' life satisfaction using the Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985). This is a heavily used 5-item scale with a 7-point response scale ranging from "Strongly disagree" to "Strongly agree", including a neutral option "Neither agree nor disagree".

Academic Performance (AP). Participants' academic performance (AP) was assessed with one item asking about their performance at school (i.e. "How do/did you do in school"). The participants provided their answers on a 7-point response scale with no neutral option (i.e. 1 = "Very poorly", 2 = "Poorly", 3 = "Slightly below average", 4 = "Average", 5 = "Slightly above average", 6 = "Well", and 7 = "Very well").

Counterproductive Work Behavior (CWB). Participants were first asked whether they were employed at the time they participated in the study, and if they were, they would be asked to fill out the 10-item CWB checklist (Spector, Bauer, & Fox, 2010). Participants were asked to indicate how frequently they performed at their current jobs a variety of counterproductive work behaviors. The 10-item CWB measure uses a 5-point response scale (i.e. 1 = "Never", 2 = "Once or twice", 3 = "Once or twice/month", 4 = "Once or twice/week", 5 = "Every day").

The Health Behavior Checklist (HBCL). I assessed participants' health-related behaviors with the Preventive Health Behaviors (PHB) and part of the Risk Taking Behavior (RTB) subscales of the HBCL (Vickers, Conway, & Hervig, 1990). The PHB scale consists of 16 items, 10 of which focus on wellness maintenance and enhancement (WME), and 6 on accident control

(AC). The 7-item Traffic Risk (TR) scale was used to measure risk taking behavior.

Training Feedback

Right after participants finished the training, I presented them with 2 questions asking for their feedback on the training (see Table 4). The first question asked how often they had been confused with items similar to the ones in the training, and the second asked if they had learned anything new from the training.

Response Time

The time each participant spent finishing the IPIP measures, the CSES, and the CPS measures was recorded, respectively.

Procedure

Training

The participants assigned to the two training groups (i.e. Groups 3 and 4) were first presented with a flowchart illustrating the ideal-point response process I would like them to adopt when responding to an item: (1) read the statement; (2) think about yourself; (3) compare yourself to the statement; (4) if **every** part of the statement applies to you, agree or strongly agree; (5) if not, disagree or strongly disagree.

Then, in order to make sure that participants paid attention to and understood the flowchart, I presented them with each of the 5 steps of the process in a random order, and asked them to put those steps in the correct order as shown in the flowchart. Participants had to answer this question correctly to move on, and they were allowed to go back to the flowchart when they were working on the ordering task.

After participants had passed the ordering task, they were asked to work on an intermediate item practice containing 3 example intermediate items of different types. These 3 example items were presented one at a time to the participants and a flowchart was used to explain when this item should be agreed with (i.e. when the item applied completely) or disagreed with (i.e. when the item didn't apply at all or part of it didn't). Following each example item, a similar intermediate item was presented as a test question, followed by 2-3 vignettes of made-up characters, and respondents were asked to respond to the test item for the made-up characters based on the vignettes. This was to test if the respondents had studied and understood the example. If the respondents managed to choose the correct answers for all the made-up characters, they would pass, otherwise, they would be provided with an explanation on why their responses were incorrect and they were asked to try again. Respondents needed to pass the practice questions for all 3 examples in order to take the official surveys. This was how I ensured that participants in the training group spent time learning the materials rather than skipping them and taking the survey directly. Details of the training for the polytomous group can be found in Appendix A. The only difference between training in the polytomous group and the dichotomous group was that all response categories (in flowcharts, instructions, and items) were changed accordingly.

No training

Participants assigned to the two non-training groups only took the surveys and received no additional training.

Dichotomous

Participants in the two dichotomous groups answered the personality surveys using a "Disagree-Agree" scale.

Polytomous

Participants in the two polytomous groups answered the personality items using a 4-point response scale (i.e. "1= Strongly disagree", "2 = Disagree", "3 = Agree", and "4 = Strongly agree").

Measures

Participants in all 4 groups were measured on all 5 personality traits (i.e. curiosity, industriousness, order, self-control, and core self-evaluation) and 4 external criteria (i.e. life satisfaction, academic performance, CWB, and health-related behaviors) with the 7 instruments mentioned above.

Randomization

Participants were randomly assigned to one of the four groups. Personality measures were presented in a random order, and so were the outcome measures. Within each scale, items were also randomly presented.

Analyses

Reliability

As shown in Table 3, all dominance measures and criterion measures had acceptable Cronbach's alpha (i.e. ⩾ 0.7) except for the Accident Control scale of the HBCL for the untrained group using the polytomous response scale, the reliability of which was very slightly below 0.7 (i.e. 0.658). Considering that this scale showed acceptable reliability in all the other 3 groups, we decided to keep this measure in our analyses for all groups.

Scales and Items

Since there were 11 personality scales included in the analyses, I gave them each a shorter name, and will use them in the rest of this paper:

CSES: The Core Self-Evaluation Scale.

IPIP_Cur: The Adventurousness scale of the IPIP measuring curiosity.

IPIP_Ind: The Achievement-Striving scale of the IPIP measuring industriousness.

IPIP_Ord: The Orderliness scale of the IPIP measuring order.

IPIP_SC: The Cautiousness scale of the IPIP measuring self-control.

CPS_Cur20: The 20-item Curiosity scale of the CPS.

CPS_Ind20: The 20-item Industriousness scale of the CPS.

CPS_Ord20: The 20-item Order scale of the CPS.

CPS_SC20: The 20-item Self-Control scale of the CPS.

Cur28: The 20-item CPS Curiosity scale combine with the 8 intermediate items developed by Cao et al. (2015).

Ord29: The 20-item CPS Order scale combine with the 9 intermediate items developed by Cao et al. (2015).

In Group 2 (i.e. untrained and polytomous), response category 1 (i.e. "Strongly Disagree") was not selected by any participant for Item 3 of the IPIP_Cur scale (i.e. "*Am interested in many things*"), which was a problem for running the IRT software (esp. GGUM2004), so the item was dropped from all further analyses from all groups. This was also to ensure that exactly the same materials were included in the analyses for all 4 groups.

Unidimensionality

Since both dominance and ideal-point IRT models assume unidimensionality, I conducted exploratory factor analysis (EFA) with principal axis factoring. Results showed that in general unidimensionality held for most of the scales, with a few exceptions concentrating in the two dichotomous groups (i.e. Groups 1 and 3), where the total variance explained by the first factor were slightly below the 20% cutoff (Reckase, 1979), ranging from 15.16% to 19.97. However, sometimes when a response scale is dichotomous, and there is a large proportion of 1's, it's possible to have a smaller first eigenvalue but at the same time data satisfy unidimensionality

(Drasgow, personal communication, Aug 20, 2018). Therefore, I decided to run IRT analyses with all data.

Model Fit

Since GGUM2004 (Roberts & Shim, 2008) failed to converge for all the polytomous CPS-related data, I had to exclude them from the analyses. Therefore, I obtained both dichotomous and polytomous GGUM item and person parameter estimates for all the IPIP data, while only the dichotomous GGUM estimates for the CPS-related data were obtained. No reverse coding was needed for GGUM estimation.

After negatively-worded items were reversed in SPSS, I obtained item and person estimates with MULTILOG 7.0 (Thissen, Chen, & Bock, 2003) for the dominance models.

With the estimated item parameters and the item responses, I ran model fit analyses in MODFIT (Stark, 2007) for both the dominance models and GGUM. MODFIT is an Excel macro developed specifically for analyzing IRT model fit. Adequate fit was indicated by Chi-square-to-degree-of-freedom ratios less than 3 (Tay, et al., 2011).

Criterion-Related Validity

I correlated the estimated person parameters obtained with different models with participants' scores on the outcome measures. I also compared the criterion-related validity under different conditions to examine the effects of training, number of response categories, and the IRT model applied. A Bonferroni correction was used to control the family-wise error rate (FWER) induced by performing multiple hypotheses tests.

Intermediate Items

The number of good intermediate items is an important indicator of the training effect, so I looked at the GGUM α- (i.e. the discrimination parameter) and δ-parameters (i.e. the location parameters) of each of the items, and an item with an acceptable α would be considered an intermediate item if its δ fell between the 10^{th} and 90^{th} percentile of the estimated theta distribution (Roberts & Shim, 2008). I compared between the untrained and the trained group the numbers of intermediate items, and how many of them were actually developed to be intermediate. I also compared the means of the α-parameters of the intermediate items in the trained and untrained groups. The larger the mean α-parameters, the more informative the set of items were on average. Mean discrimination along with the number of good intermediate items were used to examine the effectiveness of training on item responding.

In addition to item parameters, I also referred to the GGUM ICCs to examine more intuitively if an item was displaying different properties in the trained and the untrained groups.

Response Time

The mean time participants spent on finishing the IPIP measures, CPS measures, and the CSES was compared across conditions.

CHAPTER 4: RESULTS

Model Fit

Estimated Item Parameters

For all 4 groups, estimated item parameters for each of the 11 scales mentioned in Chapter 3 can be found in Tables 5-26. When GGUM was fitted to the data, the scale developing method didn't seem to matter, and most items had discrimination parameters larger than 0.5. The only exception was Item 13 of CPS_Ind20 (i.e. "***There is too much to be done to waste time relaxing***"), which had discrimination parameters slightly below 0.4 in both the trained and untrained group.

When the dominance models were fitted to the data obtained using the dominance measures (i.e. the 4 IPIP scales and the SCES), most of the items turned out to be discriminating enough (i.e. with alphas larger than 0.51). However, when 2PL and SGR were fitted to CPS-related data, many intermediately-worded items turned out to be barely discriminating as expected.

Model Fit Comparison

Model fit results can be found in Tables 27 and 28. Chi-square-to-degree-of-freedom ratios are reported for item singles, doubles and triples. In the current study, I focused on the fit of item doubles and triples. This is because item singles are insensitive to misfit when item parameters and fit are computed using the same sample (Drasgow, Levine, Tsien, William, & Mead, 1995). Also, when there was misfit for more than one model, relative misfit of the two models were compared (Stark et al., 2006).

Dominance Measures. As shown in Table 27, both GGUM and the dominance model showed much better fit for dichotomous data than polytomous data. With dichotomous data, GGUM showed either similar or better fit than 2PL, in both the trained and untrained groups, which was consistent with previous findings (e.g., Chernyshenko et al., 2001).

Also, compared to the untrained group, GGUM showed similarly good fit in the trained group for IPIP_Cur, IPIP_Ind, and IPIP_SC, slightly worse fit for CSES, and slightly better fit for IPIP_Ord.

On the other hand, 2PL in the trained group showed either worse fit (for CSES and IPIP_Cur) or similar fit (for IPIP_Ind, IPIP_Ord, and IPIP_SC), compared to the untrained group. For IPIP_Cur, 2PL had much worse fit in the trained group, while the fit for GGUM remained equally good.

However, when the response scale was polytomous, a more consistent pattern was observed that GGUM fitted worse than SGR in both the trained and untrained groups, and fit in the trained group was worse than that in the untrained group, regardless of the model used.

Ideal-Point Measures. As shown in Table 28, when measures were developed under the ideal-point method, 2PL consistently had fit that was much worse than GGUM, whether participants were trained or not.

Compared to the untrained group, in the trained group, 2PL had similar fit for Ord29, slight better fit for CPS_Cur20 and CPS_SC20, and worse fit for the other scales. GGUM, on the other hand, showed worse fit for CPS_Ind20, CPS_Ord20, and CPS_SC20, and similar fit for the rest.

Summary

In general, fit of GGUM stayed the same or got worse with training, so Hypothesis 1a was not supported.

Fit of the 2PL was better with training for CPS_Cur20, CPS_SC20, and fit of the SGR was better with training for CPS_Ind20, while for the other personality measures, compared with the untrained group, fit of the dominance models (i.e. 2PL and SGR) in the trained group was either the same or worse, so Hypothesis 1b was partially supported.

Hypothesis 1c was partially supported: when the response scale was dichotomous, GGUM had similar or better fit than 2PL for the dominance measure. However, when the response scale was polytomous, with the dominance scales, GGUM always had poorer fit than SGR. In fact, both models had bad fit, but GGUM fitted even worse.

I was unable to run polytomous GGUM with the ideal-point measures, but according to the results obtained with the dichotomous data, Hypothesis 1d was fully supported that GGUM fitted better than the dominance model (i.e. 2PL) for data obtained with ideal-point measures.

As for Research Question 1, I concluded that for the dominance measures, both GGUM and the dominance models had better fit for dichotomous data than for polytomous data. With ideal-point measures, fit of the dominance models was generally better when the response scale was dichotomous, except for CPS_SC20 and Ord29 without training.

Intermediate items

A good intermediate item, in the current study, would be characterized as having (a) an acceptable discrimination parameter (i.e. $\alpha \geqslant 0.5$), and (b) a location parameter (i.e. δ) falling between the 10th and 90th percentile of the estimated theta distribution (Roberts & Shim, 2008).

Dominance Measures

According to Chernyshenko et al. (2001), when GGUM was applied to dominance measures containing extreme items only, some items would end up having unfolding ICCs, one of the key features of intermediate items. In Tables 5-9, intermediate items have been identified by a superscript letter "I" by the location parameters based on Rules (a) and (b) mentioned above. When the response scale was dichotomous, I found that out of the 51 items that were written to be extreme, 5 of them turned out intermediate in both groups, 7 of them were intermediate in the trained group but not the untrained group, and 6 of them were intermediate in the untrained group only. This finding of extreme items turning out to be unfolding was consistent with what Chernyshenko et al. (2001) discovered. No intermediate items were observed when the polytomous scale was used. Training didn't seem to affect the extremity of items, as we observed almost identical numbers of intermediate items in the trained and untrained groups.

Ideal-Point Measures

Since some of the items on the ideal-point measures were written to measure moderate trait levels, I believed that compared to the dominance measures, the CPS-related measures containing a nontrivial amount of intermediate items were the better materials for examining the effects of training on response behaviors. The ideal-point measures I examined were Cur28, CPS_Ind20, Ord29, and CPS_SC20. I adopted the longer version of the Curiosity and Order scales because they contained more intermediate items, which allowed me to more clearly examine the effects of training, if there was any.

Estimated Item Parameters. Intermediate items again were identified based on Rules (a) and (b). As shown in Tables 15-18, out of the 97 items, 29 (i.e. 29.90%) turned out to be intermediate in both the trained and untrained dichotomous groups, 3 (i.e. 3.09%) were intermediate in the untrained group only, and 24 (i.e. 24.74%) were intermediate items in the trained group only. Moreover, among the 24 intermediate items unique to the trained group, I recognized 18 well-performed intermediate items (6 "Frequency", 6 "Condition", 2 "Transition", 1 "Average", 1 "Frequency + Condition", 1 "Frequency + Double-barreled", 1 "Double-barreled"), and 6 seemingly extreme items. Considering that the training covered the "Double-barreled" type (in the training Introduction part), the "Frequency" type (Example 3), the "Condition" type (Example 1), and the "Average" type (Example 2), I believed that it was the training that led to the presence of overwhelmingly more well-performed intermediate items in the trained group than in the untrained group.

The 3 items that were intermediate only in the untrained group all belonged to the "Frequency" domain.

I also ran analyses using only the 4 original CPS measures (i.e. CPS_Cur20, CPS_Ind20, CPS_Ord20, and CPS_SC20), and results (see Tables 17-20) were similar considering the reduced set of intermediate items: out of 80 items, 21 (i.e. 26.25%) turned out intermediate only in the trained group, 3 (i.e. 3.75%) in the untrained group only, and 19 (i.e. 23.75%) in both.

Based on Rule (a), I compared the means of the discrimination parameters of different sets of intermediate items between the trained and untrained groups, and detailed results can be found in Table 39. As shown in the table, mean alpha values were higher in the trained group than in the untrained group in 25 out of the 30 (i.e. 83.33%) comparisons, and the difference was larger than 0.1 in 14 out of the 25 cases.

Visual Aids. In addition to item parameter estimates, I also obtained GGUM ICCs of intermediate items unique to either the trained or the untrained group (see Figures 1-27). For some items, the shapes of the ICCs differed substantially (e.g. Figures 7, 11, 14, and 19), while for some items, the difference was less obvious (e.g. Figures 18 and 23).

Summary

With dominance measure, I didn't observe the training effect on item responses, so Hypothesis 3a was not supported.

However, with ideal-point measures consisting of both extreme and intermediate items, I found that a lot more items turned out to be intermediate in the trained group than in the untrained group. Also, intermediate items in the trained group were on average more discriminating than in the untrained group. Thus, Hypothesis 3b was supported.

As to Research Question 3, I believed that the answer would be that when measures were dominance, intermediate items were found only with a dichotomous scale, but not a polytomous one. I didn't have any results with the ideal-point measure, as GGUM2004 wouldn't converge when the response scale was polytomous.

Training Feedback

Table 4 contains the results of the 2 feedback questions in the trained group using the dichotomous response scale. According to Table 9, 56.3% of all participants reported being confused at least sometimes about items similar to the ones in the training. 40.7% reported having barely or very rarely seen such question, and only 2.6% reported that they had seen similar questions and had the same understanding as explained in the training.

When asked if they had learned anything new from the training, most participants (45.7%) reported their knowledge about survey responding becoming more systematic instead of learning new things. About 30% of all participants reported having their major confusing removed by the training, followed by 15.8% reporting that the training had answered some of their questions. Only approximately 9% of all participants reported that they had already known everything covered in the training.

Mean Response of the Ideal-point Measure Items

For each of the four groups, the mean of the responses for each of the ideal-point measure items used in the current study can be found in APPENDIX C. In general, the means did not differ between the trained and the untrained groups. The average absolute difference between the trained and the untrained group for all ideal-point measure items was 0.04 when the response scale was dichotomous and 0.05 when the response scale was polytomous.

Criterion-Related validity

Individual Correlations

Dominance Measures. Tables 29-33 contain criterion-related validity results for all dominance measures in all 4 groups. According to the significant test, curiosity in general was the weakest predictor across outcome variables, groups, and IRT models. It barely predicted CWB, WME, and AC, and it was the only trait that negatively related to TR.

CSE and industriousness both predicted all outcomes except for TR. CSE seemed to be an excellent predictor of life satisfaction and a good one of CWB, while industriousness turned out to be the mostly strongly associated with AP.

Self-control was the strongest predictor of TR and CWB: the higher the level of self-control, the less likely someone would engage in CWB or risk-taking behaviors on the road. Order was the best predictor across all 4 groups, 6 outcome variables, and 2 IRT models, with all but only 4 non-significant correlations. It was not the strongest predictor of any outcome among the 5 traits, but it seemed to predict everything.

Ideal-Point Measures. Criterion-related validity results obtained with the CPS-related measures can be found in Tables 34-37. In general, order and self-control turned out to be better predictors than curiosity and industriousness. Curiosity seemed to be the best predictor of all for academic performance, which industriousness failed to predict consistently across groups and IRT models. Order and self-control both consistently predicted life satisfaction, WME, and AC across groups and models.

Correlation Comparisons

I also compared correlations across different training conditions (i.e. trained vs. untrained), numbers of response categories (i.e. dichotomous vs. polytomous), and the IRT model applied (i.e. dominance vs. GGUM).

I tested 612 hypotheses in total for all measures and groups, so the corrected α = 0.05/612 = 0.0000817. Tables 38 contains detailed results of the comparisons where the differences were significant. I found out that out of the 612 pairs of correlations that were compared, only 14 were significant, which was 2.29%, suggesting that in general none of the factors examined had an effect on criterion-related validity.

Mean Correlations. I also examined the trends of criterion-related validity across experimental conditions (i.e. training, response scale, and model) by comparing the mean criterion-related validity (see Table 29-37) without statistically testing the differences. It turned out that when personality measures were dominance, the dominance models and GGUM yielded very similar mean validity. The number of response categories and training generally didn’t seem to affect mean validity, but there were a few exceptions, especially between IPIP_Ind and SWLS, where training and the polytomous scale were both associated with lower mean validity. In addition, compared to the polytomous scale, much larger correlations were observed with the dichotomous scale between IPIP_SC and the CWB and Risk Taking scales, as well as between CSES and the Accident Control scale.

As to the ideal-point personality measures, I found that when GGUM was applied, in general higher criterion-related validity was observed than when dominance models were used. Training, on the other hand, did not seem to influence the mean correlations. Since I was not able to apply GGUM to the polytomous data, I couldn’t compare mean correlations obtained using different response scales for the ideal-point measures.

Additionally, criterion-related validity was averaged across all personality measures and outcome measures and compared between the trained and untrained groups, the dichotomous and polytomous groups, and the GGUM and the dominance models. As can be found in Table 41, between different conditions, correlations were very similar averaged over all personality dimensions and outcome measures.

Summary

I found that criterion-related validity was largely unaffected by training, response scale, or the model applied. With the small number of comparisons of correlations that were significant after a Bonferroni correction, I failed to find any consistent pattern, and I was not certain if these significant results would be replicated. When mean correlations were compared without statistical tests, I found that GGUM was generally associated with higher mean validity than the dominance models for ideal-point measures.

Therefore, Hypothesis 2a was not supported, while Hypothesis 2b was. The answer to Research Question 2 would be "No, the number of response categories did not seem to affect criterion-related validity, given that the other conditions were the same."

Mean Response Time

The average response time for the CPS measures, the IPIP measures, and the CSES for different conditions can be found in Table 40. When the response scale was dichotomous, the trained participants on average spent approximately the same time on the IPIP measures and the CSES as the untrained participants, while 32 seconds less on the CPS measures and the extra intermediate items. When the response scale was polytomous, the pattern was less obvious: compared to the untrained participants, the trained participants spent approximately the same time on the IPIP measures, 8 seconds shorter on the CPS measure, and 3 seconds more on the CSES.

CHAPTER 5: DISCUSSION

The Training Effects

Intermediate Items

The most important finding of the current study is that when trained about ideal-point response process and how to interpret common types of intermediate items, participants responded to ideal-point measure items differently from those who were untrained, leading to more items (esp. intermediate items) turning out to be intermediate items that were more discriminating on average. Since I was unable to run GGUM with polytomous data obtained with the ideal-point measures, this finding is limited to dichotomous data.

This finding converged with some of the feedback from participants regarding personality items and the training. First of all, over half of the participants reported being confused by personality items similar to the ones in the training, suggesting that confusion among participants about personality items was real, and that the confusion was nontrivial, as many participants were aware of it. For the other 40% who had barely seen such questions, I believed it was partially due to the fact that some of them were new MTurk workers who had participated in only a few studies before.

Second, a lot of the trained participants reported either their knowledge becoming more systematic (45.7%) or confusion being removed more or less (45.6%), which was consistent with what was reflected by their personality survey responses, especially when compared to the untrained: The trained participants seemed to interpret intermediate items in a way that was closer to the expectations of the developers of the ideal-point measures.

Participants' responses to these two feedback questions suggested that confusion did exist among participants, and that the training did a good job targeting such uncertainty. My supposition upon which this project was built has been supported: Participants are not as knowledgeable as researchers about the items, and their interpretation of some intermediate items can be different from ours. By providing a training session explaining how different types of intermediate items were expected to be interpreted and emphasizing the ideal-point response process, I observed more intermediate items working as expected than in the untrained group.

Interestingly, no such training effect was observed with the dominance measures (i.e. IPIP measures and CSES), probably suggesting that the training about intermediate item interpretation (i.e. the 3 examples and tests) was the more effective part compared to the training on the ideal-point response process. Approximately equal numbers of items turned out to be uniquely intermediate in the trained group and the untrained group, which was consistent with what Chernyshenko and colleagues (2001) found when applying GGUM to the IPIP measures.

Model-Data Fit

Training seemed to have a much less obvious effect on model-data fit. It led to worse fit of the dominance models for dominance measures and some ideal-point measures, and similar or worse fit of GGUM for both types of measures. Training was also associated with better dominance model fit for some of the ideal-point measures. In general, training didn't help with the fit of GGUM as hypothesized, and I was not quite sure why. My guess was that since with the dichotomous data, GGUM already had pretty good fit without training, it was probably hard for GGUM fit in the trained group to top that.

Criterion-Related Validity

Hypotheses regarding training leading to higher criterion-related validity were not supported either. Criterion-related validity in general was unaffected by training, which was not a surprise. Criterion-related validity of self-report personality measures has never been found to be easily influenced by factors such as the model applied, the method used to develop the measure, or the response scale used. However, I found that ideal-point measures had a tendency of having higher mean criterion-related validity when GGUM was applied than when dominance models were used. This trend we found was consistent with what Cao et al. (2015) reported: GGUM yielded higher predictive validity when a measure was developed with the ideal-point method containing several intermediate items.

Mean Response Time

When the response scale was dichotomous, not only did training relate to a lot more well-performing intermediate items, but also to quicker responses to the ideal-point measures. This finding was consistent with what was reported by the participants: the training helped with item interpretation and responding.

Dichotomous vs Polytomous

Model-Data Fit

For the dominance measures, model-data fit was always better when the response scale was dichotomous, and for the ideal-point measures, dichotomous scale was also associated with better fit of the dominance models in general. This is consistent with what was found in Cao et al. (2015), which is also an empirical study, but is inconsistent with Tay et al. (2011), a simulation study where better fit was observed with polytomous IRT models.

I was unable to run any analyses with polytomous GGUM, as GGUM2004 wouldn't converge. I tried dropping some of the poorly discriminating items in order for GGUM2004 to run, but there were very few items that I could justify dropping, and GGUM2004 still couldn't converge even after the removal.

I think that dichotomous data is more software-friendly, especially when GGUM is used. However, more studies, especially empirical studies, are needed to see if the pattern can be replicated that when data are dichotomous, IRT models fit better and less hassles occur than when polytomous data are analyzed.

Criterion-Related Validity

Same as training, the number of response categories didn't appear to make a difference to criterion-related validity. However, among the 14 significantly different comparisons of criterion-related validity, 8 had to do with the response scale, and in 7 of them the polytomous scale was associated with higher correlations with the outcome variables. Perhaps this was because polytomous data contained more information than the dichotomous data (Chernyshenko et al., 2001). Still, this number is trivial compared to the 612 comparisons conducted, so more evidence is needed before any conclusion could be drawn regarding the effects of response scales on criterion-related validity.

Construct Validity

With training, when a dichotomous response scale was used and the GGUM was applied, the ideal-point measures had more well-performing and more discriminating and informative intermediate items than the untrained group, indicating better construct validity.

Limitations and Future Research

First of all, all participants of the current study were MTurk workers, which limited the generalizability of the conclusions. Also, the survey taking experience may have made these MTurk workers more resistant to the training. Therefore, In the future, samples containing a wider range of participants, especially those who are less experienced in survey responding (e.g. Freshmen during their first few weeks in college) should be used to improve generalizability.

Secondly, compared to the multi-session, face-to-face rater training, by which I was inspired to conduct the current study, I felt that our single-session online training with only words and flowcharts might not have been as powerful. In the future, researchers should consider giving the respondent training in the same way as rater training to see if the strength of training will improve.

Thirdly, now that respondent training has been shown to be effective for the non-adaptive online personality survey responding, in the future, researchers should consider applying the training to the adaptive testing environment.

Lastly, GGUM2004 failed to run properly with polytomous data obtained with ideal-point measures, so I think it's time that scientists considered developing GGUM software that is able to handle different types of data more stably, and perhaps adopts alternative methods such as the Bayesian estimation (Wang, 2013), instead of the maximum likelihood estimation used by GGUM2004.

Conclusion

This study has proved that a knowledge gap exists between researchers and participants regarding self-report personality items, especially the intermediate items, should be processed, interpreted, and responded to. There are things about personality items that participants don't understand entirely but researchers have long been assuming that they do. With a short online training session, this gap can be removed, indicated by the positive participants' feedback, less mean response time for the ideal-point measures, and more well-performing and more discriminating and informative intermediate items.

TABLES

Table 1. Experiment design

	No Training	**Training**
Dichotomous	Group 1	Group 3
Polytomous	Group 2	Group 4

Table 2. Information for Samples (by Group)

	N	Emp.*N*	Female (% of N)	Mean Age	SD Age	Racial Makeup	Education
Group1	490	340	63.7%	38.87	12.98	8.8% African American	11.2% High School or lower
						3.9% Asian	31.3% Some College
						6.3% Hispanic/Latino	37.4% B.A. Degree
						79.8% White	3.3% Some Graduate School
						1.2% Other	13.3% Master's Degree
							3.5% Doctoral Degree
Group2	495	349	62.6%	38.39	11.90	8.5% African American	8.3% High School or lower
						4.7% Asian	31.7% Some College
						6.5% Hispanic/Latino	36.3% B.A. Degree
						78.7% White	4.5% Some Graduate School
						1.6% Other	16.2% Master's Degree
							3.0% Doctoral Degree
Group3	494	343	65.1%	38.76	13.26	11.6% African American	10.7% High School or lower
						7.1% Asian	34.6% Some College
						5.1% Hispanic/Latino	34.8% B.A. Degree
						75.5% White	3.2% Some Graduate School
						0.7% Other	15.0% Master's Degree
							1.6% Doctoral Degree
Group4	498	348	59.8%	37.97	12.39	8.7% African American	10.7% High School or lower
						7.3% Asian	34.6% Some College
						5.2% Hispanic/Latino	34.8% B.A. Degree
						77.8% White	3.2% Some Graduate School
						1.0% Other	15.0% Master's Degree
							1.6% Doctoral Degree

Note: *N*: the sample sizes of correlations involving all criteria except for CWB; Emp.*N*: the sample sizes of correlations involving CWB; SD Age: the standard deviation of age.

Table 3. Cronbach's alpha for the dominance measures and the criterion a measures.

	Untrained & Dichotomous	**Trained & Dichotomous**	**Untrained & Polytomous**	**Trained & Polytomous**
1. IPIP_Ord	.86	.84	.86	.87
2. IPIP_Ind	.77	.80	.83	.85
3. IPIP_SC	.83	.80	.84	.85
4. IPIP_Cur	.80	.76	.80	.76
5. CSES	.85	.86	.90	.90
6. SWLS	.93	.93	.92	.92
7. CWB	.82	.87	.87	.85
8. HBCL_WME	.82	.81	.76	.78
9. HBCL_AC	.72	.72	.66	.70
10. HBCL_TR	.79	.76	.75	.75

Note: IPIP_Ord: the Orderliness scale of the IPIP; IPIP_Ind: the Achievement-Striving scale of the IPIP used for measuring industriousness; IPIP_SC: the Cautiousness scale of the IPIP used for measuring self-control; IPIP_Cur: the Adventurousness scale of the IPIP used for measuring curiosity;

Table 4. Results on the 23 training feedback questions for the trained group using a dichotomous response scale.

Feedback Questions	Frequency	Proportion
Have you ever been confused about personality survey questions similar to the ones you see in the training?		
1. No, I’ve seen similar questions, but my understanding is the same as explained in the training.	13	2.6%
2. No, I’ve barely seen questions similar to the ones in the training.	112	22.7%
3. Yes, but very rarely.	89	18.0%
4. Yes, sometimes.	130	26.3%
5. Yes, frequently.	148	30.0%
6. Yes, all the time.	2	0.4%
Did you learn anything new from the training?		
1. No, I had already known pretty much everything before I had this training.	44	8.9%
2. No, but my existing knowledge about personality survey responding is more systematic after the training.	226	45.7%
3. Yes, the training answered some of the questions I have about personality surveys.	78	15.8%
4. Yes, the training answered my major confusion about personality surveys.	146	29.6%

Note: *N* = 494.

Table 5. Estimated GGUM item parameters and intermediate items of the IPIP Orderliness scale

		Untrained & Dichotomous			Trained & Dichotomous			Untrained & Polytomous					Trained & Polytomous				
ID	Content	α	δ	τ	α	δ	τ	α	δ	τ_1	τ_2	τ_3	α	δ	τ_1	τ_2	τ_3
1	Like order.	2.81	0.79^{I}	-2.02	2.71	0.78	-2.02	1.81	2.03	-4.93	-3.93	-1.30	1.27	2.26	-4.83	-4.16	-1.41
2	Like to tidy up.	1.99	1.30	-2.16	1.74	0.81	-1.64	1.63	2.13	-4.54	-3.16	-1.15	1.74	2.16	-3.84	-3.05	-1.29
3	Want everything to be "just right."	1.32	0.77^{I}	-1.71	1.26	0.74^{I}	-1.62	0.82	2.35	-5.60	-4.13	-0.81	0.75	2.14	-5.20	-3.56	-0.58
4	Love order and regularity.	2.43	0.80^{I}	-1.84	2.87	0.61^{I}	-1.54	1.30	1.91	-5.21	-3.37	-0.72	1.05	2.24	-5.11	-3.60	-1.04
5	Do things according to a plan.	2.05	0.90	-2.19	1.22	0.99	-2.16	0.66	2.71	-5.26	-5.98	-0.01	0.73	2.04	-5.69	-4.23	-0.16
6	Often forget to put things back in their proper place.	2.87	-1.58	-1.18	3.97	-1.27^{I}	-0.92	2.58	-1.65	-2.32	-1.25	-0.21	1.77	-2.05	-2.86	-1.49	-0.43
7	Leave a mess in my room.	3.99	-1.37^{I}	-0.98	4.09	-1.34^{I}	-0.91	2.51	-1.71	-2.42	-1.27	-0.22	3.20	-2.41	-3.11	-1.93	-0.91
8	Leave my belongings around.	3.92	-1.70	-1.39	4.18	-1.28^{I}	-0.96	2.71	-1.63	-2.38	-1.33	-0.17	3.03	-2.01	-2.68	-1.65	-0.51
9	Am not bothered by messy people.	1.38	-1.99	-1.22	1.16	-2.37	-1.26	1.16	-2.48	-3.65	-1.78	-0.48	1.04	-2.76	-3.91	-2.06	-0.59
10	Am not bothered by disorder.	1.94	-1.97	-1.40	1.11	-2.57	-1.45	1.57	-2.20	-3.08	-1.67	-0.34	1.12	-2.84	-3.85	-2.02	-0.71

Note: Items with location parameters between the 10th and 90th percentile of the estimated theta distribution and discrimination parameters larger than 0.5 were considered intermediate items, and were identified by ^{I}Intermediate.

Table 6. Estimated GGUM item parameters and intermediate items of the IPIP Achievement-Striving scale (Industriousness)

		Untrained & Dichotomous			**Trained & Dichotomous**			**Untrained & Polytomous**					**Trained & Polytomous**				
ID	**Content**	**α**	**δ**	**τ**	**α**	**δ**	**τ**	**α**	**δ**	**τ_1**	**τ_2**	**τ_3**	**α**	**δ**	**τ_1**	**τ_2**	**τ_3**
1	Go straight for the goal.	1.65	-0.57	-1.81	1.71	-0.86	-1.45	1.30	-1.95	-5.06	-3.22	-0.77	0.98	-2.13	-5.33	-3.38	-0.54
2	Work hard.	2.71	-1.23	-2.97	2.95	-0.65^{I}	-2.12	3.22	-1.57	-3.55	-3.29	-1.41	2.69	-1.78	-4.46	-3.29	-1.68
3	Turn plans into actions.	1.79	-0.88	-2.29	1.78	-1.05	-2.16	1.78	-1.76	-4.11	-3.18	-0.88	1.34	-1.89	-4.58	-3.47	-0.73
4	Plunge into tasks with all my heart.	1.70	-0.38	-1.75	2.37	-0.66	-1.33	1.57	-1.90	-4.93	-3.22	-1.01	1.43	-1.97	-4.24	-3.11	-0.91
5	Do more than what's expected of me.	2.24	-0.85	-2.17	2.63	-0.61^{I}	-1.48	2.13	-1.59	-3.71	-2.92	-0.98	2.23	-1.89	-4.01	-2.97	-1.25
6	Set high standards for myself and others.	2.01	-0.35^{I}	-1.84	1.99	-0.92	-1.89	1.43	-1.81	-4.52	-3.42	-1.26	1.41	-1.98	-4.03	-3.51	-1.26
7	Demand quality.	1.88	-0.33^{I}	-1.86	1.64	-1.23	-2.26	1.32	-1.79	-4.30	-3.63	-0.71	1.42	-1.97	-4.63	-3.42	-1.01
8	Am not highly motivated to succeed.	1.87	1.72	-0.63	1.24	2.44	-1.18	1.03	3.18	-3.60	-1.91	-0.58	0.71	3.40	-4.03	-1.47	-0.54
9	Do just enough work to get by.	1.84	2.67	-1.54	1.33	2.59	-1.35	0.83	3.06	-3.94	-1.66	-0.53	1.28	2.80	-3.43	-1.65	-0.54
10	Put little time and effort into my work.	1.12	2.98	-0.76	1.00	2.95	-0.79	0.66	3.52	-3.83	-1.03	-0.74	1.47	3.07	-3.11	-1.38	-0.89

Note: Items with location parameters between the 10th and 90th percentile of the estimated theta distribution and discrimination parameters larger than 0.5 were considered intermediate items, and were identified by ^{I}Intermediate.

Table 7. Estimated GGUM item parameters and intermediate items of the IPIP Cautiousness scale (Self-Control)

		Untrained & Dichotomous			Trained & Dichotomous			Untrained & Polytomous					Trained & Polytomous				
ID	**Content**	α	δ	τ	α	δ	τ	α	δ	τ_1	τ_2	τ_3	α	δ	τ_1	τ_2	τ_3
1	Avoid mistakes.	1.19	0.72[I]	-2.26	0.77	0.95	-2.49	0.57	2.91	-6.71	-6.07	-0.27	0.63	2.53	-6.72	-4.89	-0.34
2	Choose my words with care.	1.13	1.12	-2.69	0.79	1.28	-2.69	0.55	2.17	-6.89	-5.01	-0.37	0.29	4.32	-11.34	-9.17	-0.01
3	Stick to my chosen path.	0.72	1.25	-2.43	0.53	1.68	-2.59	0.52	3.95	-9.44	-5.97	-0.01	0.78	2.55	-6.02	-3.91	-0.01
4	Jump into things without thinking.	3.17	-1.76	-0.84	3.91	-1.75	-0.93	2.69	-2.37	-3.01	-1.50	-0.48	2.91	-2.71	-3.34	-1.72	-0.86
5	Make rash decisions.	3.88	-1.76	-0.93	2.91	-1.74	-0.74	2.64	-3.08	-3.64	-2.00	-0.95	2.73	-1.98	-2.53	-0.89	-0.17
6	Like to act on a whim.	2.29	-1.44	-0.92	2.31	-1.25[I]	-0.66	1.49	-3.04	-4.28	-2.51	-0.65	1.54	-2.93	-4.15	-2.18	-0.80
7	Rush into things.	2.94	-1.88	-1.04	3.67	-2.24	-1.44	2.47	-1.98	-2.78	-1.10	-0.18	1.90	-2.17	-3.05	-1.03	-0.27
8	Do crazy things.	2.77	-1.39[I]	-0.73	1.47	-1.50	-0.45	1.45	-2.43	-3.24	-1.63	-0.36	1.28	-2.98	-3.62	-2.13	-0.77
9	Act without thinking.	3.79	-2.03	-1.11	3.12	-1.69	-0.76	2.49	-2.80	-3.37	-1.75	-0.35	2.17	-2.91	-3.62	-1.82	-0.99
10	Often make last-minute plans.	2.09	-1.22[I]	-0.85	2.27	-1.15[I]	-0.96	1.17	-2.44	-4.05	-2.16	-0.42	1.02	-2.86	-4.37	-2.16	-0.75

Note: Items with location parameters between the 10th and 90th percentile of the estimated theta distribution and discrimination parameters larger than 0.5 were considered intermediate items, and were identified by [I]Intermediate.

Table 8. Estimated GGUM item parameters and intermediate items of the IPIP Adventurousness scale (Curiosity)

		Untrained & Dichotomous			Trained & Dichotomous			Untrained & Polytomous					Trained & Polytomous				
ID	Content	α	δ	τ	α	δ	τ	α	δ	τ_1	τ_2	τ_3	α	δ	τ_1	τ_2	τ_3
1	Prefer variety to routine.	1.46	2.08	-1.97	1.37	1.57	-1.47	0.90	2.70	-5.78	-2.70	-0.54	0.69	2.95	-5.73	-2.79	-0.41
2	Like to visit new places.	2.23	1.14	-2.76	1.60	1.47	-3.30	0.51	2.24	-6.03	-6.84	-1.50	0.71	1.85	-4.72	-4.57	-1.39
3	Am interested in many things.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	Like to begin new things.	2.01	1.23	-2.56	1.68	1.11[I]	-2.23	1.34	2.36	-5.87	-4.14	-0.98	0.67	2.24	-6.99	-4.65	-0.15
5	Prefer to stick with things that I know.	1.94	-1.85	-2.46	1.55	-0.98	-1.52	1.14	-2.42	-5.03	-3.34	-0.50	1.35	-2.10	-4.63	-2.63	-0.51
6	Dislike changes.	4.24	-1.42	-1.44	5.08	-1.05	-1.09	3.04	-2.12	-3.55	-2.13	-0.96	2.09	-1.80	-3.43	-1.88	-0.34
7	Don't like the idea of change.	3.40	-1.46	-1.34	4.58	-0.97	-0.90	3.18	-2.24	-3.56	-2.12	-0.82	2.51	-1.72	-3.22	-1.67	-0.41
8	Am a creature of habit.	1.33	-1.56	-2.84	1.62	-0.85[I]	-1.82	0.83	-2.37	-5.70	-4.22	-0.69	0.86	-2.03	-5.04	-3.57	-0.39
9	Dislike new foods.	1.14	-2.23	-0.42	1.11	-2.38	-0.38	0.64	-3.49	-4.49	-1.21	-0.52	0.54	-3.98	-4.60	-1.55	-0.67
10	Am attached to conventional ways.	1.23	-1.70	-1.59	1.15	-1.45	-1.22	0.88	-3.12	-5.21	-3.54	-0.31	0.67	-3.20	-5.85	-2.89	-0.34

Note: Items with location parameters between the 10th and 90th percentile of the estimated theta distribution and discrimination parameters larger than 0.5 were considered intermediate items, and were identified by [I]Intermediate. X: the item was dropped from the analysis.

Table 9. Estimated GGUM item parameters and intermediate items of the CSES

		Untrained & Dichotomous			**Trained & Dichotomous**			**Untrained & Polytomous**					**Trained & Polytomous**				
ID	**Content**	**α**	**δ**	τ	**α**	**δ**	τ	**α**	**δ**	τ_1	τ_2	τ_3	**α**	**δ**	τ_1	τ_2	τ_3
1	I am confident I get the success I deserve in life	1.95	1.78	-2.39	1.87	1.25	-1.64	1.64	1.77	-3.88	-2.79	-0.50	1.38	-2.17	-4.14	-2.75	-0.35
2	Sometimes I feel depressed	2.01	-1.42	-1.91	2.50	-1.43	-1.83	1.82	-2.29	-3.01	-2.61	-1.02	1.15	2.23	-3.55	-3.06	-0.91
3	When I try, I generally succeed	1.79	1.20^{I}	-2.97	1.78	1.60	-3.09	1.46	1.51	-3.97	-3.73	-0.65	1.39	-1.83	-4.73	-3.49	-0.77
4	Sometimes when I fail I feel worthless	2.49	-2.12	-2.15	1.89	-2.07	-2.08	2.09	-1.87	-2.45	-1.81	-0.56	1.48	2.37	-3.39	-2.44	-1.18
5	I complete tasks successfully	1.76	1.32	-3.53	1.74	1.33	-2.86	1.44	1.61	-4.40	-4.15	-0.87	1.34	-2.02	-5.40	-3.89	-0.85
6	Sometimes, I do not feel in control of my work	1.48	-2.39	-2.10	1.59	-1.34	-1.17	1.46	-2.94	-3.88	-2.54	-0.71	1.26	2.70	-3.99	-2.59	-0.79
7	Overall, I am satisfied with myself	2.89	1.59	-2.43	2.84	1.70	-2.37	2.14	1.60	-3.59	-2.68	-0.92	2.48	-2.12	-3.75	-2.92	-1.29
8	I am filled with doubts about my competence	1.88	-2.40	-1.83	2.19	-1.61	-1.13	2.63	-1.93	-2.42	-1.36	-0.41	1.86	2.00	-2.71	-1.41	-0.61
9	I determine what will happen in my life	1.22	1.28	-2.64	1.11	1.44	-2.24	0.93	1.77	-4.68	-3.63	-0.21	1.09	-2.00	-4.70	-3.11	-0.35
10	I do not feel in control of my success in my career	1.73	-2.19	-1.39	2.07	-1.51	-0.76	1.73	-2.77	-3.58	-1.87	-0.94	1.57	2.71	-3.77	-2.03	-0.82
11	I am capable of coping with most of my problems	1.76	1.14^{I}	-3.01	2.17	0.95^{I}	-2.36	1.52	1.64	-3.80	-3.58	-0.79	1.27	-2.02	-4.27	-3.85	-0.86
12	There are times when things look pretty bleak and hopeless to me	3.25	-2.01	-2.05	2.47	-1.66	-1.65	1.93	-2.01	-2.74	-1.77	-0.52	1.90	2.38	-3.31	-2.37	-1.16

Note: Items with location parameters between the 10th and 90th percentile of the estimated theta distribution and discrimination parameters larger than 0.5 were considered intermediate items, and were identified by ^{I}Intermediate.

Table 10. Estimated 2PL and SGR item parameters of the IPIP Orderliness scale

		Untrained & Dichotomous		Trained & Dichotomous		Untrained & Polytomous				Trained & Polytomous			
ID	**Content**	**a**	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	Like order.	1.46	-1.21	1.27	-1.24	1.07	-3.27	-1.96	0.72	0.91	-3.23	-1.71	0.76
2	Like to tidy up.	1.31	-0.81	1.05	-0.75	1.06	-2.65	-1.08	0.97	1.33	-1.88	-0.81	0.83
3	Want everything to be "just right."	0.68	-0.99	0.65	-0.90	0.61	-3.97	-1.64	1.31	0.60	-3.57	-1.26	1.36
4	Love order and regularity.	1.27	-1.00	1.15	-0.89	0.87	-3.63	-1.47	1.11	0.81	-3.20	-1.22	1.07
5	Do things according to a plan.	1.15	-1.26	0.75	-1.16	0.50	-4.46	-2.54	2.11	0.53	-4.80	-1.99	1.57
6	Often forget to put things back in their proper place.	1.71	-0.41	2.10	-0.41	1.62	-1.74	-0.49	0.74	1.40	-1.66	-0.51	0.84
7	Leave a mess in my room.	1.97	-0.45	2.54	-0.47	1.56	-1.81	-0.53	0.77	2.17	-1.54	-0.48	0.72
8	Leave my belongings around.	2.59	-0.31	2.35	-0.37	1.61	-1.78	-0.40	0.83	2.05	-1.54	-0.36	0.73
9	Am not bothered by messy people.	0.90	-0.66	0.77	-0.92	0.88	-2.54	-0.70	1.24	0.91	-2.16	-0.62	1.14
10	Am not bothered by disorder.	1.26	-0.52	0.73	-0.93	1.09	-2.29	-0.60	0.98	0.96	-2.18	-0.74	1.02

Table 11. Estimated 2PL and SGR item parameters of the IPIP Achievement-Striving scale (Industriousness)

		Untrained & Dichotomous		**Trained & Dichotomous**		**Untrained & Polytomous**				**Trained & Polytomous**			
ID	**Content**	**a**	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	Go straight for the goal.	0.85	-1.21	0.98	-0.54	0.92	-3.42	-1.23	1.10	0.82	-3.27	-1.10	1.30
2	Work hard.	1.89	-1.63	1.73	-1.39	2.01	-2.36	-1.72	0.20	1.82	-2.89	-1.51	0.12
3	Turn plans into actions.	1.06	-1.38	1.08	-1.08	1.19	-2.81	-1.39	0.85	0.98	-3.15	-1.44	1.03
4	Plunge into tasks with all my heart.	0.77	-1.34	1.22	-0.56	1.07	-3.36	-1.31	0.85	1.06	-2.58	-1.06	0.98
5	Do more than what's expected of me.	1.46	-1.23	1.34	-0.77	1.35	-2.52	-1.33	0.64	1.50	-2.35	-1.08	0.63
6	Set high standards for myself and others.	0.92	-1.43	1.19	-0.92	0.98	-3.30	-1.59	0.53	1.04	-2.59	-1.39	0.65
7	Demand quality.	0.82	-1.52	0.99	-1.03	0.91	-3.30	-1.72	0.99	1.01	-3.08	-1.37	0.88
8	Am not highly motivated to succeed.	0.85	-1.21	0.74	-1.17	0.84	-2.89	-1.27	0.53	0.63	-3.48	-1.68	0.70
9	Do just enough work to get by.	1.15	-1.07	0.79	-1.16	0.72	-2.88	-1.24	0.91	0.98	-2.54	-1.13	0.67
10	Put little time and effort into my work.	0.70	-2.04	0.56	-2.11	0.60	-3.69	-2.08	0.42	1.09	-2.68	-1.65	0.09

Table 12. Estimated 2PL and SGR item parameters of the IPIP Cautiousness scale (Self-Control)

		Untrained & Dichotomous		**Trained & Dichotomous**		**Untrained & Polytomous**				**Trained & Polytomous**			
ID	**Content**	**a**	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	Avoid mistakes.	0.59	-1.71	0.38	-1.96	0.38	-6.29	-2.98	2.38	0.48	-5.26	-2.10	1.78
2	Choose my words with care.	0.61	-1.75	0.50	-1.51	0.36	-7.25	-3.01	1.65	0.43	-5.35	-2.19	1.76
3	Stick to my chosen path.	0.43	-1.40	0.32	-1.32	0.35	-6.93	-1.87	3.75	0.65	-3.63	-1.06	2.05
4	Jump into things without thinking.	2.13	-0.91	2.84	-0.80	1.76	-2.04	-0.85	0.66	1.99	-1.97	-0.95	0.62
5	Make rash decisions.	2.70	-0.84	1.90	-0.98	1.74	-2.26	-1.09	0.57	1.63	-2.21	-1.10	0.58
6	Like to act on a whim.	1.24	-0.57	1.05	-0.74	1.07	-2.41	-0.54	1.26	1.13	-2.19	-0.68	1.19
7	Rush into things.	1.99	-0.82	2.38	-0.78	1.53	-2.15	-0.89	0.83	1.32	-2.33	-1.06	0.86
8	Do crazy things.	1.42	-0.77	0.82	-1.08	1.12	-2.20	-0.76	0.85	1.04	-2.24	-0.79	0.71
9	Act without thinking.	2.58	-0.90	2.05	-0.92	1.67	-2.56	-1.04	0.58	1.54	-2.06	-1.01	0.69
10	Often make last-minute plans.	0.95	-0.49	0.99	-0.28	0.88	-2.07	-0.27	1.66	0.87	-2.14	-0.56	1.44

Table 13. Estimated 2PL and SGR item parameters of the IPIP Adventurousness scale (Curiosity)

		Untrained & Dichotomous		Trained & Dichotomous		Untrained & Polytomous				Trained & Polytomous			
ID	**Content**	**a**	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	Prefer variety to routine.	0.94	0.06	0.75	0.02	0.72	-2.93	-0.09	2.08	0.59	-2.67	0.08	2.44
2	Like to visit new places.	1.24	-1.63	0.62	-2.40	0.32	-8.62	-5.08	0.47	0.48	-4.99	-2.79	0.39
3	Am interested in many things.	X	X	X	X	X	X	X	X	X	X	X	X
4	Like to begin new things.	1.17	-1.32	0.69	-1.40	0.88	-4.10	-1.71	1.28	0.47	-6.07	-2.28	1.75
5	Prefer to stick with things that I know.	1.23	0.59	0.82	0.53	0.84	-1.74	0.79	2.83	0.98	-1.51	0.51	2.61
6	Dislike changes.	2.82	0.00	4.37	-0.05	2.05	-1.19	0.01	1.44	1.54	-1.49	0.10	1.63
7	Don't like the idea of change.	2.10	-0.13	2.45	-0.21	2.11	-1.44	-0.12	1.34	1.70	-1.41	-0.04	1.53
8	Am a creature of habit.	0.85	1.25	0.82	1.00	0.63	-1.41	1.62	4.02	0.69	-1.35	1.32	3.45
9	Dislike new foods.	0.68	-1.70	0.55	-2.18	0.50	-4.12	-1.94	1.12	0.40	-5.07	-2.35	0.96
10	Am attached to conventional ways.	0.79	-0.02	0.67	-0.16	0.68	-2.60	0.27	2.35	0.56	-2.79	-0.20	2.61

Note: X: the item was dropped from the analysis.

Table 14. Estimated 2PL and SGR item parameters of the CSES

ID	Content	Untrained & Dichotomous		Trained & Dichotomous		Untrained & Polytomous				Trained & Polytomous			
		a	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	I am confident I get the success I deserve in life	1.27	-0.59	1.12	-0.38	1.07	-2.48	-0.95	1.31	0.99	-2.17	-0.49	1.77
2	Sometimes I feel depressed	1.18	0.49	1.55	0.39	1.41	-1.24	0.20	0.98	0.98	-1.16	0.60	1.70
3	When I try, I generally succeed	1.14	-1.68	1.17	-1.40	0.87	-3.64	-2.21	0.89	0.92	-3.52	-1.58	1.02
4	Sometimes when I fail I feel worthless	1.61	0.03	1.21	0.02	1.46	-1.37	-0.13	0.77	1.19	-1.21	0.07	1.20
5	I complete tasks successfully	1.14	-2.09	1.13	-1.45	0.85	-4.16	-2.61	0.75	0.86	-4.17	-1.81	1.14
6	Sometimes, I do not feel in control of my work	0.96	-0.23	0.89	-0.14	1.11	-2.19	-0.45	1.05	0.99	-1.85	-0.16	1.38
7	Overall, I am satisfied with myself	1.94	-0.79	1.84	-0.64	1.36	-2.31	-1.05	0.74	1.73	-1.78	-0.75	0.84
8	I am filled with doubts about my competence	1.22	-0.51	1.37	-0.45	1.71	-1.67	-0.58	0.58	1.41	-1.53	-0.52	0.78
9	I determine what will happen in my life	0.76	-1.36	0.68	-0.86	0.62	-4.08	-1.79	1.48	0.84	-2.95	-0.95	1.50
10	I do not feel in control of my success in my career	1.13	-0.72	1.17	-0.75	1.24	-2.03	-0.86	0.86	1.18	-1.98	-0.62	1.07
11	I am capable of coping with most of my problems	1.14	-1.76	1.15	-1.43	0.94	-3.13	-1.88	0.87	0.92	-2.95	-1.57	1.03
12	There are times when things look pretty bleak and hopeless to me	2.16	0.04	1.57	0.00	1.40	-1.55	-0.26	0.87	1.46	-1.25	-0.04	1.03

Table 15. Estimated GGUM item parameters and intermediate items of the CPS Order scale and items from Cao et al. (2015)

		Untrained & Dichotomous			**Trained & Dichotomous**		
ID	**Content**	α	δ	τ	α	δ	τ
1	I can never find the things I want at home	1.28	-2.01	-0.32	0.80	-3.49	-0.71
2	I am an unorganized person	3.80	-1.89	-1.22	2.90	-1.66	-0.75
3	I try to keep track of my bills, but I'm not too accurate	1.00	-1.61	-0.52	1.30	-1.35	-0.35
4	Being in a clean room makes me feel uncomfortable	0.70	-4.78	-0.72	0.71	-4.58	-0.74
5	Organizing and arranging things is extremely fulfilling	1.48	1.64	-2.48	1.57	1.47	-2.19
6	It's hard for me to keep things in order	1.89	-1.77	-1.04	1.92	-1.70	-0.88
7	I can ignore a mess for a long time, but eventually I have to clean it up	1.27	-1.37	-1.44	1.86	-1.12^{I}	-1.28
8	I plan my time very carefully	1.40	1.55	-1.91	1.55	1.22^{I}	-1.26
9	I prefer not to plan ahead and instead take life as it comes	0.90	-2.23	-0.86	0.56	-2.89	-0.61
10	Every book on my bookshelf is in a specific order	0.83	2.01	-0.57	1.06	1.45	-0.24
11	I follow a strict daily schedule	1.12	2.17	-1.46	1.37	1.16^{I}	-0.53
12	I try to keep my room clean and tidy but I don't always have time to do so	1.30	-0.01^{I}	-1.13	1.70	-0.22^{I}	-1.05
13	When I have many things to do, I try to focus on the task with the highest priority first	0.89	1.37	-4.21	0.81	1.27^{I}	-3.61
14	Sometimes I wish that everyone was as organized as me	1.83	1.35^{I}	-1.36	1.56	1.36	-1.31
15	Being messy helps my creativity	1.39	-2.33	-0.98	0.99	-3.02	-0.87
16	Being clean helps me to focus	2.18	1.36	-2.44	1.71	1.43	-2.59
17	It bothers me a lot when my plans are disturbed	0.47	1.00^{I}	-2.35	0.53	1.06^{I}	-2.12
18	Organizing things is a waste of time	1.40	-2.68	-0.39	1.19	-3.86	-0.81
19	I am about average in regard to details	1.19	-0.63^{I}	-0.67	0.93	-0.64^{I}	-0.54
20	A little bit of disorganization is good for people	1.41	-1.43	-1.45	1.33	-1.59	-1.44

Table 15. (cont.)

21	Occasionally I miss a deadline or two.	1.06	-1.52	-0.95	1.14	-1.30	-0.77
22	Sometimes I do not put things in their proper place.	1.88	-2.19	-2.48	1.97	-1.60	-1.85
23	Sometimes I can tolerate the messiness of my room.	1.82	-2.25	-2.61	1.94	-1.04[I]	-1.41
24	I spend time cleaning and organizing things when I am not busy.	1.89	1.48	-2.06	1.89	1.14[I]	-1.69
25	I deviate from my routines when needed.	0.47	-0.92[I]	-4.86	0.63	-0.81[I]	-3.40
26	When my desk gets too messy, I will clean it up.	1.05	0.71[I]	-2.78	1.05	0.59[I]	-2.56
27	I am about average in regard to details.	0.97	-0.68[I]	-0.58	0.81	-0.89[I]	-0.71
28	My room neatness is about average.	1.54	-0.08[I]	-0.95	1.63	-0.06[I]	-1.01
29	I consider myself as organized as most other people.	1.71	0.47[I]	-1.34	1.94	0.33[I]	-1.04

Note: Items 1-20 came from the CPS, and Items 21-29 came from Cao et al. (2015). Items with location parameters between the 10th and 90th percentile of the estimated theta distribution and discrimination parameters larger than 0.5 were considered intermediate items, and were identified by [I]Intermediate.

Table 16. Estimated GGUM item parameters and intermediate items of the CPS Curiosity scale and items from Cao et al. (2015)

ID	Content	Untrained & Dichotomous			Trained & Dichotomous		
		α	**δ**	**τ**	**α**	**δ**	**τ**
1	I am open to new concepts but only if they are not hard to understand	1.75	-1.36	-0.89	1.01	1.78	-0.91
2	I learn new things only when I have to	1.86	-2.51	-1.23	1.19	2.55	-0.85
3	I am not really interested in new technology	0.77	-3.17	-0.83	0.73	3.07	-0.85
4	I am usually intrigued by what I learn in classes	1.29	1.13	-3.04	1.53	-0.54^{I}	-1.94
5	I only care about information that is relevant to me	1.16	-1.56	-0.40	0.80	2.39	-0.70
6	I sometimes try new things just so I can learn more about them	1.85	0.86^{I}	-2.56	1.73	-0.76^{I}	-2.40
7	I can be persuaded to try some new things, but most of the time I am reluctant to do so	1.11	-1.64	-1.10	0.86	2.08	-1.29
8	I sometimes read non-fiction books to learn something new	0.67	1.28	-3.22	1.36	-0.75^{I}	-1.80
9	I prefer to explore new concepts rather than apply them	0.49	-0.63^{I}	-0.66	0.73	0.30^{I}	-0.26
10	I am interested in what is happening around the world	1.13	1.30	-3.78	1.47	-0.86^{I}	-2.55
11	I am excited about new knowledge	2.79	0.71^{I}	-2.60	2.94	-0.56^{I}	-2.15
12	I like to learn new things whenever I have time	3.06	1.17	-2.75	2.51	-0.50^{I}	-1.86
13	I am as curious as anybody else I know	1.51	-0.14^{I}	-1.84	1.49	0.39^{I}	-1.22
14	I am not curious about the things that I don't know	0.83	-3.19	-0.88	1.11	3.22	-0.89
15	I would prefer a job where I don't have to learn anything new	1.70	-2.64	-1.16	1.37	2.50	-0.84
16	I prefer to read fiction books rather than non-fiction	0.52	-0.57^{I}	-0.61	0.43	1.04^{I}	-0.68
17	I am fascinated by science	1.28	0.73^{I}	-2.21	1.31	-0.71^{I}	-1.81
18	I am not interested in learning new things	1.42	-2.85	-0.60	1.25	3.30	-0.79
19	I like to experience new things, but find myself limited by my obligations	1.19	-0.18^{I}	-1.32	1.03	0.10^{I}	-1.14
20	I try new restaurants only when other people recommend them	0.72	-1.40	-0.25	1.45	0.83^{I}	-0.19

Table 16. (cont.)

21	I like to experience new things, but seldom have time.	0.99	-0.33^{I}	-0.95	0.98	0.29^{I}	-0.72
22	I am not excited about new technology, but I become interested when others show me how to	0.79	-1.65	-0.62	0.70	1.67	-0.37
23	At times I prefer to try new things rather than stick to old choices.	0.96	0.95^{I}	-2.22	1.09	-0.75^{I}	-2.02
24	Occasionally I find myself interested in information that I really don't need.	0.72	1.49	-3.86	0.89	-0.50^{I}	-2.78
25	I do not mind trying new things when there are not many choices.	0.83	0.47^{I}	-2.05	0.63	-0.03^{I}	-1.90
26	I am about as curious as my friends.	1.51	-0.33^{I}	-1.55	1.59	0.55^{I}	-1.07
27	I am about average in curiosity about new knowledge.	1.12	-1.34	-1.45	1.82	0.97^{I}	-0.93
28	I have a moderate interest in learning new skills	1.03	-0.22^{I}	-1.93	1.33	0.56^{I}	-1.29

Note: Items 1-20 came from the CPS, and Items 21-29 came from Cao et al. (2015). Items with location parameters between the 10th and 90th percentile of the estimated theta distribution and discrimination parameters larger than 0.5 were considered intermediate items, and were identified by ^{I}Intermediate.

Table 17. Estimated GGUM item parameters and intermediate items of the CPS Industriousness scale

		Untrained & Dichotomous			**Trained & Dichotomous**		
ID	**Content**	**α**	**δ**	**τ**	**α**	**δ**	**τ**
1	I am competitive and play to win	0.61	1.96	-2.29	1.33	-0.01^{I}	-0.75
2	I find it easy to stick to my plans	0.86	1.08^{I}	-2.69	1.38	0.29^{I}	-1.39
3	I am average at the things I do	1.01	-1.49	-1.24	0.77	-1.75	-1.37
4	I frequently make up believable excuses for not finishing my work	1.22	-2.83	-0.94	1.12	-2.29	-0.49
5	I finish my work on time but try not to work more than I have to	1.04	-1.56	-1.03	2.10	-1.02^{I}	-0.91
6	I work hard, but I know when it's time to quit	1.09	0.09^{I}	-2.08	1.02	-0.04^{I}	-2.15
7	I enjoy the process of doing things and don't care much about the results	0.92	-2.70	-1.03	0.75	-2.87	-0.63
8	Being successful is more important than most other things in my life	0.51	2.82	-0.85	0.18	-5.13	-1.14
9	I don't care very much about the quality of my work	2.61	-2.51	-0.39	1.29	-3.54	-0.83
10	I hardly ever finish the tasks I start	1.64	-2.81	-1.20	1.56	-2.46	-0.62
11	I tend to do just what is expected of me when doing a job	0.82	-2.20	-1.44	1.23	-1.40	-1.03
12	I always want to be better than others in the things I do	0.55	1.98	-2.60	1.01	-0.07^{I}	-0.81
13	There is too much to be done to waste time relaxing	0.36	3.21	-0.66	0.39	3.65	-1.26
14	When I set my mind on achieving a goal, I can always reach it	1.04	1.35	-2.34	1.88	0.10^{I}	-0.89
15	I always try to do my best work even when no one will know	1.47	0.86^{I}	-2.55	1.82	0.48^{I}	-1.75
16	If I am interested in something I don't mind working hard	1.30	0.42^{I}	-3.25	1.22	0.40^{I}	-3.17
17	To me, being moderately successful is enough	1.42	-1.02	-1.53	0.52	-1.78	-2.42
18	I don't really care about being successful	1.00	-1.93	-0.37	0.62	-3.22	-0.70
19	People should not sacrifice too much for work	0.61	-0.84^{I}	-2.10	0.60	-1.92	-2.88
20	I try to do the minimal amount of work possible to maintain my current status	1.76	-1.73	-0.58	2.87	-1.48	-0.61

Note: Items with location parameters between the 10th and 90th percentile of the estimated theta distribution and discrimination parameters larger than 0.5 were considered intermediate items, and were identified by ^{I}Intermediate.

Table 18. Estimated GGUM item parameters and intermediate items of the CPS Self-control scale

		Untrained & Dichotomous			**Trained & Dichotomous**		
ID	**Content**	**α**	**δ**	**τ**	**α**	**δ**	**τ**
1	I try to consider all of the consequences of my actions, but sometimes can't help acting on impulse	1.21	-0.91^{I}	-1.43	1.19	-0.71^{I}	-1.20
2	I have often missed important meetings because I forgot them	2.78	-1.30	-0.15	4.09	-0.78^{I}	-0.09
3	I usually control my impulses	1.48	1.76	-3.05	0.90	1.52	-3.28
4	It is hard to distract me when I am focused on a task	0.84	1.76	-2.56	1.07	2.20	-2.63
5	Keeping a careful record of things is not my strength	1.36	-1.37	-0.67	1.80	-1.06^{I}	-0.67
6	An impulsive decision isn't always bad	1.32	-0.48^{I}	-2.00	0.61	-1.18	-2.97
7	I always think twice before saying something	0.74	2.30	-2.42	0.80	1.75	-1.21
8	I am usually cautious	1.16	1.46	-3.21	0.97	0.85^{I}	-2.85
9	I often make careless mistakes	2.37	-1.36	-0.54	2.59	-0.94^{I}	-0.27
10	I can keep my concentration only on short tasks	1.53	-1.38	-0.29	2.10	-0.88^{I}	-0.29
11	I don't think that being impulsive is a fault	1.10	-0.57^{I}	-0.63	1.00	-0.43^{I}	-0.41
12	I am meticulous at most things I do	0.95	1.57	-2.36	1.02	1.42	-1.92
13	I don't mind waiting for something better to come along	0.96	0.50^{I}	-2.13	0.81	0.66^{I}	-1.94
14	My mind wanders a lot when I'm working on something	1.75	-1.33	-1.13	1.63	-1.17	-1.01
15	I don't usually think before I talk	0.80	-2.82	-0.46	1.67	-1.12	-0.14
16	I am more careful in places I am not familiar with	0.87	0.55^{I}	-3.51	1.01	0.12^{I}	-3.23
17	I always have a detailed plan for my daily activities	0.87	2.68	-2.25	0.62	1.92	-0.84
18	I believe people can never be too careful	0.71	1.25	-2.24	0.78	0.63^{I}	-1.44
19	I am only careful on tasks that are important to me	0.69	-2.79	-0.81	1.72	-0.69^{I}	-0.12
20	I make plans if I have enough time	1.01	0.15^{I}	-2.40	1.20	0.29^{I}	-1.65

Note: Items with location parameters between the 10th and 90th percentile of the estimated theta distribution and discrimination parameters larger than 0.5 were considered intermediate items, and were identified by ^{I}Intermediate.

Table 19. Estimated GGUM item parameters and intermediate items of the CPS Order scale

ID	Content	Untrained & Dichotomous			Trained & Dichotomous		
		α	δ	τ	α	δ	τ
1	I can never find the things I want at home	1.92	-1.41	-0.21	0.89	-3.20	-0.65
2	I am an unorganized person	4.66	-1.66	-1.04	2.86	-2.16	-1.24
3	I try to keep track of my bills, but I'm not too accurate	1.23	-1.33	-0.46	1.38	-1.36	-0.38
4	Being in a clean room makes me feel uncomfortable	0.73	-4.70	-0.74	0.72	-4.51	-0.75
5	Organizing and arranging things is extremely fulfilling	1.43	1.64	-2.49	1.62	1.29	-2.01
6	It's hard for me to keep things in order	2.38	-1.41	-0.81	1.90	-1.88	-1.04
7	I can ignore a mess for a long time, but eventually I have to clean it up	1.16	-1.45	-1.50	1.68	-1.12^{I}	-1.27
8	I plan my time very carefully	1.43	2.05	-2.42	1.86	1.09^{I}	-1.19
9	I prefer not to plan ahead and instead take life as it comes	1.10	-1.92	-0.78	0.62	-2.68	-0.62
10	Every book on my bookshelf is in a specific order	0.74	2.53	-0.95	1.31	1.07^{I}	-0.18
11	I follow a strict daily schedule	1.08	2.39	-1.67	1.56	0.95^{I}	-0.48
12	I try to keep my room clean and tidy but I don't always have time to do so	1.08	0.09^{I}	-1.17	1.74	-0.17^{I}	-1.03
13	When I have many things to do, I try to focus on the task with the highest priority first	0.96	1.21^{I}	-3.91	0.90	1.01^{I}	-3.21
14	Sometimes I wish that everyone was as organized as me	1.78	1.81	-1.81	1.92	1.09^{I}	-1.13
15	Being messy helps my creativity	1.50	-2.11	-0.83	0.98	-2.82	-0.65
16	Being clean helps me to focus	2.18	1.26	-2.33	1.71	1.70	-2.86
17	It bothers me a lot when my plans are disturbed	0.48	1.30	-2.54	0.54	1.14^{I}	-2.15
18	Organizing things is a waste of time	1.83	-2.09	-0.21	1.20	-3.82	-0.81
19	I am about average in regard to details	0.94	-0.79^{I}	-0.69	0.63	-0.83^{I}	-0.47
20	A little bit of disorganization is good for people	1.36	-1.72	-1.73	1.46	-1.55	-1.40

Table 20. Estimated GGUM item parameters and intermediate items of the CPS Curiosity scale

		Untrained & Dichotomous			**Trained & Dichotomous**		
ID	**Content**	**α**	**δ**	**τ**	**α**	**δ**	**τ**
1	I am open to new concepts but only if they are not hard to understand	1.71	-1.61	-1.10	1.07	1.85	-1.02
2	I learn new things only when I have to	2.11	-2.38	-1.18	1.41	1.97	-0.46
3	I am not really interested in new technology	0.76	-3.26	-0.88	0.67	3.20	-0.80
4	I am usually intrigued by what I learn in classes	1.44	0.80^{I}	-2.63	1.57	-0.48^{I}	-1.89
5	I only care about information that is relevant to me	0.99	-1.94	-0.58	1.12	1.69	-0.43
6	I sometimes try new things just so I can learn more about them	1.73	1.08	-2.81	1.77	-0.76^{I}	-2.39
7	I can be persuaded to try some new things, but most of the time I am reluctant to do so	1.25	-1.67	-1.18	1.06	1.89	-1.23
8	I sometimes read non-fiction books to learn something new	0.67	1.91	-3.84	1.48	-0.54^{I}	-1.64
9	I prefer to explore new concepts rather than apply them	0.66	-0.69^{I}	-0.74	1.05	0.50^{I}	-0.42
10	I am interested in what is happening around the world	1.33	0.69^{I}	-2.99	1.50	-0.75^{I}	-2.44
11	I am excited about new knowledge	2.91	0.67^{I}	-2.54	3.02	-0.60^{I}	-2.16
12	I like to learn new things whenever I have time	2.92	1.22	-2.80	2.61	-0.48^{I}	-1.83
13	I am as curious as anybody else I know	0.78	-0.01^{I}	-2.47	0.86	0.36^{I}	-1.33
14	I am not curious about the things that I don't know	0.93	-3.12	-1.03	1.10	3.08	-0.74
15	I would prefer a job where I don't have to learn anything new	1.95	-2.53	-1.16	2.06	1.73	-0.43
16	I prefer to read fiction books rather than non-fiction	0.59	-0.58^{I}	-0.63	0.37	1.50	-0.89
17	I am fascinated by science	1.55	0.48^{I}	-1.93	1.35	-0.48^{I}	-1.67
18	I am not interested in learning new things	1.69	-2.63	-0.60	1.19	3.29	-0.70
19	I like to experience new things, but find myself limited by my obligations	0.95	-0.21^{I}	-1.40	0.82	0.12^{I}	-1.18
20	I try new restaurants only when other people recommend them	0.74	-1.47	-0.29	1.44	0.92^{I}	-0.20

Table 21. Estimated 2PL and SGR item parameters of the CPS Order scale and items from Cao et al. (2015)

ID	Content	Untrained & Dichotomous		Trained & Dichotomous		Untrained & Polytomous				Trained & Polytomous			
		a	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	I can never find the things I want at home	0.73	-1.69	0.44	-2.83	0.81	-2.71	-1.51	0.70	0.88	-2.69	-1.61	0.70
2	I am an unorganized person	2.38	-0.69	1.56	-0.93	1.83	-1.66	-0.81	0.38	1.61	-1.74	-0.87	0.34
3	I try to keep track of my bills, but I'm not too accurate	0.01	63.70	0.01	46.00	0.04	-11.64	13.14	46.48	0.02	-25.31	22.39	75.52
4	Being in a clean room makes me feel uncomfortable	0.34	-4.73	0.29	-5.26	0.76	-3.45	-2.11	-0.38	0.75	-3.33	-2.27	-0.45
5	Organizing and arranging things is extremely fulfilling	0.99	-0.84	1.16	-0.67	1.04	-2.45	-0.94	0.90	1.28	-1.78	-0.70	0.77
6	It's hard for me to keep things in order	1.16	-0.71	1.01	-0.83	1.36	-1.89	-0.74	0.75	1.13	-2.06	-0.80	0.82
7	I can ignore a mess for a long time, but eventually I have to clean it up	-0.63	0.13	-0.73	0.20	0.04	-29.63	-7.83	26.40	0.03	-33.72	-7.89	30.06
8	I plan my time very carefully	0.93	-0.40	0.97	-0.04	0.80	-2.57	-0.56	1.73	0.80	-2.21	-0.39	1.59
9	I prefer not to plan ahead and instead take life as it comes	0.64	-1.07	0.37	-1.77	0.47	-4.04	-1.26	1.67	0.45	-3.85	-1.62	1.30
10	Every book on my bookshelf is in a specific order	0.57	1.10	0.57	1.28	0.47	-1.59	0.86	3.04	0.40	-1.39	0.98	2.92
11	I follow a strict daily schedule	0.68	0.58	0.77	0.68	0.43	-2.79	0.51	3.57	0.55	-2.07	0.55	2.97
12	I try to keep my room clean and tidy but I don't always have time to do so	0.12	-2.77	0.00	-979.29	0.06	-24.16	-7.47	19.17	0.07	-17.41	-3.86	14.15
13	When I have many things to do, I try to focus on the task with the highest priority first	0.53	-2.95	0.44	-2.71	0.69	-4.76	-2.67	0.47	0.51	-5.91	-3.13	0.49
14	Sometimes I wish that everyone was as organized as me	1.19	-0.04	1.07	0.02	1.14	-1.60	-0.28	1.29	1.17	-1.51	-0.13	1.10
15	Being messy helps my creativity	0.85	-1.26	0.52	-2.21	1.07	-2.38	-1.22	0.38	0.94	-2.66	-1.21	0.46
16	Being clean helps me to focus	1.39	-1.05	1.20	-1.08	1.20	-2.47	-1.39	0.56	1.37	-2.23	-1.01	0.51
17	It bothers me a lot when my plans are disturbed	0.22	-2.07	0.25	-1.66	0.16	-10.93	-3.08	5.13	0.32	-6.11	-2.07	2.58
18	Organizing things is a waste of time	0.92	-2.13	0.75	-2.88	1.01	-2.99	-1.97	0.11	1.09	-2.59	-1.81	-0.06
19	I am about average in regard to details	-0.27	-0.23	0.00	511.35	0.03	-41.33	-2.39	58.01	0.02	-41.14	4.69	60.73
20	A little bit of disorganization is good for people	-0.80	0.06	-0.78	-0.08	0.03	-37.01	0.40	55.01	0.03	-31.23	1.43	49.68
21	Occasionally I miss a deadline or two.	0.00	49.32	0.00	67.93	0.01	-55.77	24.98	134.44	0.01	-51.30	27.73	139.75
22	Sometimes I do not put things in their proper place.	0.00	-73.76	0.00	-70.04	0.02	-53.41	-3.04	74.23	0.02	-41.69	-10.31	48.65
23	Sometimes I can tolerate the messiness of my room.	0.00	-56.06	0.00	-58.38	0.02	-49.75	-11.70	51.40	0.03	-38.17	-12.85	36.58
24	I spend time cleaning and organizing things when I am not busy.	1.22	-0.58	1.31	-0.51	1.06	-1.97	-0.63	1.25	1.15	-1.73	-0.63	1.09
25	I deviate from my routines when needed.	0.00	-279.05	0.01	-107.99	0.13	-17.67	-8.30	7.36	0.11	-20.33	-10.40	8.38
26	When my desk gets too messy, I will clean it up.	0.51	-2.40	0.48	-2.33	0.79	-3.45	-1.97	0.53	1.21	-2.50	-1.39	0.30
27	I am about average in regard to details.	0.00	120.98	-0.25	-0.21	0.03	-36.35	-5.30	55.98	0.02	-45.92	8.68	81.70
28	My room neatness is about average.	0.01	-16.05	0.05	-5.26	0.08	-20.89	-5.55	20.81	0.07	-16.35	-3.04	19.65
29	I consider myself as organized as most other people.	0.52	-1.02	0.39	-0.74	0.69	-3.08	-1.17	1.72	0.61	-2.99	-0.85	1.91

Table 22. Estimated 2PL and SGR item parameters of the CPS Curiosity scale and items from Cao et al. (2015)

ID	Content	Untrained & Dichotomous		Trained & Dichotomous		Untrained & Polytomous				Trained & Polytomous			
		a	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	I am open to new concepts but only if they are not hard to understand	0.00	64.08	0.01	68.96	0.02	-68.54	13.30	104.58	0.02	-47.58	20.36	74.64
2	I learn new things only when I have to	0.16	22.59	0.56	-1.85	0.86	-2.93	-1.23	0.81	1.16	-2.48	-1.39	0.43
3	I am not really interested in new technology	0.54	-1.91	0.34	-2.48	0.59	-3.65	-1.82	0.57	0.57	-3.85	-1.90	0.62
4	I am usually intrigued by what I learn in classes	0.82	-1.86	0.92	-1.27	0.97	-3.42	-1.83	0.78	1.08	-3.23	-1.71	0.68
5	I only care about information that is relevant to me	0.01	49.11	0.01	46.11	0.02	-58.44	33.14	100.00	0.01	-56.23	37.66	106.41
6	I sometimes try new things just so I can learn more about them	1.40	-1.52	1.11	-1.53	1.46	-2.61	-1.50	0.68	1.39	-2.54	-1.50	0.61
7	I can be persuaded to try some new things, but most of the time I am reluctant to do so	0.00	67.69	0.00	71.13	0.02	-62.40	10.84	78.71	0.01	-65.87	17.61	105.81
8	I sometimes read non-fiction books to learn something new	0.34	-2.51	0.87	-0.97	0.74	-2.62	-1.42	1.00	0.65	-2.95	-1.50	0.77
9	I prefer to explore new concepts rather than apply them	0.01	-2.80	0.05	2.95	0.09	-20.64	-0.92	17.64	0.06	-27.45	1.96	25.99
10	I am interested in what is happening around the world	0.75	-2.35	0.90	-1.64	0.84	-4.15	-1.94	0.64	0.91	-2.90	-1.71	0.44
11	I am excited about new knowledge	1.75	-1.80	1.99	-1.43	1.62	-3.01	-2.00	0.24	1.76	-3.12	-1.66	0.13
12	I like to learn new things whenever I have time	2.54	-1.44	1.47	-1.21	1.39	-2.95	-1.77	0.52	1.47	-2.48	-1.40	0.42
13	I am as curious as anybody else I know	0.36	-2.55	0.06	-5.74	0.46	-4.60	-2.31	1.89	0.23	-8.69	-2.55	4.06
14	I am not curious about the things that I don't know	0.44	-2.37	0.65	-2.26	0.81	-2.92	-1.69	0.36	1.03	-2.64	-1.70	0.16
15	I would prefer a job where I don't have to learn anything new	0.98	-1.44	0.84	-1.53	0.79	-2.94	-1.45	0.57	0.90	-2.75	-1.56	0.41
16	I prefer to read fiction books rather than non-fiction	0.00	-599.66	0.00	639.95	0.05	-22.47	-1.31	17.33	0.04	-23.55	-0.78	18.22
17	I am fascinated by science	0.70	-1.56	0.70	-1.12	0.73	-3.07	-1.47	0.58	0.73	-3.18	-1.72	0.55
18	I am not interested in learning new things	0.83	-2.21	0.70	-2.52	1.15	-2.47	-1.76	-0.10	1.47	-2.44	-1.69	-0.21
19	I like to experience new things, but find myself limited by my obligations	0.35	-1.29	0.19	-1.60	0.12	-12.61	-2.58	9.40	0.15	-9.55	-2.09	5.68
20	I try new restaurants only when other people recommend them	0.00	386.04	0.00	204.10	0.02	-48.35	16.31	69.35	0.02	-28.89	18.35	53.01
21	I like to experience new things, but seldom have time.	0.22	-0.80	0.11	-0.38	0.11	-15.62	-2.19	10.92	0.11	-13.86	-1.84	11.40
22	I am not excited about new technology, but I become interested when others show me how to use it	0.00	121.17	0.00	179.43	0.01	-73.34	24.32	144.53	0.02	-42.58	21.41	99.76
23	At times I prefer to try new things rather than stick to old choices.	0.60	-1.30	0.69	-1.21	0.52	-4.27	-1.78	2.30	0.81	-3.55	-1.35	1.41
24	Occasionally I find myself interested in information that I really don’t need.	0.31	-3.41	0.48	-2.40	0.66	-3.84	-2.09	1.03	0.58	-3.70	-2.12	0.89
25	I do not mind trying new things when there are not many choices.	0.31	-2.23	0.18	-2.88	0.38	-5.56	-2.87	2.81	0.46	-4.10	-1.86	2.37
26	I am about as curious as my friends.	0.24	-2.64	0.00	-1550.72	0.26	-8.02	-3.01	4.18	0.13	-13.15	-3.27	8.91
27	I am about average in curiosity about new knowledge.	0.00	-787.13	0.00	128.33	0.03	-39.78	-4.55	50.81	0.02	-50.65	7.97	89.78
28	I have a moderate interest in learning new skills	0.19	-3.88	0.00	-1850.92	0.16	-10.75	-4.88	6.53	0.11	-12.66	-4.41	9.22

Table 23. Estimated 2PL and SGR item parameters of the CPS Industriousness scale

ID	Content	Untrained & Dichotomous		Trained & Dichotomous		Untrained & Polytomous				Trained & Polytomous			
		a	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	I am competitive and play to win	0.75	-0.39	0.97	-0.13	0.44	-3.23	-0.80	2.15	0.39	-3.29	-0.73	2.27
2	I find it easy to stick to my plans	0.52	-1.73	0.56	-1.07	0.77	-3.40	-1.41	1.46	0.59	-4.08	-1.66	1.81
3	I am average at the things I do	-0.47	-0.13	0.00	479.95	0.04	-35.14	-0.31	44.26	0.02	-66.30	12.20	109.53
4	I frequently make up believable excuses for not finishing my work	0.01	123.85	0.00	1951.31	0.01	-19.86	68.26	144.61	0.02	-9.86	41.58	88.87
5	I finish my work on time but try not to work more than I have to	0.00	272.40	0.00	1091.17	0.03	-33.22	1.17	40.51	0.03	-31.84	0.85	37.29
6	I work hard, but I know when it's time to quit	0.13	-7.06	0.20	-4.43	0.33	-6.56	-3.68	2.16	0.40	-5.31	-2.37	1.83
7	I enjoy the process of doing things and don't care much about the results	0.34	-2.28	0.00	-10884.91	0.48	-4.29	-1.98	1.84	0.51	-4.45	-2.01	1.29
8	Being successful is more important than most other things in my life	0.61	0.89	0.73	0.85	0.14	-6.80	2.51	11.16	0.19	-4.10	2.32	7.55
9	I don't care very much about the quality of my work	1.00	-2.60	0.20	-8.28	0.95	-3.21	-2.26	-0.28	1.24	-3.20	-2.09	-0.35
10	I hardly ever finish the tasks I start	0.47	-2.59	0.32	-3.99	0.96	-2.75	-1.61	0.39	0.91	-3.13	-1.80	0.40
11	I tend to do just what is expected of me when doing a job	0.00	1084.92	0.06	2.96	0.03	-33.29	2.67	39.91	0.03	-32.02	3.39	40.75
12	I always want to be better than others in the things I do	0.69	-0.57	0.94	-0.19	0.52	-3.66	-0.97	1.82	0.53	-3.23	-0.80	1.81
13	There is too much to be done to waste time relaxing	0.17	2.33	0.30	1.65	0.06	-12.33	7.30	22.59	0.06	-13.63	7.89	24.55
14	When I set my mind on achieving a goal, I can always reach it	0.68	-1.04	0.94	-0.29	1.01	-3.19	-1.23	1.04	0.92	-3.07	-0.89	1.24
15	I always try to do my best work even when no one will know	0.55	-2.30	0.64	-1.53	0.92	-3.19	-1.90	0.44	1.12	-2.87	-1.54	0.34
16	If I am interested in something I don't mind working hard	0.28	-6.43	0.28	-5.78	1.01	-3.88	-2.62	-0.03	1.17	-2.95	-2.19	-0.04
17	To me, being moderately successful is enough	0.00	-72.90	0.00	-74.71	0.05	-28.51	-7.16	25.45	0.05	-26.33	-7.13	22.12
18	I don't really care about being successful	0.78	-1.22	0.38	-2.21	0.47	-4.50	-1.67	1.05	0.53	-3.65	-1.61	0.85
19	People should not sacrifice too much for work	0.01	-67.21	0.01	-73.16	0.06	-25.27	-6.37	16.88	0.07	-23.12	-6.83	12.34
20	I try to do the minimal amount of work possible to maintain my current status	0.01	76.41	0.01	64.02	0.02	-29.45	35.69	98.88	0.01	-38.47	68.34	163.67

Table 24. Estimated 2PL and SGR item parameters of the CPS Self-control scale

ID	Content	Untrained & Dichotomous		Trained & Dichotomous		Untrained & Polytomous				Trained & Polytomous			
		a	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	I try to consider all of the consequences of my actions, but sometimes can't help acting on impulse	0.00	-90.27	0.00	-96.26	0.04	-36.34	-7.13	36.80	0.05	-24.04	-5.22	20.63
2	I have often missed important meetings because I forgot them	0.01	215.97	0.01	91.86	0.02	-0.70	68.76	130.78	0.05	1.21	26.91	42.03
3	I usually control my impulses	1.00	-1.22	0.65	-1.65	0.61	-4.20	-1.91	1.45	0.76	-3.31	-1.51	1.15
4	It is hard to distract me when I am focused on a task	0.65	-0.81	0.78	-0.45	0.55	-3.24	-0.86	1.98	0.64	-2.69	-0.74	1.53
5	Keeping a careful record of things is not my strength	0.63	-0.76	0.66	-0.64	0.58	-2.96	-0.83	1.45	0.78	-2.16	-0.60	1.18
6	An impulsive decision isn't always bad	0.01	-62.39	0.02	-36.40	0.08	-25.80	-10.78	16.96	0.09	-20.89	-9.04	12.97
7	I always think twice before saying something	0.51	-0.30	0.61	0.26	0.69	-3.29	-0.50	1.94	0.55	-3.27	-0.39	2.23
8	I am usually cautious	0.88	-1.55	0.46	-2.45	0.94	-3.23	-1.75	0.87	0.85	-3.58	-1.72	0.98
9	I often make careless mistakes	1.04	-0.99	0.63	-1.52	0.62	-3.62	-1.51	1.30	0.70	-3.12	-1.31	1.06
10	I can keep my concentration only on short tasks	0.00	155.31	0.02	42.06	0.02	-37.69	37.14	95.35	0.01	-48.13	65.99	133.18
11	I don't think that being impulsive is a fault	0.27	-0.18	0.16	-0.93	0.17	-10.59	0.47	8.10	0.28	-5.29	-0.12	4.50
12	I am meticulous at most things I do	0.64	-0.85	0.82	-0.52	0.75	-2.78	-1.00	1.58	1.07	-2.46	-0.72	1.15
13	I don't mind waiting for something better to come along	0.23	-3.43	0.19	-3.02	0.53	-4.24	-1.68	2.27	0.33	-5.91	-2.16	3.35
14	My mind wanders a lot when I'm working on something	0.84	-0.21	0.67	-0.23	0.39	-3.52	-0.28	3.12	0.48	-2.86	-0.35	2.39
15	I don't usually think before I talk	0.01	77.95	0.01	108.65	0.01	-28.04	55.16	124.63	0.02	-22.66	53.34	113.87
16	I am more careful in places I am not familiar with	0.31	-4.58	0.10	-14.67	0.50	-5.15	-3.56	0.52	0.63	-4.56	-2.60	0.45
17	I always have a detailed plan for my daily activities	0.58	0.27	0.61	0.50	0.54	-2.45	0.00	2.58	0.57	-2.35	0.17	2.40
18	I believe people can never be too careful	0.45	-1.15	0.36	-0.98	0.57	-4.08	-1.45	1.77	0.60	-3.45	-1.12	1.53
19	I am only careful on tasks that are important to me	0.01	44.04	0.01	92.51	0.02	-40.29	34.90	80.84	0.03	-28.17	23.24	56.51
20	I make plans if I have enough time	0.20	-5.04	0.26	-2.45	0.49	-4.93	-2.53	1.87	0.49	-4.26	-2.04	2.09

Table 25. Estimated 2PL and SGR item parameters of the CPS Order scale

ID	Content	Untrained & Dichotomous		Trained & Dichotomous		Untrained & Polytomous				Trained & Polytomous			
		a	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	I can never find the things I want at home	0.78	-1.62	0.56	-2.34	0.89	-2.57	-1.44	0.67	0.91	-2.63	-1.59	0.67
2	I am an unorganized person	2.78	-0.68	1.98	-0.88	1.92	-1.68	-0.83	0.37	1.92	-1.64	-0.86	0.29
3	I try to keep track of my bills, but I'm not too accurate	0.01	62.78	0.01	60.26	0.02	-27.95	32.03	113.87	0.01	-39.75	35.44	119.25
4	Being in a clean room makes me feel uncomfortable	0.36	-4.53	0.32	-4.80	0.80	-3.33	-2.05	-0.37	0.81	-3.13	-2.15	-0.45
5	Organizing and arranging things is extremely fulfilling	0.90	-0.88	0.94	-0.74	0.90	-2.70	-1.03	0.99	1.16	-1.87	-0.75	0.79
6	It's hard for me to keep things in order	1.19	-0.71	1.25	-0.76	1.43	-1.87	-0.74	0.75	1.26	-1.95	-0.78	0.77
7	I can ignore a mess for a long time, but eventually I have to clean it up	-0.64	0.12	-0.79	0.18	0.03	-31.19	-8.22	27.79	0.03	-37.61	-8.79	33.53
8	I plan my time very carefully	0.93	-0.41	0.96	-0.05	0.74	-2.73	-0.60	1.84	0.76	-2.30	-0.42	1.66
9	I prefer not to plan ahead and instead take life as it comes	0.68	-1.03	0.45	-1.52	0.51	-3.75	-1.18	1.56	0.47	-3.72	-1.58	1.25
10	Every book on my bookshelf is in a specific order	0.56	1.12	0.52	1.38	0.42	-1.74	0.96	3.36	0.39	-1.43	1.00	2.99
11	I follow a strict daily schedule	0.71	0.56	0.68	0.73	0.40	-2.96	0.56	3.82	0.57	-2.03	0.53	2.93
12	I try to keep my room clean and tidy but I don't always have time to do so	0.10	-3.45	0.00	-2119.27	0.06	-26.26	-8.10	20.85	0.06	-21.22	-4.70	17.28
13	When I have many things to do, I try to focus on the task with the highest priority first	0.53	-2.95	0.44	-2.69	0.65	-5.00	-2.81	0.49	0.45	-6.54	-3.45	0.53
14	Sometimes I wish that everyone was as organized as me	1.21	-0.05	1.04	0.01	1.03	-1.71	-0.29	1.38	1.11	-1.54	-0.14	1.14
15	Being messy helps my creativity	0.89	-1.24	0.59	-2.02	1.14	-2.32	-1.20	0.38	1.02	-2.54	-1.17	0.43
16	Being clean helps me to focus	1.29	-1.08	1.00	-1.18	1.05	-2.69	-1.50	0.61	1.20	-2.35	-1.08	0.52
17	It bothers me a lot when my plans are disturbed	0.23	-2.00	0.22	-1.90	0.16	-11.20	-3.15	5.27	0.29	-6.67	-2.27	2.80
18	Organizing things is a waste of time	0.94	-2.10	0.77	-2.83	1.09	-2.87	-1.90	0.11	1.18	-2.47	-1.74	-0.08
19	I am about average in regard to details	-0.31	-0.21	-0.23	-0.42	0.03	-44.70	-2.61	62.77	0.03	-32.01	3.52	47.16
20	A little bit of disorganization is good for people	-0.86	0.05	-0.84	-0.08	0.03	-47.48	0.53	70.80	0.05	-19.91	0.92	31.23

Table 26. Estimated 2PL and SGR item parameters of the CPS Curiosity scale

ID	Content	Untrained & Dichotomous		Trained & Dichotomous		Untrained & Polytomous				Trained & Polytomous			
		a	**b**	**a**	**b**	**a**	**b1**	**b2**	**b3**	**a**	**b1**	**b2**	**b3**
1	I am open to new concepts but only if they are not hard to understand	0.01	58.10	0.01	62.95	0.01	-79.96	15.42	122.03	0.02	-54.71	23.35	85.54
2	I learn new things only when I have to	1.08	-1.24	0.60	-1.78	0.88	-2.89	-1.22	0.81	1.17	-2.50	-1.40	0.44
3	I am not really interested in new technology	0.54	-1.93	0.36	-2.35	0.62	-3.53	-1.76	0.55	0.56	-3.89	-1.92	0.63
4	I am usually intrigued by what I learn in classes	0.89	-1.78	0.92	-1.28	0.94	-3.52	-1.88	0.80	1.09	-3.23	-1.72	0.69
5	I only care about information that is relevant to me	0.01	72.01	0.01	57.37	0.01	-60.76	34.51	104.22	0.02	-54.19	36.10	101.85
6	I sometimes try new things just so I can learn more about them	1.16	-1.64	1.05	-1.57	1.34	-2.74	-1.56	0.71	1.29	-2.65	-1.56	0.64
7	I can be persuaded to try some new things, but most of the time I am reluctant to do so	0.00	56.56	0.00	76.02	0.02	-63.64	11.14	80.23	0.02	-57.12	15.19	91.73
8	I sometimes read non-fiction books to learn something new	0.33	-2.56	0.83	-1.00	0.71	-2.70	-1.46	1.03	0.63	-3.06	-1.54	0.80
9	I prefer to explore new concepts rather than apply them	0.00	-1043.84	0.01	12.50	0.07	-26.24	-1.16	22.49	0.05	-29.82	2.13	28.34
10	I am interested in what is happening around the world	0.76	-2.34	0.92	-1.62	0.83	-4.23	-1.97	0.65	0.89	-2.96	-1.75	0.46
11	I am excited about new knowledge	1.88	-1.77	2.06	-1.42	1.66	-3.01	-2.00	0.24	1.75	-3.15	-1.69	0.14
12	I like to learn new things whenever I have time	2.33	-1.47	1.53	-1.20	1.34	-3.03	-1.80	0.54	1.40	-2.56	-1.45	0.44
13	I am as curious as anybody else I know	0.24	-3.74	0.07	-4.99	0.41	-5.07	-2.54	2.08	0.20	-9.82	-2.86	4.60
14	I am not curious about the things that I don't know	0.54	-2.01	0.69	-2.16	0.84	-2.87	-1.67	0.35	1.04	-2.65	-1.70	0.17
15	I would prefer a job where I don't have to learn anything new	1.16	-1.33	0.89	-1.48	0.82	-2.90	-1.43	0.57	0.92	-2.74	-1.56	0.42
16	I prefer to read fiction books rather than non-fiction	0.00	-662.77	-0.09	-0.11	0.05	-22.44	-1.32	17.32	0.04	-24.43	-0.79	18.90
17	I am fascinated by science	0.70	-1.55	0.70	-1.13	0.71	-3.16	-1.51	0.60	0.71	-3.27	-1.76	0.57
18	I am not interested in learning new things	1.03	-1.96	0.74	-2.44	1.20	-2.43	-1.73	-0.10	1.52	-2.43	-1.69	-0.21
19	I like to experience new things, but find myself limited by my obligations	0.21	-2.09	0.14	-2.10	0.09	-17.45	-3.55	13.04	0.16	-8.88	-1.97	5.32
20	I try new restaurants only when other people recommend them	0.00	307.48	0.00	124.51	0.02	-55.81	18.86	80.18	0.03	-28.76	18.28	52.80

Table 27. Model fit of the dominance measures

	Untrained & Dichotomous		**Trained & Dichotomous**		**Untrained & Polytomous**		**Trained & Polytomous**	
	2PL	**GGUM**	**2PL**	**GGUM**	**SGR**	**GGUM**	**SGR**	**GGUM**
CSES								
Singlets	0	0	0	0	0	0.47	0	0.02
Doublets	0.91	1.22	2.06	2.42	5.49	9.12	11.99	14.65
Triplets	1.92	1.82	3.70	3.40	7.31	10.80	9.98	12.99
IPIP_Cur								
Singlets	0	0	0	0	0	0	0	0
Doublets	0.36	0.72	3.03	1.35	6.35	7.33	15.59	17.36
Triplets	0.58	0.65	4.17	1.90	7.08	8.25	11.36	13.88
IPIP_Ind								
Singlets	0	0	0	0	0	0.79	0	0.19
Doublets	1.99	1.90	1.95	1.89	6.15	9.97	11.01	15.15
Triplets	3.38	2.91	3.04	2.95	7.94	13.37	11.33	18.23
IPIP_Ord								
Singlets	0	0	0	0	0	3.74	0	0.48
Doublets	4.61	3.31	4.76	3.00	7.65	12.68	15.59	20.00
Triplets	8.14	5.45	8.44	4.46	8.06	12.37	14.80	19.23
IPIP_SC								
Singlets	0	0	0	0	0	0	0	0.20
Doublets	0.69	0.73	1.04	0.89	5.16	6.62	9.56	12.49
Triplets	1.25	0.59	1.87	1.12	6.40	8.54	6.85	11.36

Note: Singlets: item singlets; Doublets: item doublets; Triplets: item triplets.

Table 28. Model fit of the ideal-point measures

	Untrained & Dichotomous		**Trained & Dichotomous**		**Untrained & Polytomous**	**Trained & Polytomous**
	2PL	**GGUM**	**2PL**	**GGUM**	**SGR**	**SGR**
CPS_Cur20						
Singlets	0	0	0	0	11.87	0
Doublets	8.33	0.40	7.65	0.59	15.66	19.60
Triplets	10.43	0.55	9.79	0.84	15.57	19.15
CPS_Ind20						
Singlets	0	0	0	3.98	25.68	2.01
Doublets	11.47	4.33	12.91	8.00	28.19	24.12
Triplets	15.30	6.04	17.46	8.92	25.61	23.18
CPS_Ord20						
Singlets	0	0.14	0	0.21	0.12	0.31
Doublets	5.11	2.35	6.96	3.85	12.99	22.47
Triplets	6.73	3.19	9.31	5.35	13.51	21.39
CPS_SC20						
Singlets	0	0	0.00	2.79	1.67	408.77
Doublets	10.24	1.65	7.97	3.75	11.57	229.41
Triplets	12.80	2.26	10.36	3.96	11.97	139.47
Cur28						
Singlets	0	0.02	0	0.03	0	0
Doublets	6.79	1.42	8.06	1.43	11.27	22.12
Triplets	9.19	1.89	10.45	1.89	11.57	20.06
Ord29						
Singlets	0	0	0	0	0.16	0.87
Doublets	14.86	2.65	14.71	3.14	15.35	24.80
Triplets	18.79	3.59	19.09	4.36	14.79	22.28

Note: Singlets: item singlets; Doublets: item doublets; Triplets: item triplets.

Table 29. Criterion-related validity of the IPIP Orderliness scale, modeled by 2PL, SGR, and GGUM

	N	N.E	SWLS	CWB	HBCL_WME	HBCL_AC	HBCL_TR	AP
1. Group 1 by GGUM	490	340	0.19*	-0.22*	0.27*	0.25*	-0.19*	0.08
2. Group 2 by GGUM	495	349	0.09	-0.25*	0.19*	0.21*	-0.15*	0.10*
3. Group 3 by GGUM	494	343	0.20*	-0.17*	0.20*	0.29*	-0.16*	0.15*
4. Group 4 by GGUM	498	348	0.20*	-0.17*	0.24*	0.23*	-0.25*	0.10*
5. Group 1 by 2PL	490	340	0.21*	-0.21*	0.30*	0.27*	-0.20*	0.10*
6. Group 2 by SGR	495	349	0.08	-0.19*	0.18*	0.20*	-0.16*	0.09
7. Group 3 by 2PL	494	343	0.19*	-0.16*	0.24*	0.29*	-0.16*	0.16*
8. Group 4 by SGR	498	348	0.20*	-0.15*	0.24*	0.22*	-0.27*	0.11*
UntG			0.14	-0.23	0.23	0.23	-0.17	0.09
TG			0.20	-0.17	0.22	0.26	-0.21	0.12
UntD			0.14	-0.20	0.24	0.23	-0.18	0.09
TD			0.20	-0.16	0.24	0.26	-0.21	0.13
Untrained			0.14	-0.22	0.23	0.23	-0.18	0.09
Trained			0.20	-0.16	0.23	0.26	-0.21	0.13
Polytomous			0.14	-0.19	0.21	0.21	-0.21	0.10
Dichotomous			0.20	-0.19	0.25	0.28	-0.18	0.12

Note: N: the sample size of all correlations except for the ones involving CWB; N.E: the sample size of the correlation involving CWB; Group 1: untrained and dichotomous; Group 2: untrained and polytomous; Group 3: trained and dichotomous; Group 4: trained and polytomous; UntG: mean validity of all the untrained groups under GGUM; TG: mean validity of all the trained groups under GGUM; UntD: mean validity of all the untrained groups under dominance models; TD: mean validity of all the trained groups under dominance models; Untrained: mean validity of all the untrained groups; Trained: mean validity of all the trained groups; Polytomous: mean validity of all the polytomous groups; Dichotomous: mean validity of all the dichotomous group; SWLS: the Satisfaction with Life Scale; CWB: the 10-item CWB checklist; HBCL_WME: the Wellness Maintenance and Enhancement scale of the Health Behavior Checklist; HBCL_AC: the Accident Control scale of the Health Behavior Checklist; HBCL_TR: the Traffic Risk scale of the Health Behavior Checklist; AP: academic performance.
**p* < .05.

Table 30. Criterion-related validity of the IPIP Achievement-Striving (Industriousness) scale, modeled by 2PL, SGR, and GGUM

	N	N.E	SWLS	CWB	HBCL_WME	HBCL_AC	HBCL_TR	AP
1. Group 1 by GGUM	490	340	0.28*	-0.31*	0.27*	0.23*	-0.03	0.17*
2. Group 2 by GGUM	495	349	0.34*	-0.18*	0.33*	0.37*	0.01	0.22*
3. Group 3 by GGUM	494	343	0.30*	-0.25*	0.29*	0.31*	-0.03	0.20*
4. Group 4 by GGUM	498	348	0.15*	-0.25*	0.27*	0.22*	-0.19*	0.20*
5. Group 1 by 2PL	490	340	0.33*	-0.34*	0.31*	0.26*	0.01	0.23*
6. Group 2 by SGR	495	349	0.32*	-0.19*	0.29*	0.34*	-0.02	0.22*
7. Group 3 by 2PL	494	343	0.32*	-0.25*	0.30*	0.34*	-0.03	0.22*
8. Group 4 by SGR	498	348	0.16*	-0.24*	0.26*	0.21*	-0.18*	0.21*
UntG			0.31	-0.25	0.30	0.30	-0.01	0.20
TG			0.22	-0.25	0.28	0.27	-0.11	0.20
UntD			0.33	-0.26	0.30	0.30	-0.01	0.22
TD			0.24	-0.25	0.28	0.27	-0.10	0.21
Untrained			0.32	-0.25	0.30	0.30	-0.01	0.21
Trained			0.23	-0.25	0.28	0.27	-0.11	0.21
Polytomous			0.24	-0.21	0.29	0.28	-0.10	0.21
Dichotomous			0.31	-0.29	0.29	0.29	-0.02	0.21

Note: N: the sample size of all correlations except for the ones involving CWB; N.E: the sample size of the correlation involving CWB; Group 1: untrained and dichotomous; Group 2: untrained and polytomous; Group 3: trained and dichotomous; Group 4: trained and polytomous; UntG: mean validity of all the untrained groups under GGUM; TG: mean validity of all the trained groups under GGUM; UntD: mean validity of all the untrained groups under dominance models; TD: mean validity of all the trained groups under dominance models; Untrained: mean validity of all the untrained groups; Trained: mean validity of all the trained groups; Polytomous: mean validity of all the polytomous groups; Dichotomous: mean validity of all the dichotomous group; SWLS: the Satisfaction with Life Scale; CWB: the 10-item CWB checklist; HBCL_WME: the Wellness Maintenance and Enhancement scale of the Health Behavior Checklist; HBCL_AC: the Accident Control scale of the Health Behavior Checklist; HBCL_TR: the Traffic Risk scale of the Health Behavior Checklist; AP: academic performance.
**p* < .05.

Table 31. Criterion-related validity of the IPIP Cautiousness (Self-control) scale, modeled by 2PL, SGR, and GGUM

	N	N.E	SWLS	CWB	HBCL_WME	HBCL_AC	HBCL_TR	AP
1. Group 1 by GGUM	490	340	0.14*	-0.38*	0.19*	0.14*	-0.24*	0.19*
2. Group 2 by GGUM	495	349	0.09*	-0.29*	0.05	0.06	-0.39*	0.03
3. Group 3 by GGUM	494	343	0.18*	-0.35*	0.08	0.12*	-0.32*	0.17*
4. Group 4 by GGUM	498	348	0.13*	-0.24*	0.13*	0.17*	-0.51*	0.13*
5. Group 1 by 2PL	490	340	0.13*	-0.38*	0.22*	0.16*	-0.29*	0.17*
6. Group 2 by SGR	495	349	0.10*	-0.24*	0.06	0.08	-0.38*	0.05
7. Group 3 by 2PL	494	343	0.19*	-0.37*	0.12*	0.16*	-0.34*	0.15*
8. Group 4 by SGR	498	348	0.13*	-0.22*	0.11*	0.17*	-0.51*	0.14*
UntG			0.11	-0.33	0.12	0.10	-0.32	0.11
TG			0.16	-0.30	0.10	0.15	-0.42	0.15
UntD			0.11	-0.31	0.14	0.12	-0.34	0.11
TD			0.16	-0.30	0.11	0.16	-0.42	0.15
Untrained			0.11	-0.32	0.13	0.11	-0.33	0.11
Trained			0.16	-0.30	0.11	0.16	-0.42	0.15
Polytomous			0.11	-0.25	0.09	0.12	-0.45	0.09
Dichotomous			0.16	-0.37	0.15	0.15	-0.30	0.17

Note: N: the sample size of all correlations except for the ones involving CWB; N.E: the sample size of the correlation involving CWB; Group 1: untrained and dichotomous; Group 2: untrained and polytomous; Group 3: trained and dichotomous; Group 4: trained and polytomous; UntG: mean validity of all the untrained groups under GGUM; TG: mean validity of all the trained groups under GGUM; UntD: mean validity of all the untrained groups under dominance models; TD: mean validity of all the trained groups under dominance models; Untrained: mean validity of all the untrained groups; Trained: mean validity of all the trained groups; Polytomous: mean validity of all the polytomous groups; Dichotomous: mean validity of all the dichotomous group; SWLS: the Satisfaction with Life Scale; CWB: the 10-item CWB checklist; HBCL_WME: the Wellness Maintenance and Enhancement scale of the Health Behavior Checklist; HBCL_AC: the Accident Control scale of the Health Behavior Checklist; HBCL_TR: the Traffic Risk scale of the Health Behavior Checklist; AP: academic performance.
*p < .05.

Table 32. Criterion-related validity of the IPIP Adventurousness (Curiosity) scale, modeled by 2PL, SGR, and GGUM

	N	N.E	SWLS	CWB	HBCL_WME	HBCL_AC	HBCL_TR	AP
1. Group 1 by GGUM	490	340	0.13*	-0.04	-0.02	0.02	0.19*	0.09*
2. Group 2 by GGUM	495	349	0.14*	-0.09	0.07	0.10*	0.07	0.12*
3. Group 3 by GGUM	494	343	0.11*	-0.12*	0.12*	0.12*	0.07	0.10*
4. Group 4 by GGUM	498	348	0.04	-0.10	0.02	-0.01	0.19*	0.07
5. Group 1 by 2PL	490	340	0.12*	-0.06	-0.02	0.02	0.18*	0.12*
6. Group 2 by SGR	495	349	0.14*	-0.06	0.07	0.11*	0.08	0.16*
7. Group 3 by 2PL	494	343	0.05	-0.11*	0.07	0.08	0.09	0.11*
8. Group 4 by SGR	498	348	0.04	-0.10	0.07	0.01	0.21*	0.07
UntG			0.13	-0.07	0.02	0.06	0.13	0.11
TG			0.08	-0.11	0.07	0.05	0.13	0.08
UntD			0.13	-0.06	0.02	0.06	0.13	0.14
TD			0.04	-0.11	0.07	0.04	0.15	0.09
Untrained			0.13	-0.06	0.02	0.06	0.13	0.12
Trained			0.06	-0.11	0.07	0.05	0.14	0.09
Polytomous			0.09	-0.09	0.06	0.05	0.14	0.10
Dichotomous			0.10	-0.08	0.04	0.06	0.13	0.11

Note: N: the sample size of all correlations except for the ones involving CWB; N.E: the sample size of the correlation involving CWB; Group 1: untrained and dichotomous; Group 2: untrained and polytomous; Group 3: trained and dichotomous; Group 4: trained and polytomous; UntG: mean validity of all the untrained groups under GGUM; TG: mean validity of all the trained groups under GGUM; UntD: mean validity of all the untrained groups under dominance models; TD: mean validity of all the trained groups under dominance models; Untrained: mean validity of all the untrained groups; Trained: mean validity of all the trained groups; Polytomous: mean validity of all the polytomous groups; Dichotomous: mean validity of all the dichotomous group; SWLS: the Satisfaction with Life Scale; CWB: the 10-item CWB checklist; HBCL_WME: the Wellness Maintenance and Enhancement scale of the Health Behavior Checklist; HBCL_AC: the Accident Control scale of the Health Behavior Checklist; HBCL_TR: the Traffic Risk scale of the Health Behavior Checklist; AP: academic performance.
*p < .05.

Table 33. Criterion-related validity of the CSES, modeled by 2PL, SGR, and GGUM

	N	N.E	SWLS	CWB	HBCL_WME	HBCL_AC	HBCL_TR	AP
1. Group 1 by GGUM	490	340	0.65*	-0.32*	0.27*	0.21*	0.03	0.24*
2. Group 2 by GGUM	495	349	0.59*	-0.22*	0.16*	0.18*	-0.07	0.14*
3. Group 3 by GGUM	494	343	0.62*	-0.29*	0.24*	0.36*	-0.02	0.20*
4. Group 4 by GGUM	498	348	0.65*	-0.31*	0.23*	0.19*	-0.15*	0.10*
5. Group 1 by 2PL	490	340	0.64*	-0.33*	0.27*	0.21*	0.03	0.25*
6. Group 2 by SGR	495	349	0.61*	-0.26*	0.20*	0.22*	-0.08	0.17*
7. Group 3 by 2PL	494	343	0.63*	-0.28*	0.25*	0.35*	-0.01	0.20*
8. Group 4 by SGR	498	348	0.63*	-0.30*	0.23*	0.19*	-0.14*	0.11*
UntG			0.62	-0.27	0.22	0.20	-0.02	0.19
TG			0.64	-0.30	0.24	0.28	-0.08	0.15
UntD			0.63	-0.29	0.23	0.21	-0.02	0.21
TD			0.63	-0.29	0.24	0.27	-0.08	0.16
Untrained			0.62	-0.28	0.22	0.20	-0.02	0.20
Trained			0.63	-0.30	0.24	0.27	-0.08	0.15
Polytomous			0.62	-0.27	0.21	0.20	-0.11	0.13
Dichotomous			0.64	-0.31	0.26	0.28	0.01	0.22

Note: N: the sample size of all correlations except for the ones involving CWB; N.E: the sample size of the correlation involving CWB; Group 1: untrained and dichotomous; Group 2: untrained and polytomous; Group 3: trained and dichotomous; Group 4: trained and polytomous; UntG: mean validity of all the untrained groups under GGUM; TG: mean validity of all the trained groups under GGUM; UntD: mean validity of all the untrained groups under dominance models; TD: mean validity of all the trained groups under dominance models; Untrained: mean validity of all the untrained groups; Trained: mean validity of all the trained groups; Polytomous: mean validity of all the polytomous groups; Dichotomous: mean validity of all the dichotomous group; SWLS: the Satisfaction with Life Scale; CWB: the 10-item CWB checklist; HBCL_WME: the Wellness Maintenance and Enhancement scale of the Health Behavior Checklist; HBCL_AC: the Accident Control scale of the Health Behavior Checklist; HBCL_TR: the Traffic Risk scale of the Health Behavior Checklist; AP: academic performance.
*p < .05.

Table 34. Criterion-related validity of the Order scale of CPS and items from Cao et al. (2015), modeled by 2PL, SGR, and GGUM

	N	N.E	SWLS	CWB	HBCL_WME	HBCL_AC	HBCL_TR	AP
1. Group 1 by GGUM (I20)	490	340	0.25*	-0.30*	0.33*	0.28*	-0.20*	0.16*
2. Group 1 by GGUM (I29)	490	340	0.28*	-0.30*	0.34*	0.29*	-0.16*	0.15*
3. Group 3 by GGUM (I20)	494	343	0.27*	-0.19*	0.27*	0.34*	-0.17*	0.20*
4. Group 3 by GGUM (I29)	494	343	0.29*	-0.17*	0.26*	0.35*	-0.18*	0.21*
5. Group 1 by 2PL (I20)	490	340	0.20*	-0.21*	0.32*	0.25*	-0.19*	0.08
6. Group 1 by SGR (I29)	490	340	0.13*	-0.15*	0.25*	0.21*	-0.15*	0.03
7. Group 2 by SGR (I20)	495	349	0.11*	-0.18*	0.21*	0.26*	-0.14*	0.11*
8. Group 2 by SGR (I29)	495	349	0.11*	-0.12*	0.23*	0.27*	-0.09	0.10*
9. Group 3 by 2PL (I20)	494	343	0.20*	-0.10	0.22*	0.25*	-0.03	0.12*
10. Group 3 by 2PL (I29)	494	343	0.11*	-0.04	0.16*	0.17*	0.03	0.05
11. Group 4 by SGR (I20)	498	348	0.24*	-0.12*	0.29*	0.28*	-0.22*	0.09
12. Group 4 by SGR (I29)	498	348	0.22*	-0.07	0.25*	0.25*	-0.13*	0.05
UntG (I20)			0.25	-0.30	0.33	0.28	-0.20	0.16
TG (I20)			0.27	-0.19	0.27	0.34	-0.17	0.20
UntG (I29)			0.28	-0.30	0.34	0.29	-0.16	0.15
TG (I29)			0.29	-0.17	0.26	0.35	-0.18	0.21
UntD (I20)			0.15	-0.20	0.27	0.25	-0.16	0.09
TD(I20)			0.22	-0.11	0.25	0.27	-0.13	0.10
UntD (I29)			0.12	-0.13	0.24	0.24	-0.12	0.06
TD (I29)			0.16	-0.06	0.21	0.21	-0.05	0.05

Table 34. (cont.)

Untrained (I20)	0.19	-0.23	0.29	0.26	-0.18	0.12
Untrained (I29)	0.18	-0.19	0.27	0.26	-0.13	0.09
Trained (I20)	0.23	-0.13	0.26	0.29	-0.14	0.13
Trained (I29)	0.21	-0.10	0.22	0.26	-0.09	0.10

Note: I20: the 20-item CPS Order scale; I29: the 20-item CPS Order scale combined with 9 intermediate items from Cao et al. (2015); N: the sample size of all correlations except for the ones involving CWB; N.E: the sample size of the correlation involving CWB; Group 1: untrained and dichotomous; Group 2: untrained and polytomous; Group 3: trained and dichotomous; Group 4: trained and polytomous; UntG: mean validity of all the untrained groups under GGUM; TG: mean validity of all the trained groups under GGUM; UntD: mean validity of all the untrained groups under dominance models; TD: mean validity of all the trained groups under dominance models; Untrained: mean validity of all the untrained groups; Trained: mean validity of all the trained groups; Polytomous: mean validity of all the polytomous groups; Dichotomous: mean validity of all the dichotomous group; SWLS: the Satisfaction with Life Scale; CWB: the 10-item CWB checklist; HBCL_WME: the Wellness Maintenance and Enhancement scale of the Health Behavior Checklist; HBCL_AC: the Accident Control scale of the Health Behavior Checklist; HBCL_TR: the Traffic Risk scale of the Health Behavior Checklist; AP: academic performance.
$^*p < .05$.

Table 35. Criterion-related validity of the Curiosity scale of CPS and items from Cao et al. (2015), modeled by 2PL, SGR, and GGUM

	N	N.E	SWLS	CWB	HBCL_WME	HBCL_AC	HBCL_TR	AP
1. Group 1 by GGUM (I20)	490	340	0.14*	-0.23*	0.09	0.14*	0.02	0.23*
2. Group 1 by GGUM (I28)	490	340	0.09*	-0.18*	0.10*	0.12*	0.04	0.21*
3. Group 3 by GGUM (I20)	494	343	0.06	-0.17*	0.01	0.05	-0.05	0.19*
4. Group 3 by GGUM (I28)	494	343	-0.01	-0.17*	0.01	0.06	-0.02	0.21*
5. Group 1 by 2PL (I20)	490	340	0.08	-0.06	0.09*	0.10*	0.01	0.18*
6. Group 1 by SGR (I28)	490	340	0.07	0.03	0.05	0.07	-0.02	0.11*
7. Group 2 by SGR (I20)	495	349	0.06	-0.06	0.16*	0.25*	-0.07	0.17*
8. Group 2 by SGR (I28)	495	349	0.08	-0.01	0.18*	0.26*	-0.03	0.14*
9. Group 3 by 2PL (I20)	494	343	0.11*	-0.05	0.14*	0.16*	0.02	0.13*
10. Group 3 by 2PL (I28)	494	343	0.12*	-0.02	0.16*	0.13*	0.03	0.10*
11. Group 4 by SGR (I20)	498	348	0.00	-0.11	0.17*	0.18*	-0.04	0.18*
12. Group 4 by SGR (I28)	498	348	0.02	-0.10	0.18*	0.17*	0.02	0.16*
UntG (I20)			0.14	-0.23	0.09	0.14	0.02	0.23
TG (I20)			0.06	-0.17	0.01	0.05	-0.05	0.19
UntG (I29)			0.09	-0.18	0.10	0.12	0.04	0.21
TG (I29)			-0.01	-0.17	0.01	0.06	-0.02	0.21
UntD (I20)			0.07	-0.06	0.13	0.18	-0.03	0.18
TD(I20)			0.05	-0.08	0.16	0.17	-0.01	0.16
UntD (I29)			0.07	0.01	0.12	0.17	-0.02	0.13
TD (I29)			0.07	-0.06	0.17	0.15	0.03	0.13

Table 35. (cont.)

Untrained (I20)	0.09	-0.12	0.11	0.16	-0.01	0.19
Untrained (I29)	0.08	-0.05	0.11	0.15	0.00	0.16
Trained (I20)	0.06	-0.11	0.11	0.13	-0.02	0.17
Trained (I29)	0.04	-0.10	0.12	0.12	0.01	0.16

Note: I20: the 20-item CPS Curiosity scale; I28: the 20-item CPS Curiosity scale combined with 8 intermediate items from Cao et al. (2015); N: the sample size of all correlations except for the ones involving CWB; N.E: the sample size of the correlation involving CWB; Group 1: untrained and dichotomous; Group 2: untrained and polytomous; Group 3: trained and dichotomous; Group 4: trained and polytomous; UntG: mean validity of all the untrained groups under GGUM; TG: mean validity of all the trained groups under GGUM; UntD: mean validity of all the untrained groups under dominance models; TD: mean validity of all the trained groups under dominance models; Untrained: mean validity of all the untrained groups; Trained: mean validity of all the trained groups; Polytomous: mean validity of all the polytomous groups; Dichotomous: mean validity of all the dichotomous group; SWLS: the Satisfaction with Life Scale; CWB: the 10-item CWB checklist; HBCL_WME: the Wellness Maintenance and Enhancement scale of the Health Behavior Checklist; HBCL_AC: the Accident Control scale of the Health Behavior Checklist; HBCL_TR: the Traffic Risk scale of the Health Behavior Checklist; AP: academic performance.
$^*p < .05$.

Table 36. Criterion-related validity of the Industriousness scale of CPS, modeled by 2PL, SGR, and GGUM

	N	N.E	SWLS	CWB	HBCL_WME	HBCL_AC	HBCL_TR	AP
1. Group 1 by GGUM	490	340	0.29*	-0.39*	0.22*	0.18*	-0.04	0.24*
2. Group 3 by GGUM	494	343	0.14*	-0.27*	0.13*	0.09	-0.09	0.20*
3. Group 1 by 2PL	490	340	0.07	0.01	0.08	0.05	0.00	0.01
4. Group 2 by SGR	495	349	0.21*	-0.02	0.26*	0.26*	0.05	0.02
5. Group 3 by 2PL	494	343	0.17*	0.05	0.16*	0.21*	0.09	0.06
6. Group 4 by SGR	498	348	0.15*	-0.02	0.14*	0.12*	-0.05	0.11*
UntG			0.29	-0.39	0.22	0.18	-0.04	0.24
TG			0.14	-0.27	0.13	0.09	-0.09	0.20
UntD			0.14	-0.01	0.17	0.15	0.03	0.01
TD			0.16	0.02	0.15	0.17	0.02	0.08
Untrained			0.19	-0.14	0.19	0.16	0.00	0.09
Trained			0.16	-0.08	0.14	0.14	-0.01	0.12

Note: N: the sample size of all correlations except for the ones involving CWB; N.E: the sample size of the correlation involving CWB; Group 1: untrained and dichotomous; Group 2: untrained and polytomous; Group 3: trained and dichotomous; Group 4: trained and polytomous; UntG: mean validity of all the untrained groups under GGUM; TG: mean validity of all the trained groups under GGUM; UntD: mean validity of all the untrained groups under dominance models; TD: mean validity of all the trained groups under dominance models; Untrained: mean validity of all the untrained groups; Trained: mean validity of all the trained groups; SWLS: the Satisfaction with Life Scale; CWB: the 10-item CWB checklist; HBCL_WME: the Wellness Maintenance and Enhancement scale of the Health Behavior Checklist; HBCL_AC: the Accident Control scale of the Health Behavior Checklist; HBCL_TR: the Traffic Risk scale of the Health Behavior Checklist; AP: academic performance.

*p < .05.

Table 37. Criterion-related validity of the Self-control scale of CPS, modeled by 2PL, SGR, and GGUM

	N	N.E	SWLS	CWB	HBCL_WME	HBCL_AC	HBCL_TR	AP
1. Group 1 by GGUM	490	340	0.24*	-0.39*	0.26*	0.22*	-0.23*	0.13*
2. Group 3 by GGUM	494	343	0.28*	-0.32*	0.18*	0.32*	-0.22*	0.19*
3. Group 1 by 2PL	490	340	0.16*	-0.23*	0.22*	0.22*	-0.18*	0.07
4. Group 2 by SGR	495	349	0.17*	0.00	0.24*	0.28*	-0.08	0.03
5. Group 3 by 2PL	494	343	0.24*	-0.17*	0.24*	0.25*	-0.03	0.07
6. Group 4 by SGR	498	348	0.21*	-0.16*	0.30*	0.22*	-0.21*	0.09*
UntG			0.24	-0.39	0.26	0.22	-0.23	0.13
TG			0.28	-0.32	0.18	0.32	-0.22	0.19
UntD			0.16	-0.12	0.23	0.25	-0.13	0.05
TD			0.22	-0.17	0.27	0.23	-0.12	0.08
Untrained			0.19	-0.21	0.24	0.24	-0.16	0.08
Trained			0.24	-0.22	0.24	0.26	-0.15	0.12

Note: N: the sample size of all correlations except for the ones involving CWB; N.E: the sample size of the correlation involving CWB; Group 1: untrained and dichotomous; Group 2: untrained and polytomous; Group 3: trained and dichotomous; Group 4: trained and polytomous; UntG: mean validity of all the untrained groups under GGUM; TG: mean validity of all the trained groups under GGUM; UntD: mean validity of all the untrained groups under dominance models; TD: mean validity of all the trained groups under dominance models; Untrained: mean validity of all the untrained groups; Trained: mean validity of all the trained groups; SWLS: the Satisfaction with Life Scale; CWB: the 10-item CWB checklist; HBCL_WME: the Wellness Maintenance and Enhancement scale of the Health Behavior Checklist; HBCL_AC: the Accident Control scale of the Health Behavior Checklist; HBCL_TR: the Traffic Risk scale of the Health Behavior Checklist; AP: academic performance.
*p < .05.

Table 38. Pairs of criterion-related validity that were significantly different.

Measure	Condition	Testing	Criterion	r1	r2	p1	p2	*p*-value	Conclusion
IPIP_Ind	GGUM; Polytomous	T vs NT	SWLS	0.15	0.34	0	0	0	NT > T
IPIP_Ind	GGUM; Polytomous	T vs NT	TR	-0.19	0.01	0	0.92	0	T > NT
IPIP_SC	GGUM; Trained	Poly vs Dich	TR	-0.51	-0.32	0	0	0	Poly > Dich
IPIP_SC	Dominance; Trained	Poly vs Dich	TR	-0.51	-0.34	0	0	0	Poly > Dich
Cur28	Dominance; Untrained	Poly vs Dich	AC	0.26	0.07	0	0.13	0	Poly > Dich
CPS_Ind20	Dominance; Untrained	Poly vs Dich	WME	0.26	0.08	0	0.08	0	Poly > Dich
CPS_Ind20	Dominance; Untrained	Poly vs Dich	AC	0.26	0.05	0	0.32	0	Poly > Dich
CPS_Ind20	Dichotomous; Untrained	GGUM vs 2PL	CWB	-0.39	0.01	0	0.87	0	GGUM > 2PL
CPS_Ind20	Dichotomous; Trained	GGUM vs 2PL	CWB	-0.27	0.05	0	0.35	0	GGUM > 2PL
Ord29	Dominance; Dominance	T vs NT	TR	0.03	-0.15	0	0.50	0	NT > T
CPS_Ord20	Dominance; Trained	Poly vs Dich	TR	-0.22	-0.03	0	0.46	0	Poly > Dich
CPS_SC20	Dominance; Trained	Poly vs Dich	CWB	0.00	-0.23	0	0.97	0	Dich > Poly
CPS_SC20	Dominance; Trained	Poly vs Dich	TR	-0.21	-0.03	0	0.55	0	Poly > Dich
CPS_SC20	Dichotomous; Trained	GGUM vs 2PL	TR	-0.22	-0.03	0	0.55	0	GGUM > 2PL

Note: Condition: the conditions of the two parts being compared that are the same; Testing: the condition being tested; r1: the criterion-related validity of the condition before “vs” in the “Testing” column; r2: : the criterion-related validity of the condition after “vs” in the “Testing” column; p1: the *p*-value obtained from testing if r1 is significantly different from 0; p2: the *p*-value obtained from testing if r2 is significantly different from 0; Result: which condition has the larger correlation with the criterion. T: with training; NT: no training; Poly: response scale was polytomous; Dich: response scale was dichotomous.

Table 39. Means of discrimination parameters of different sets of intermediate items.

	CPS_Cur20	**CPS_Ind20**	**CPS_Ord20**	**CPS_SC20**	**Cur28**	**Ord29**
Shared_ length	7 items	4 items	3 items	5 items	12 items	8 items
Mean_untrained	1.37	1.18	0.99	1.03	1.25	1.09
Mean_trained	1.45	1.36	1.09	1.04	1.27	1.15
Diff	**0.08***	**0.18****	**0.10***	**0.01***	**0.03***	**0.07***
Trained_length	11 items	8 items	9 items	12 items	19 items	14 items
Mean_untrained	1.42	0.99	1.07	1.31	1.25	1.22
Mean_trained	1.59	1.47	1.35	1.60	1.39	1.33
Diff	**0.16****	**0.47****	**0.28****	**0.29****	**0.14****	**0.11****
Untrained_length	8 items	5 items	3 item	6 items	12 items	9 item
Mean_untrained	1.28	1.07	0.99	1.08	1.25	1.17
Mean_trained	1.32	1.21	1.09	0.97	1.27	1.20
Diff	**0.04***	**0.14****	**0.10***	**-0.11**	**0.03***	**0.03***
Trained_Unique_length	4 items	4 items	6 items	7 items	7 items	6 items
Mean_untrained	1.51	0.81	1.11	1.51	1.24	1.40
Mean_trained	1.82	1.58	1.48	2.01	1.58	1.57
Diff	**0.31****	**0.77****	**0.37****	**0.49****	**0.33****	**0.17****
Untrained_Unique_length	1 item	1 item	0	1 item	0	1 item
Mean_untrained	0.59	0.61	X	1.32	X	1.83
Mean_trained	0.37	0.60	X	0.61	X	1.56
Diff	**-0.22**	**-0.01**	**X**	**-0.70**	**X**	**-0.27**

Note: Shared_length: the number of items that are intermediate in both trained and untrained groups; Mean_untrained: mean alpha in the untrained group; Mean_trained: mean alpha in the trained group; Diff: Mean_untrained – Mean_trained; Trained_length: the number of all items that are intermediate in the trained group; Untrained_length: the number of all items that are intermediate in the untrained group; Trained_Unique_length: the number of items that are intermediate in the trained group only; Untrained_Unique_length: the number of items that are intermediate in the untrained group only; X: no data available. *0 < difference < 0.1; **difference ⩾0.1.

Table 40. The mean response time in seconds for personality measures across conditions.

	Length	Mean	Standard Deviation
G1: dichotomous + no training			
IPIP Measures	40	141.84	117.44
CSES	12	48.89	76.00
CPS Measures	97	472.75	369.80
G2: polytomous + no training			
IPIP Measures	40	145.32	128.36
CSES	12	51.55	56.65
CPS Measures	97	478.30	350.74
G3: dichotomous + training			
IPIP Measures	40	143.39	258.22
CSES	12	45.44	34.69
CPS Measures	97	440.68	295.62
G4: polytomous + training			
IPIP Measures	40	148.24	129.13
CSES	12	54.55	73.80
CPS Measures	97	469.92	273.85

Note: Length: the number of items; IPIP Measures: the four 10-item IPIP scales measuring industriousness, order, self-control, and curiosity; CSES: the core self-evaluation scale; CPS Measures: the four 20-item CPS measures measuring industriousness, order, self-control, and curiosity along with the intermediate items developed in Cao et al. (2015).

Table 41. The criterion-related validity averaged over all personality measures and outcome measures for different conditions.

Condition	Mean R
Untrained	0.08
Trained	0.07
Dichotomous	0.08
Polytomous	0.07
GGUM	0.07
SGR & 2PL	0.08

Note: The sample size *N* based on which the correlations were averaged over ranged from 340 (for the CWB-related correlations) to 495.

FIGURES

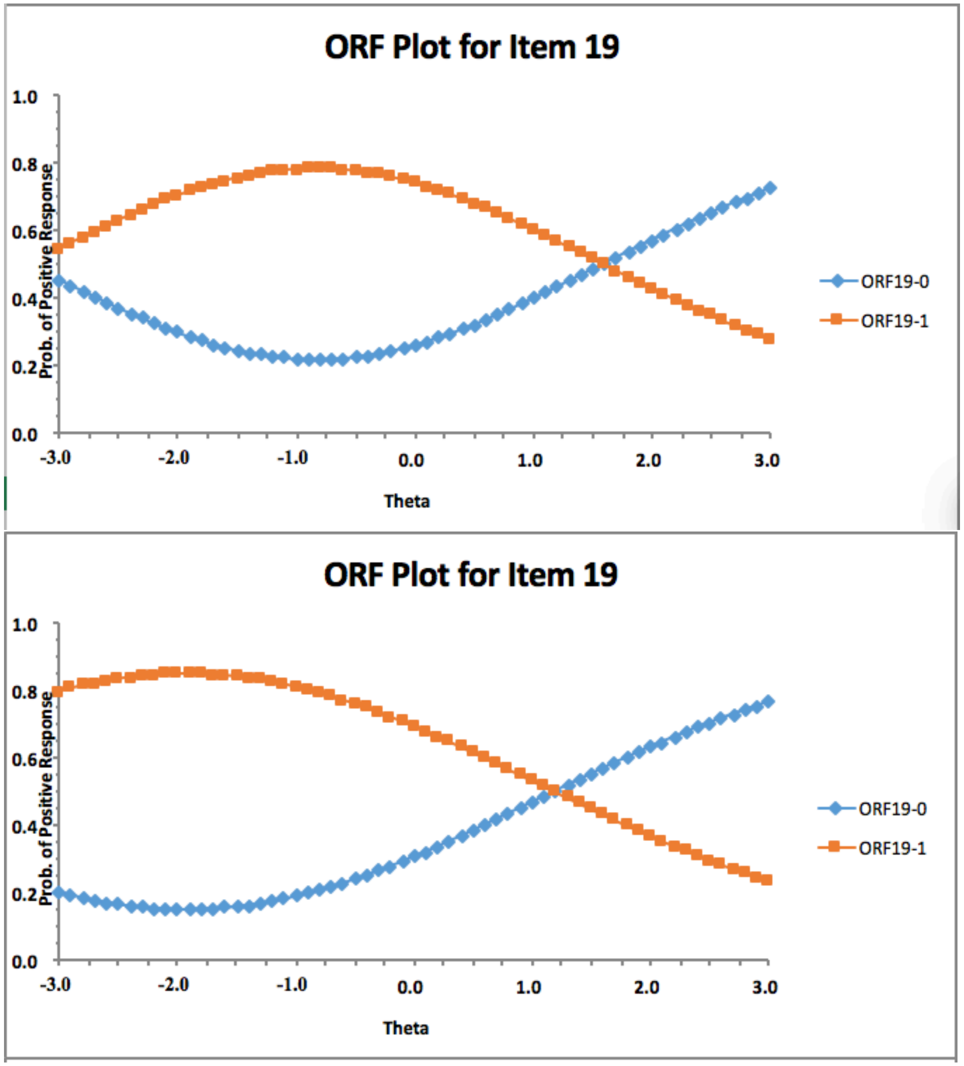


Figure 1. IRT ICCs for Item 19 (“***People should not sacrifice too much for work***”) of the Industriousness scale of CPS for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the untrained group only.

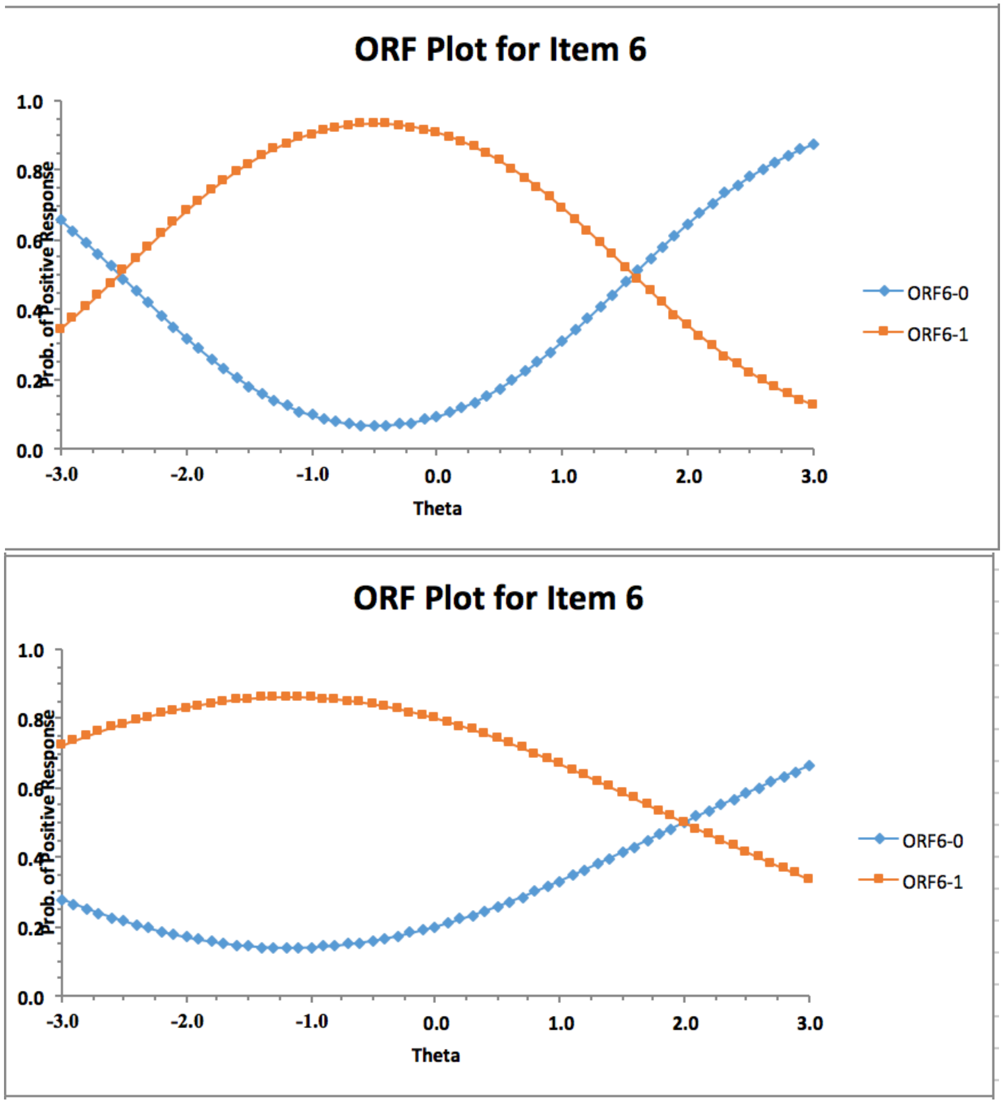


Figure 2. IRT ICCs for Item 6 ("***An impulsive decision isn't always bad***") of the Self-control scale of CPS for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the untrained group only.

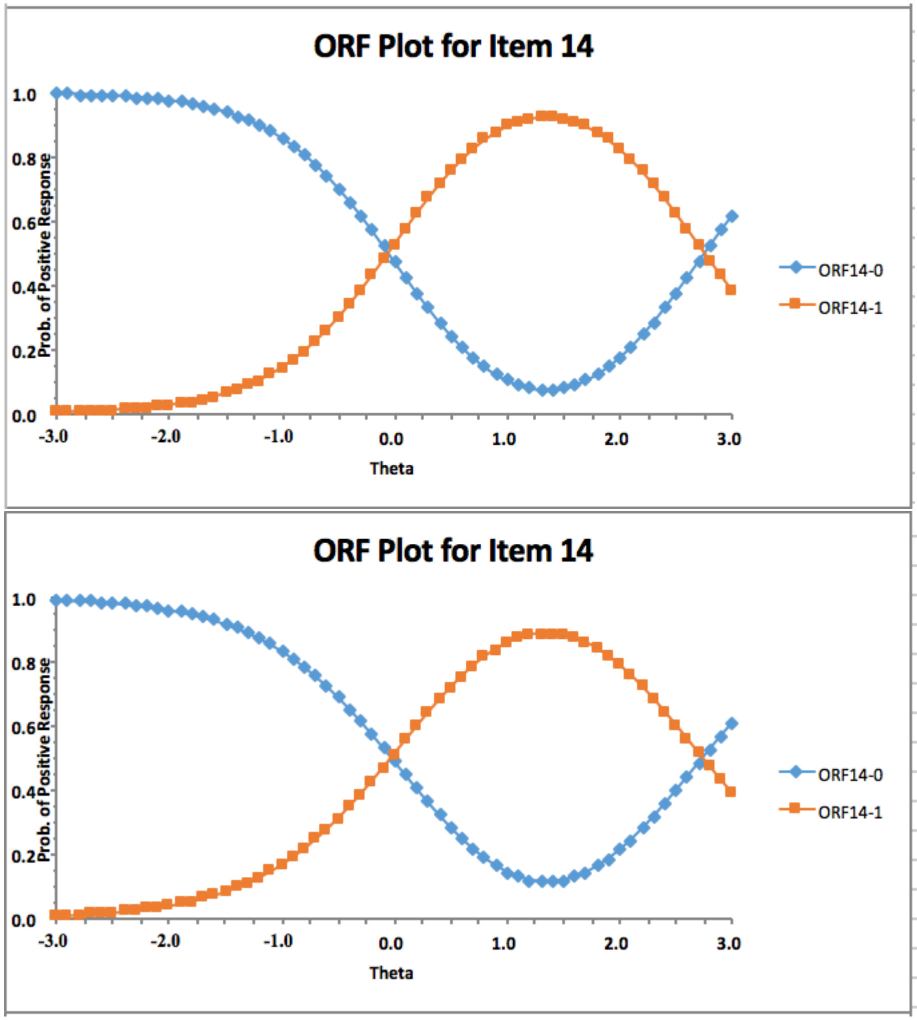


Figure 3. IRT ICCs for Item 14 ("***Sometimes I wish that everyone was as organized as me***") of the scale containing the 20 items from the CPS Order scale and 9 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the untrained group only.

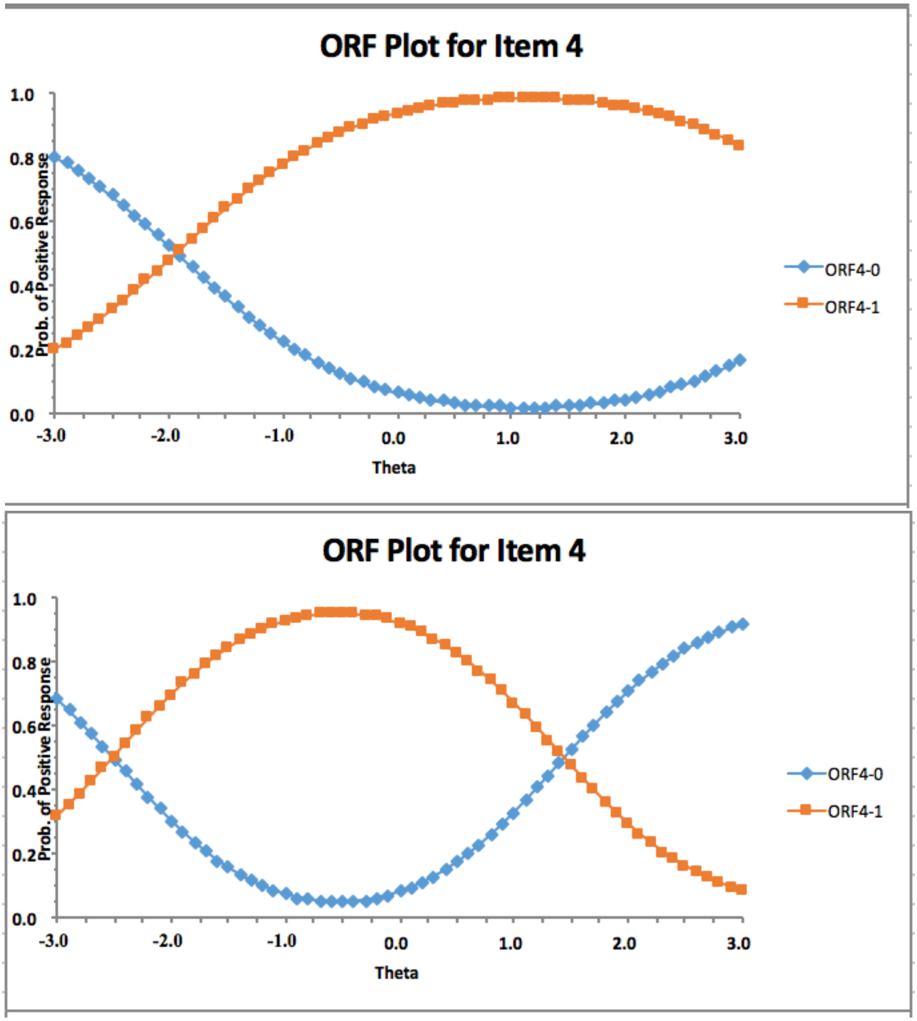


Figure 4. IRT ICCs for Item 4 ("***I am usually intrigued by what I learn in classes***") of the scale containing both the 20 items of the CPS Curiosity scale and 8 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

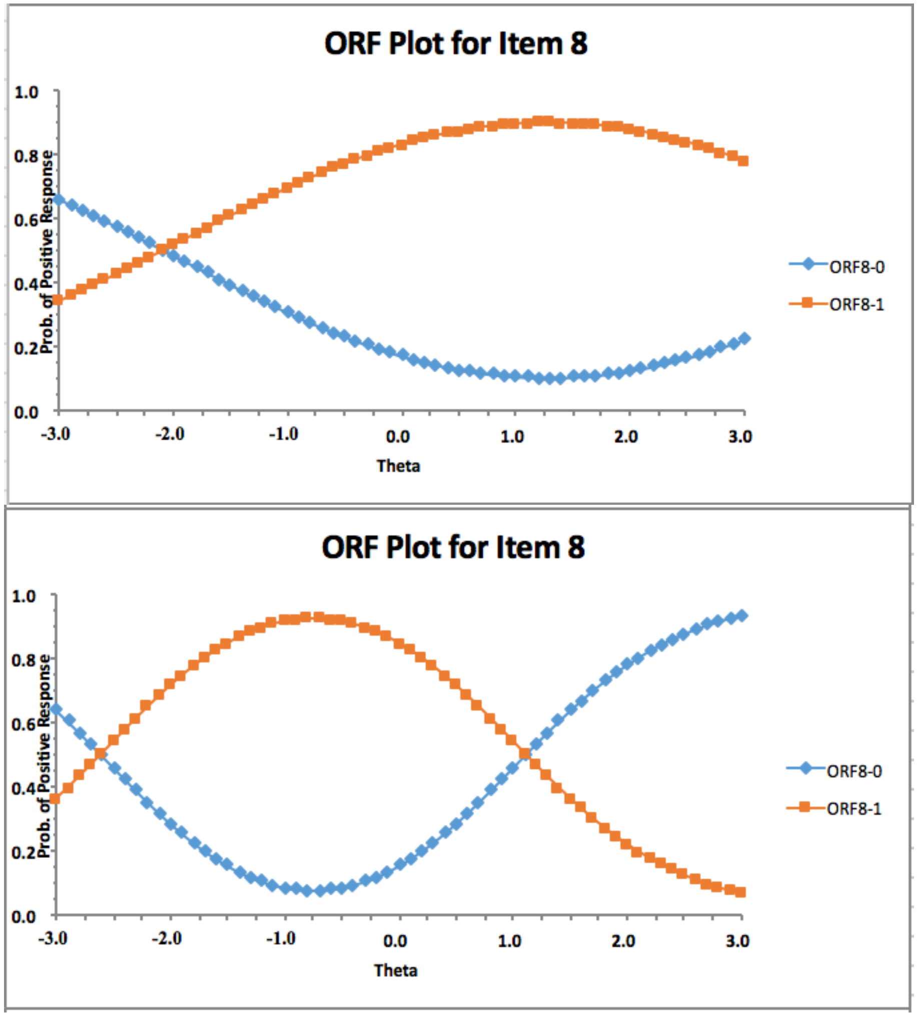


Figure 5. IRT ICCs for Item 8 ("***I sometimes read non-fiction books to learn something new***") of the scale containing both 20 items of the CPS Curiosity scale and 8 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

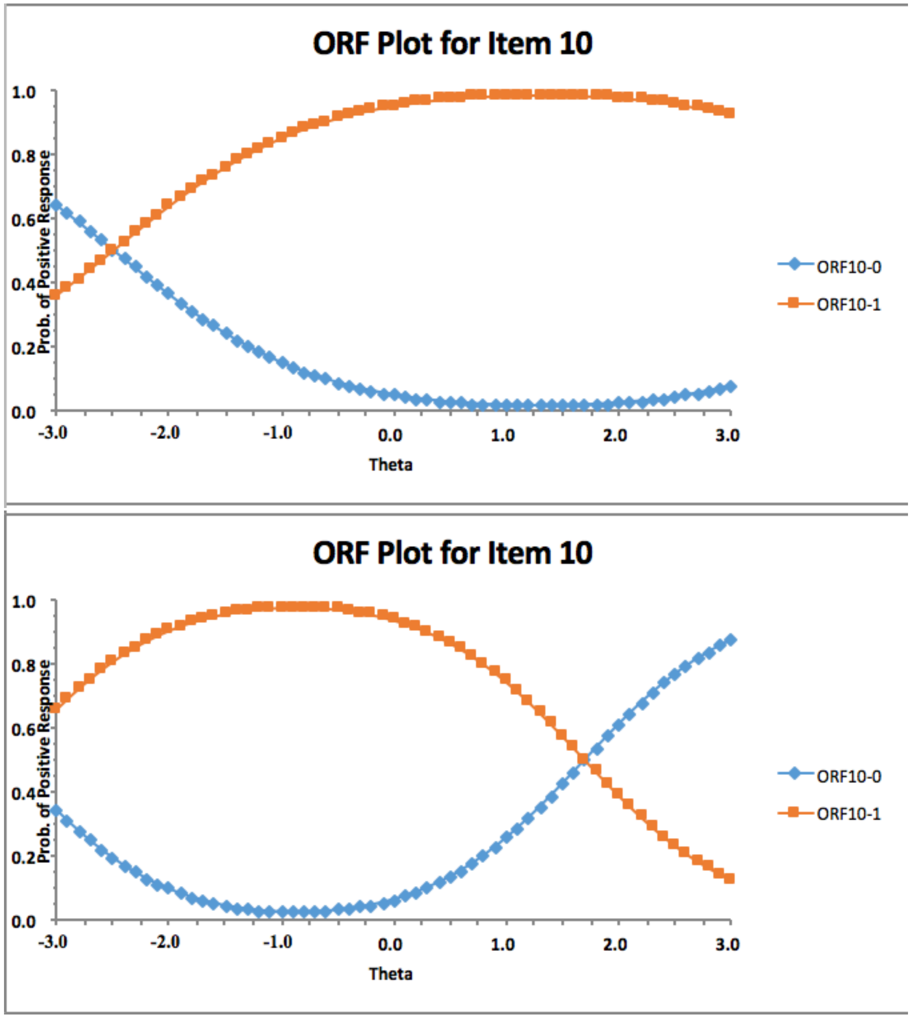


Figure 6. IRT ICCs for Item 10 ("***I am interested in what is happening around the world***") of the scale containing both 20 items of the CPS Curiosity scale and 8 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

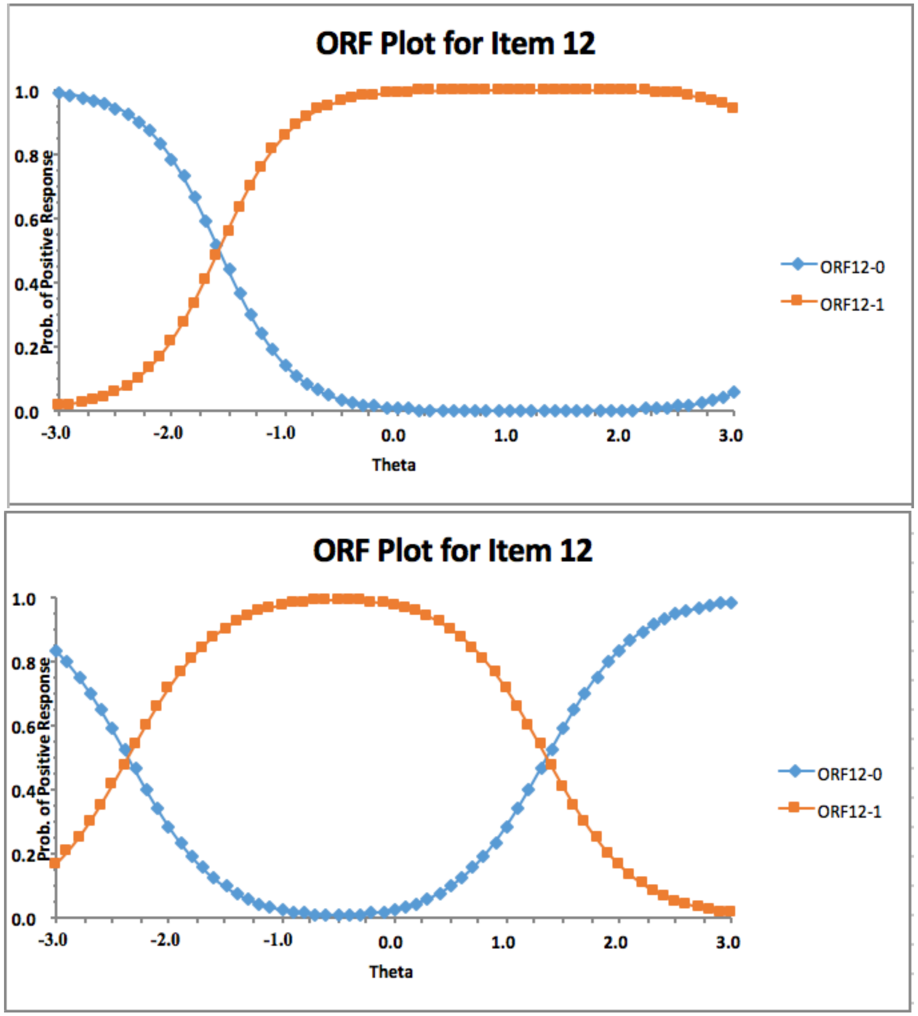


Figure 7. IRT ICCs for Item 12 ("***I like to learn new things whenever I have time***") of the scale containing both 20 items of the CPS Curiosity scale and 8 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

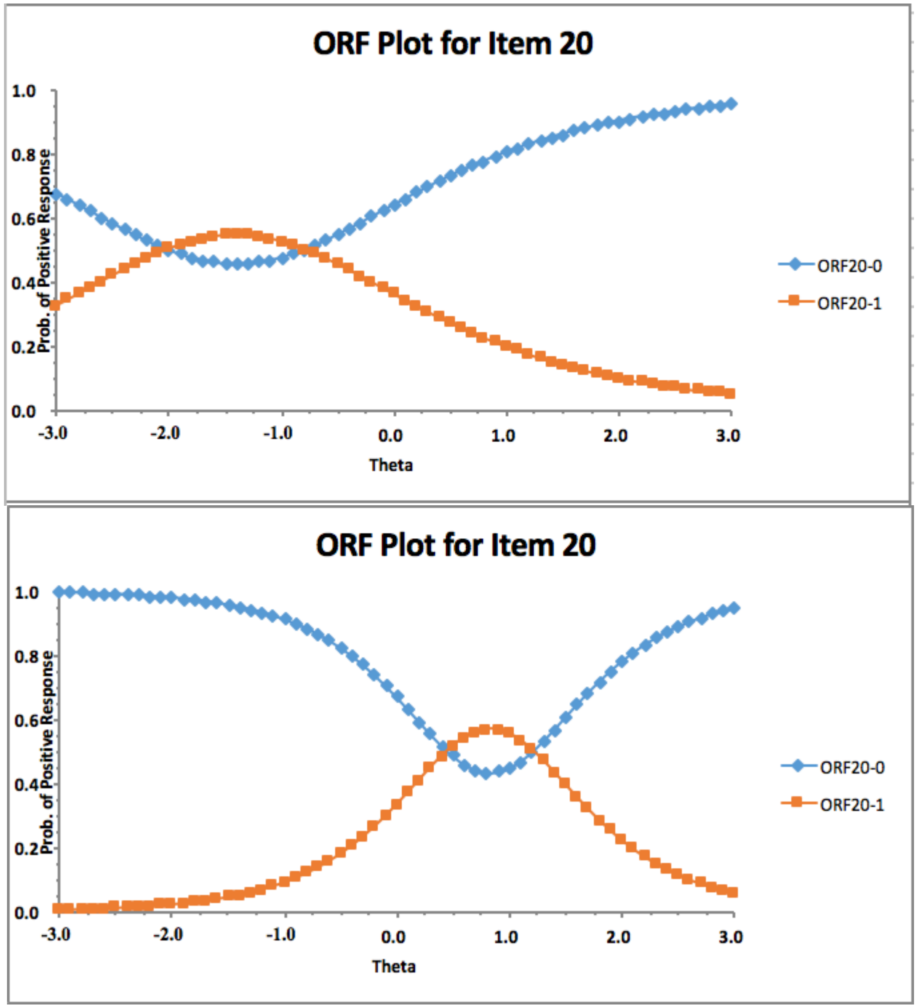
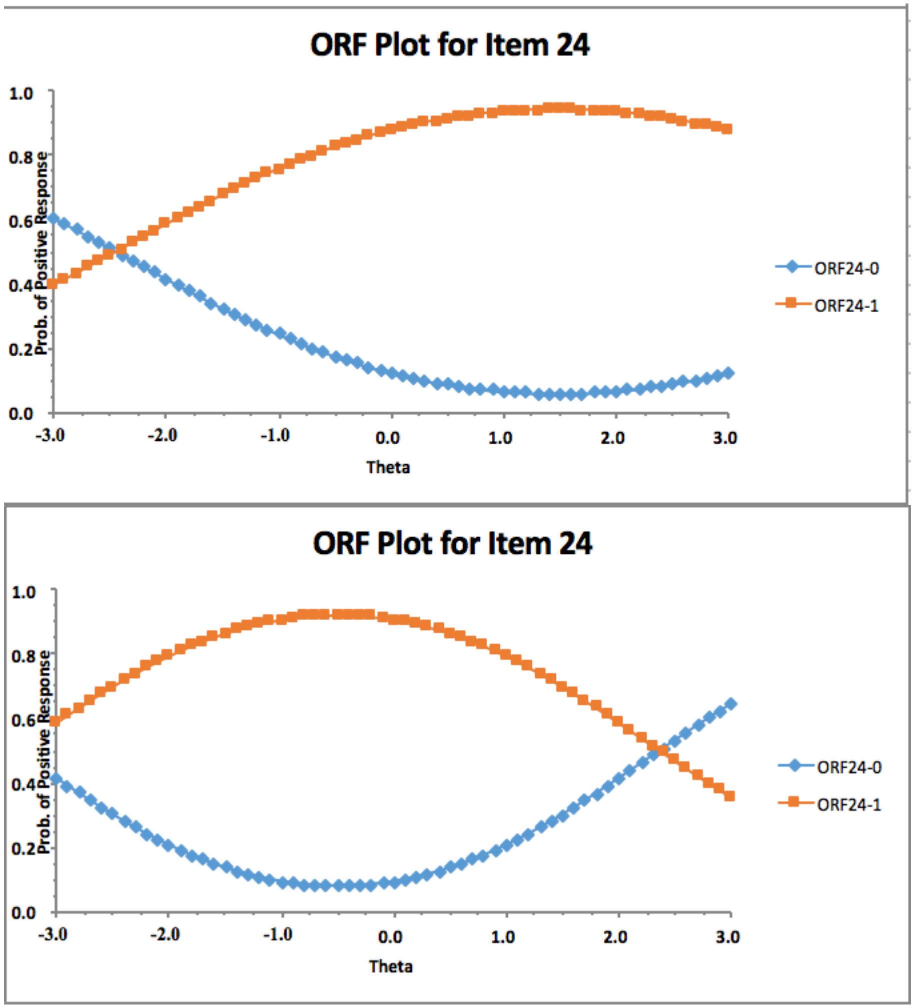


Figure 8. IRT ICCs for Item 20 ("***I try new restaurants only when other people recommend them***") of the scale containing both 20 items of the CPS Curiosity scale and 8 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.


ORF Plot for Item 24

ORF Plot for Item 24

Figure 9. IRT ICCs for Item 24 (“***Occasionally I find myself interested in information that I really don’t need***”) of the scale containing both 20 items of the CPS Curiosity scale and 8 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

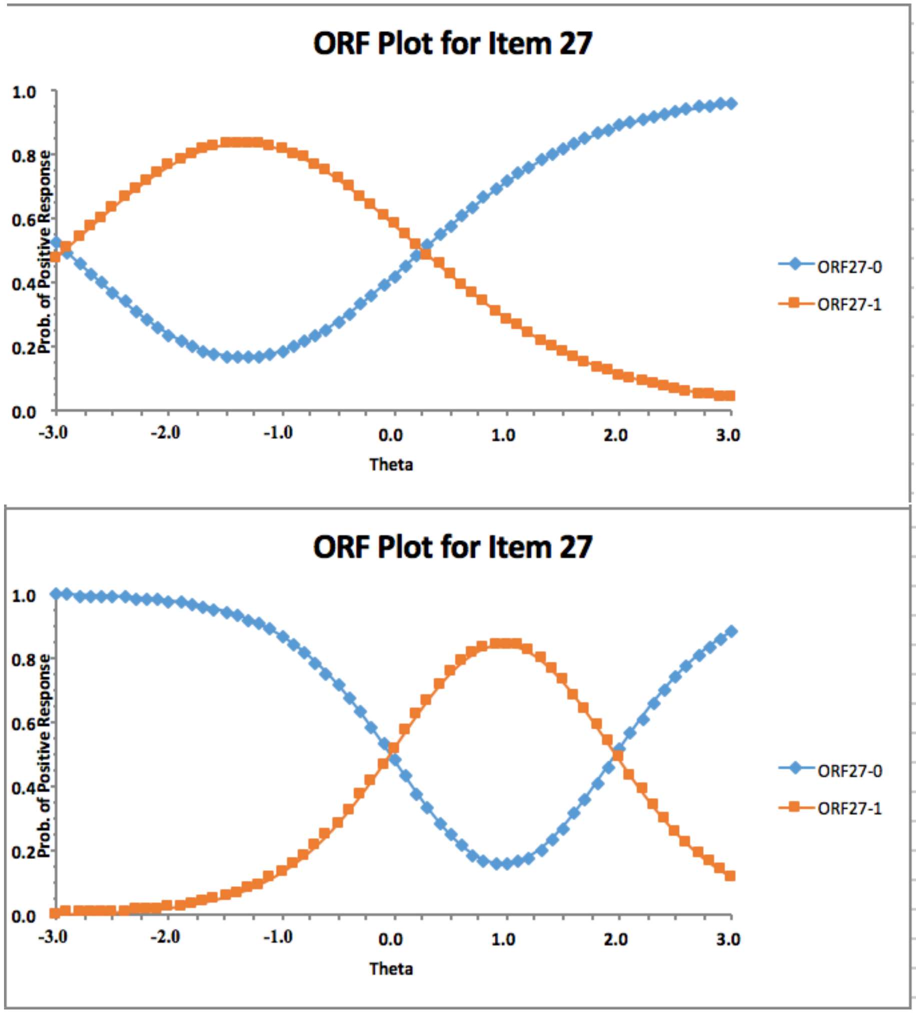


Figure 10. IRT ICCs for Item 27 ("***I am about average in curiosity about new knowledge***") of the scale containing both 20 items of the CPS Curiosity scale and 8 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

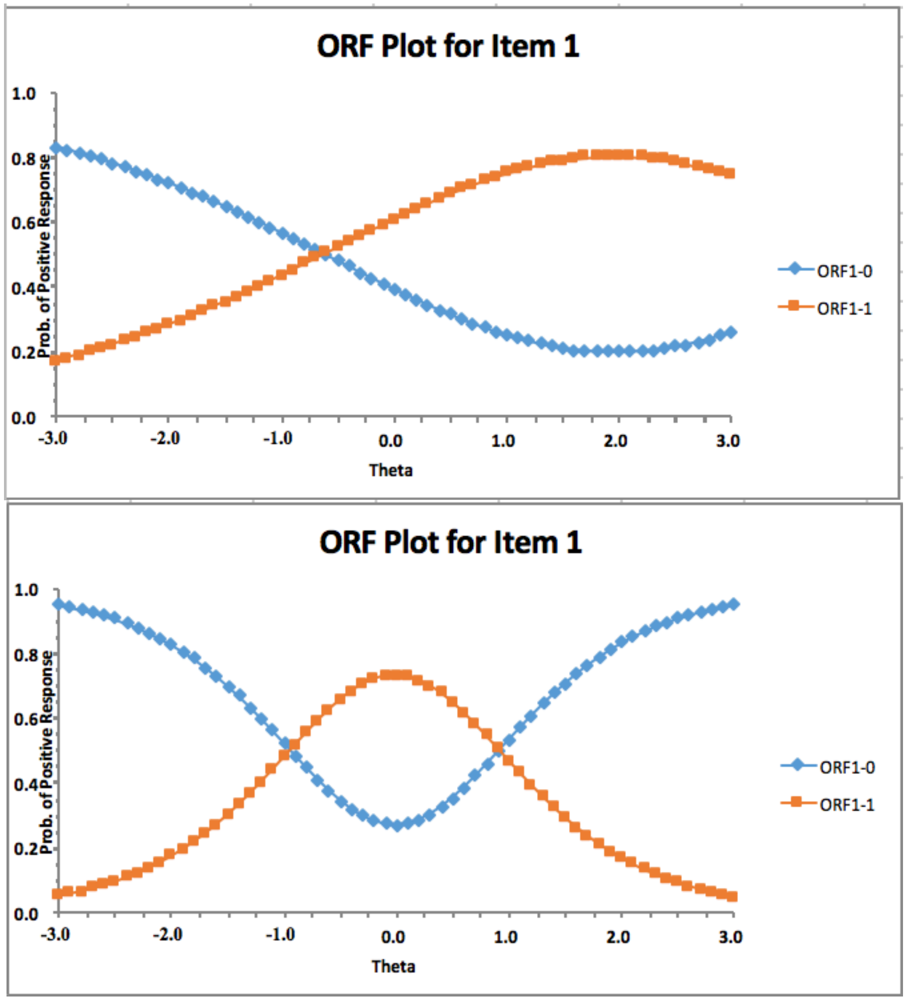


Figure 11. IRT ICCs for Item 1 ("***I am competitive and play to win***") of the CPS Industriousness scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

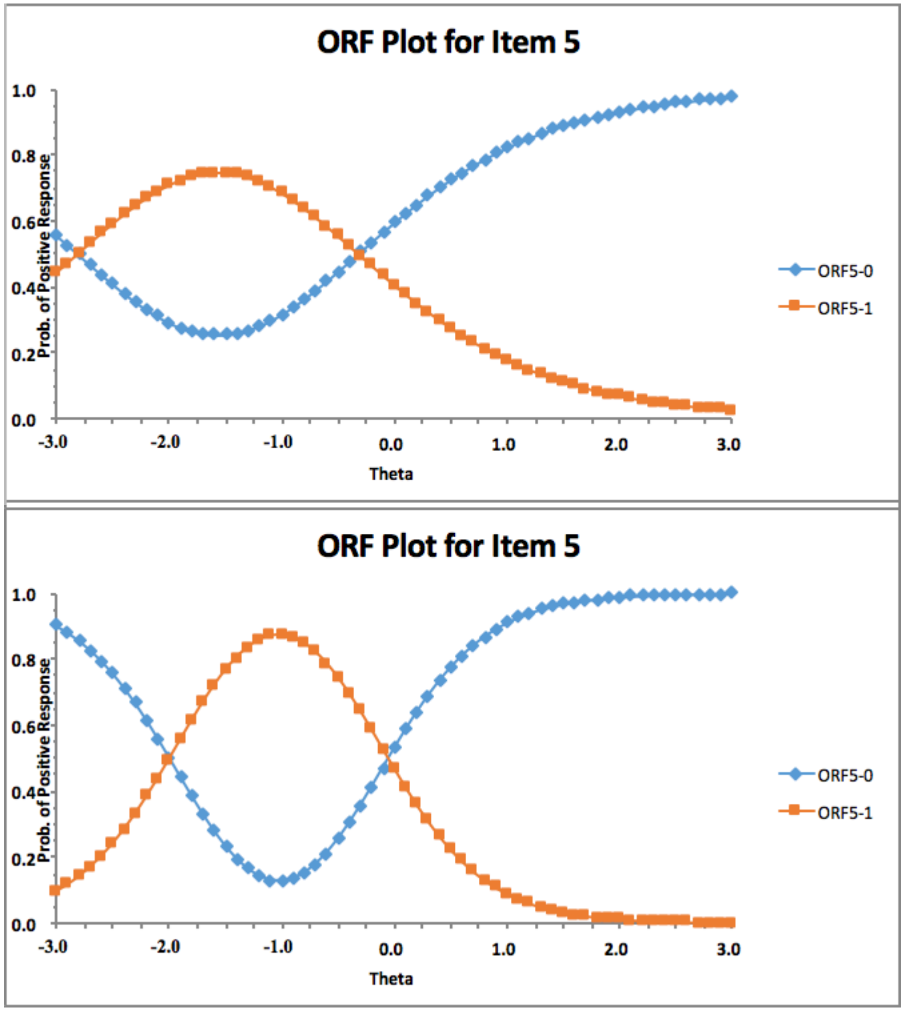


Figure 12. IRT ICCs for Item 5 ("***I finish my work on time but try not to work more than I have to***") of the CPS Industriousness scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

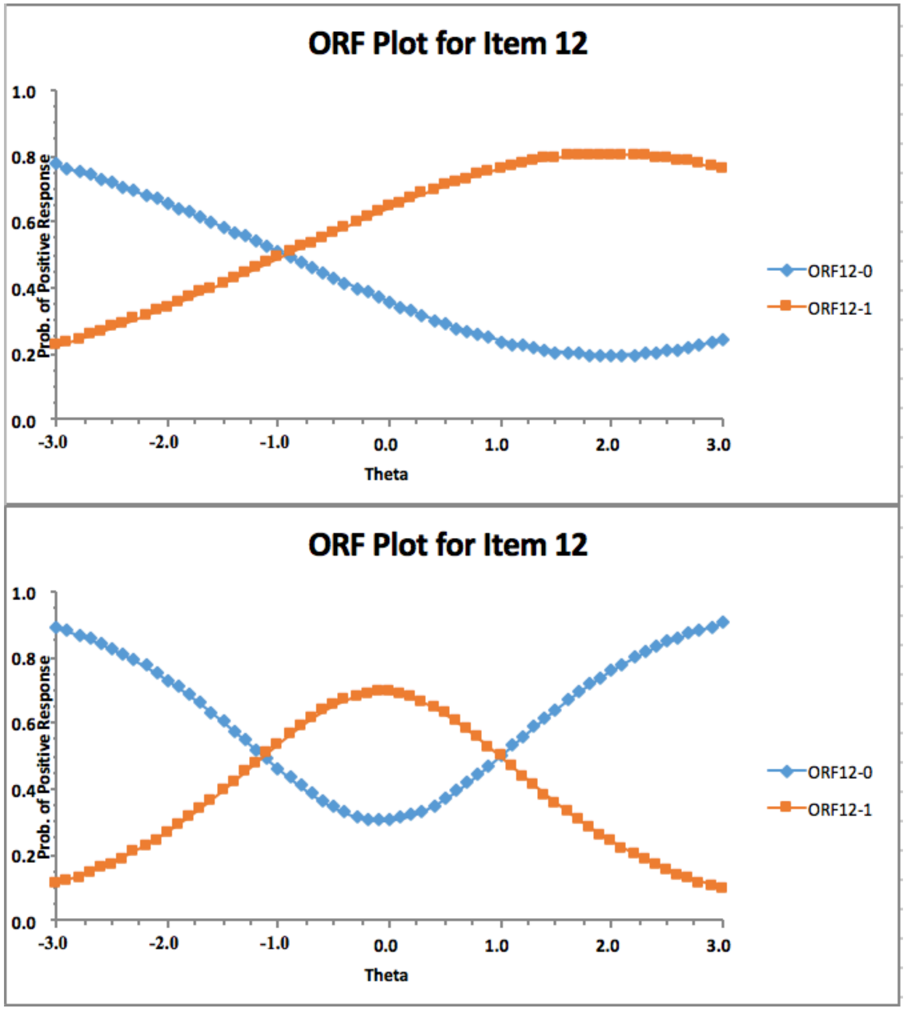


Figure 13. IRT ICCs for Item 12 ("***I always want to be better than others in the things I do***") of the CPS Industriousness scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

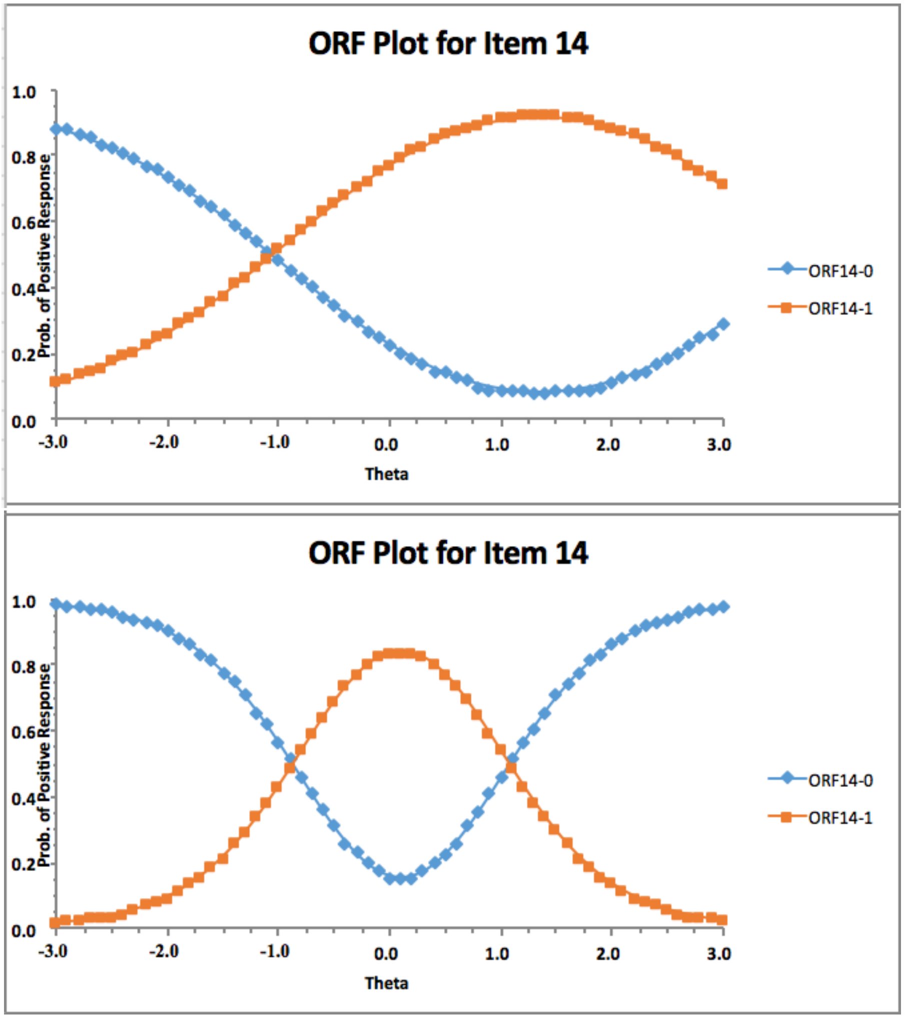


Figure 14. IRT ICCs for Item 14 ("***When I set my mind on achieving a goal, I can always reach it***") of the CPS Industriousness scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

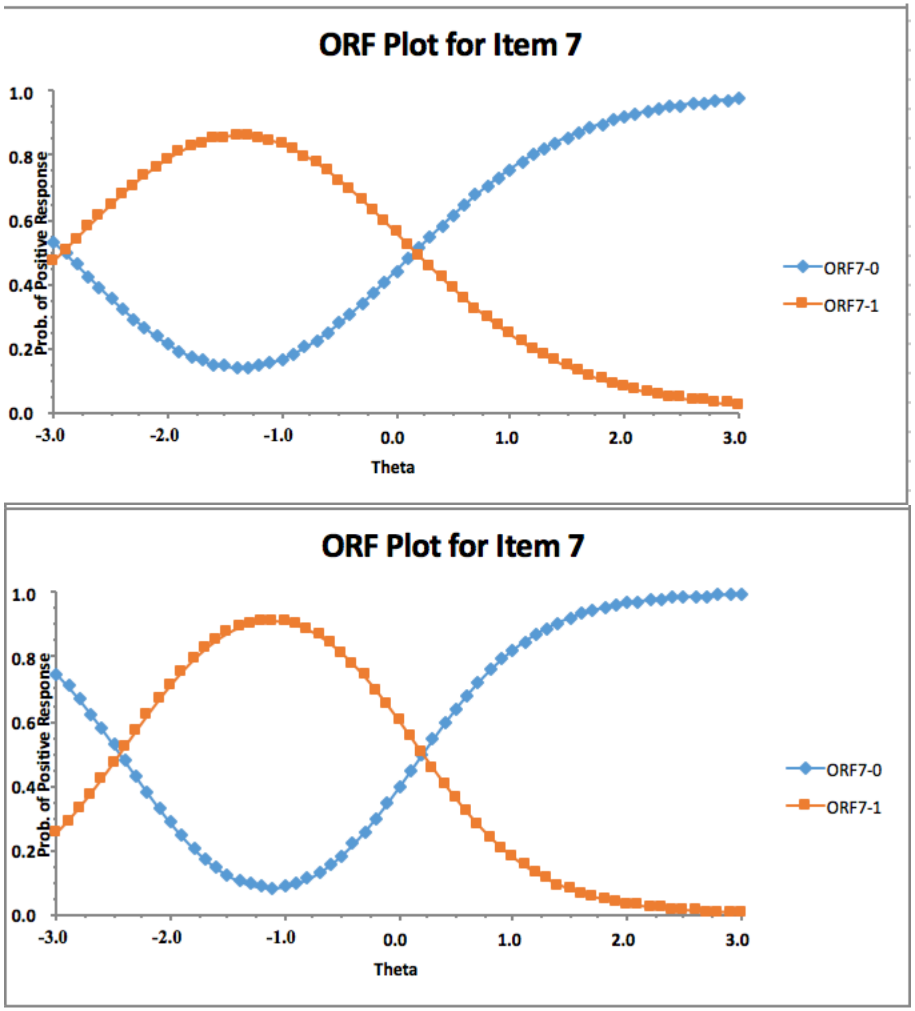


Figure 15. IRT ICCs for Item 7 ("***I can ignore a mess for a long time, but eventually I have to clean it up***") of the 20 items of the CPS Order scale and 9 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

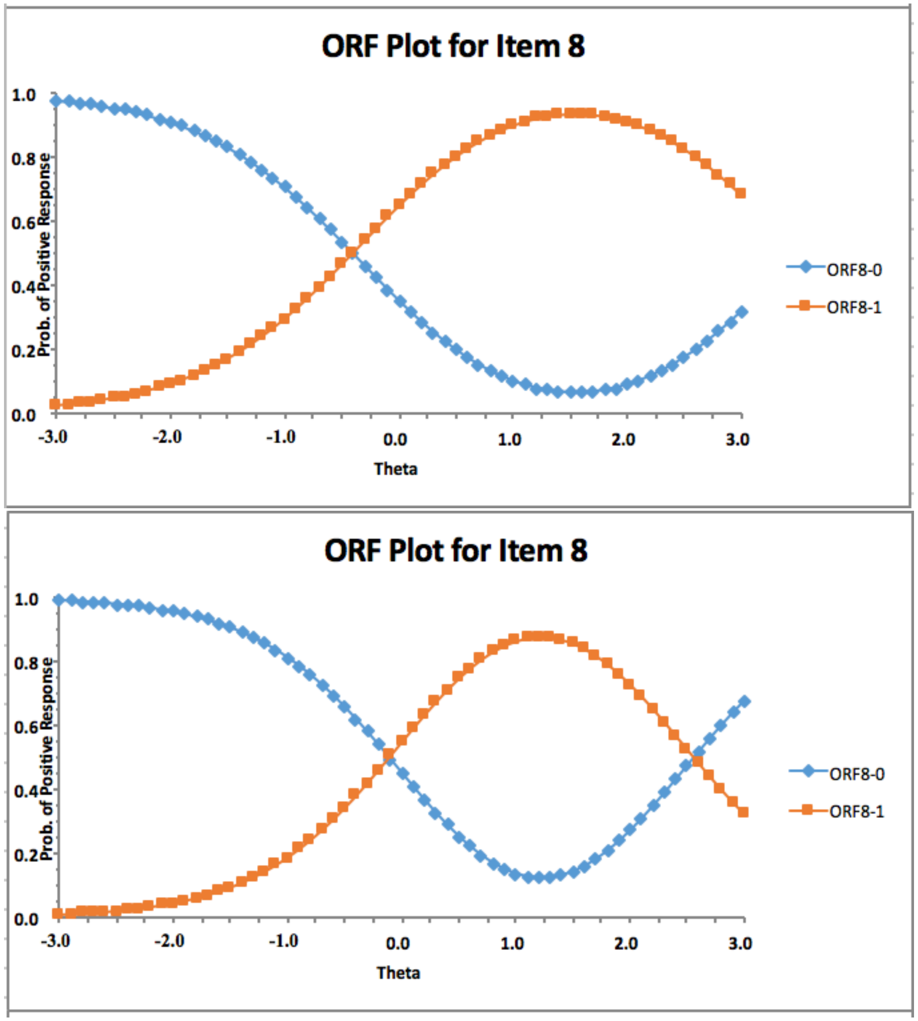


Figure 16. IRT ICCs for Item 8 ("***I plan my time very carefully***") of the 20 items of the CPS Order scale and 9 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

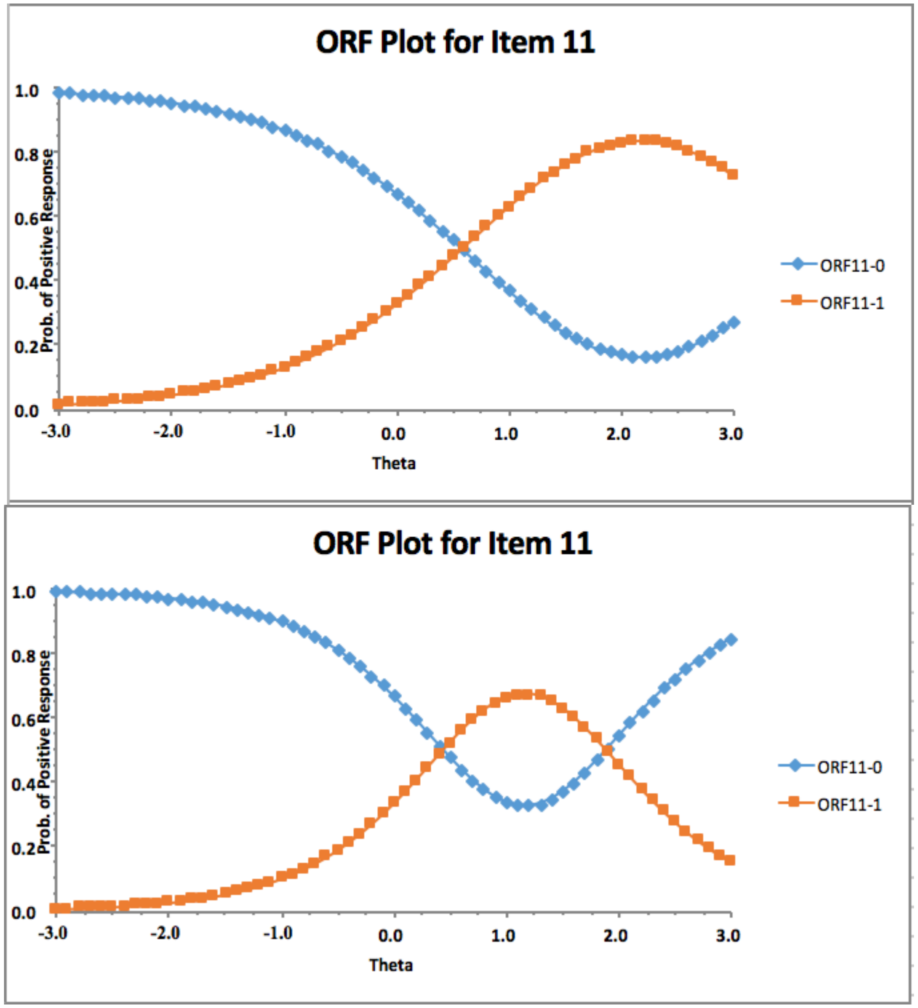


Figure 17. IRT ICCs for Item 11 ("***I follow a strict daily schedule***") of the 20 items of the CPS Order scale and 9 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

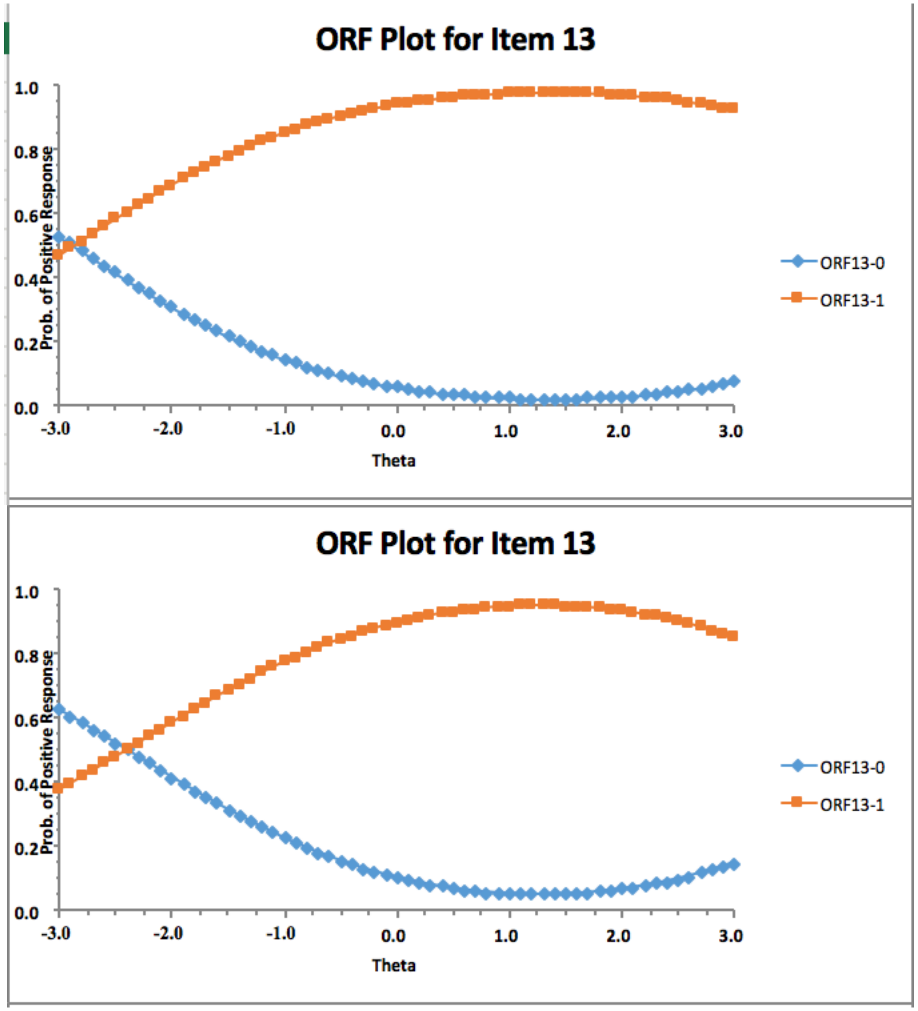
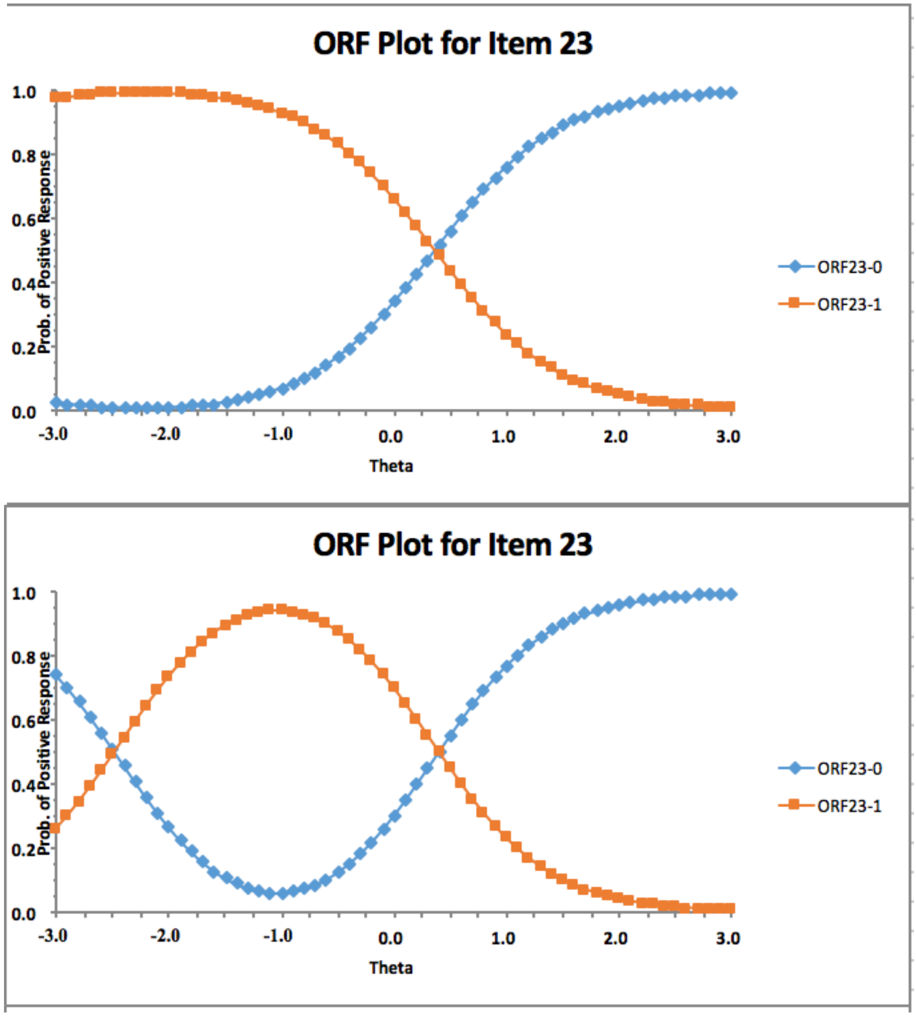


Figure 18. IRT ICCs for Item 13 ("***When I have many things to do, I try to focus on the task with the highest priority first***") of the 20 items of the CPS Order scale and 9 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.


ORF Plot for Item 23

ORF Plot for Item 23

Figure 19. IRT ICCs for Item 23 ("***Sometimes I can tolerate the messiness of my room***") of the 20 items of the CPS Order scale and 9 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

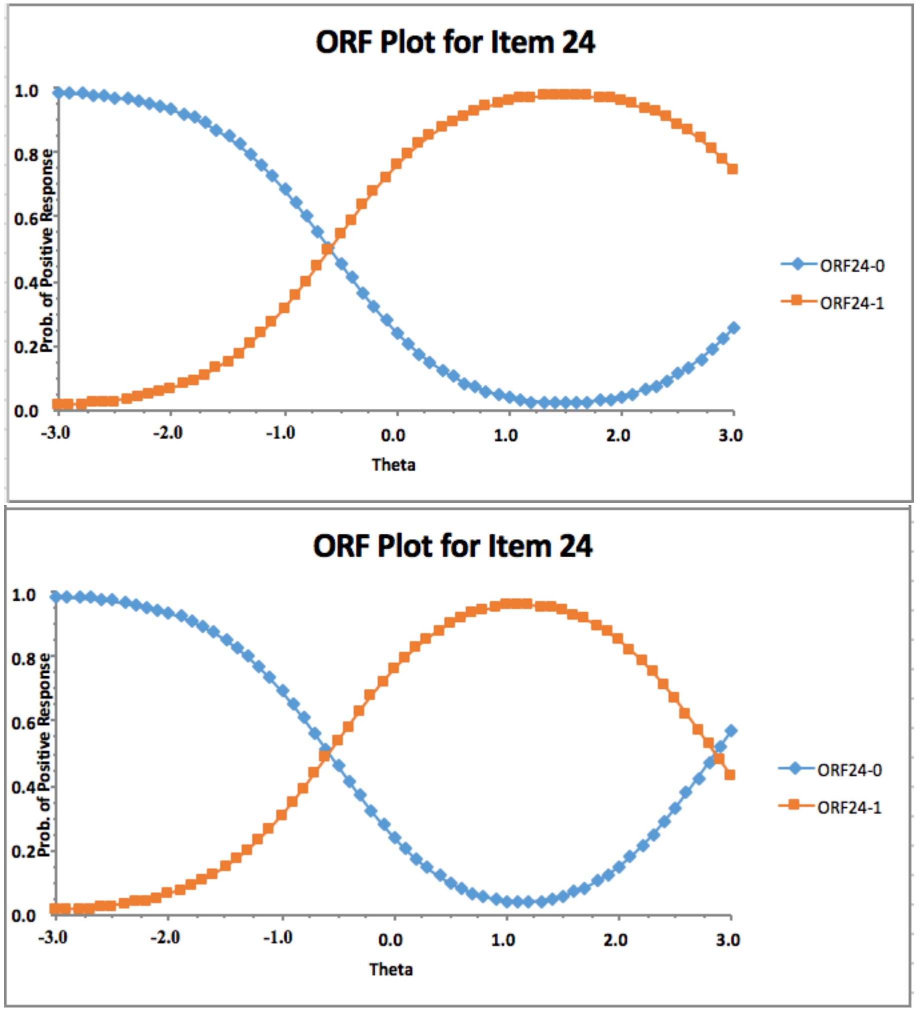


Figure 20. IRT ICCs for Item 24 ("***I spend time cleaning and organizing things when I am not busy***") of the 20 items of the CPS Order scale and 9 intermediate items from Cao et al. (2015) for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

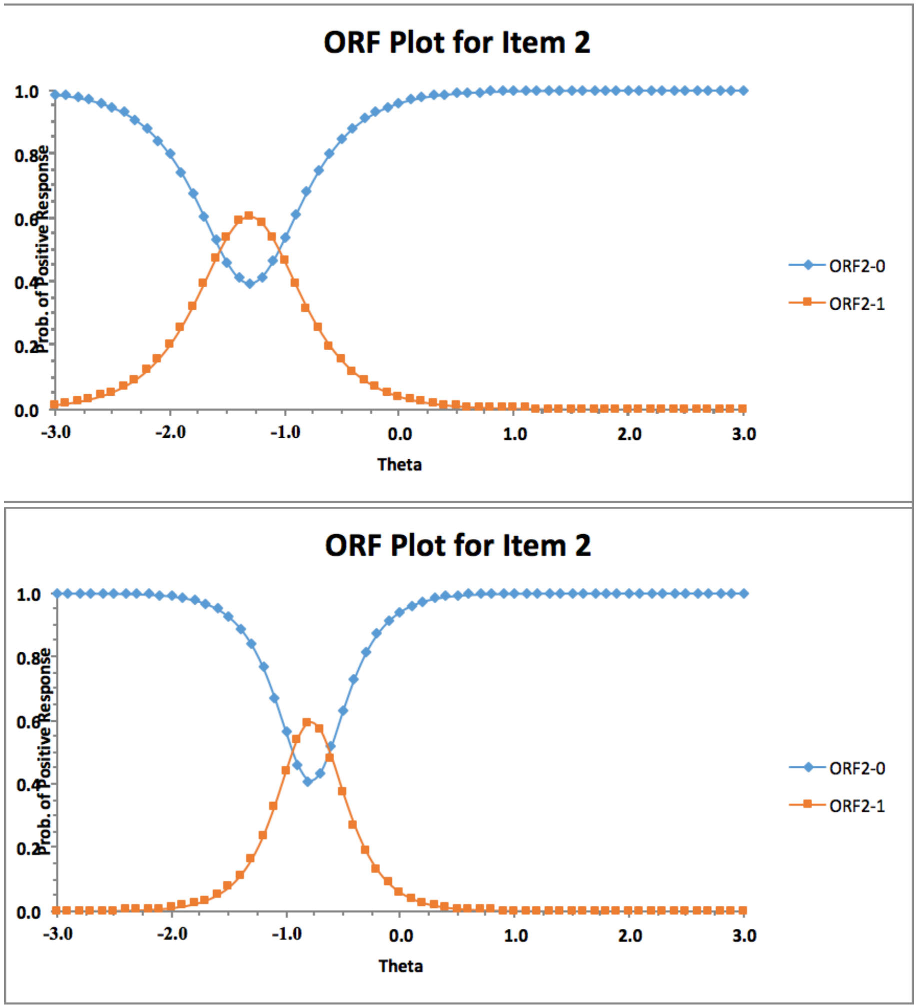


Figure 21. IRT ICCs for Item 2 ("***I have often missed important meetings because I forgot them***") of the CPS Self-control scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

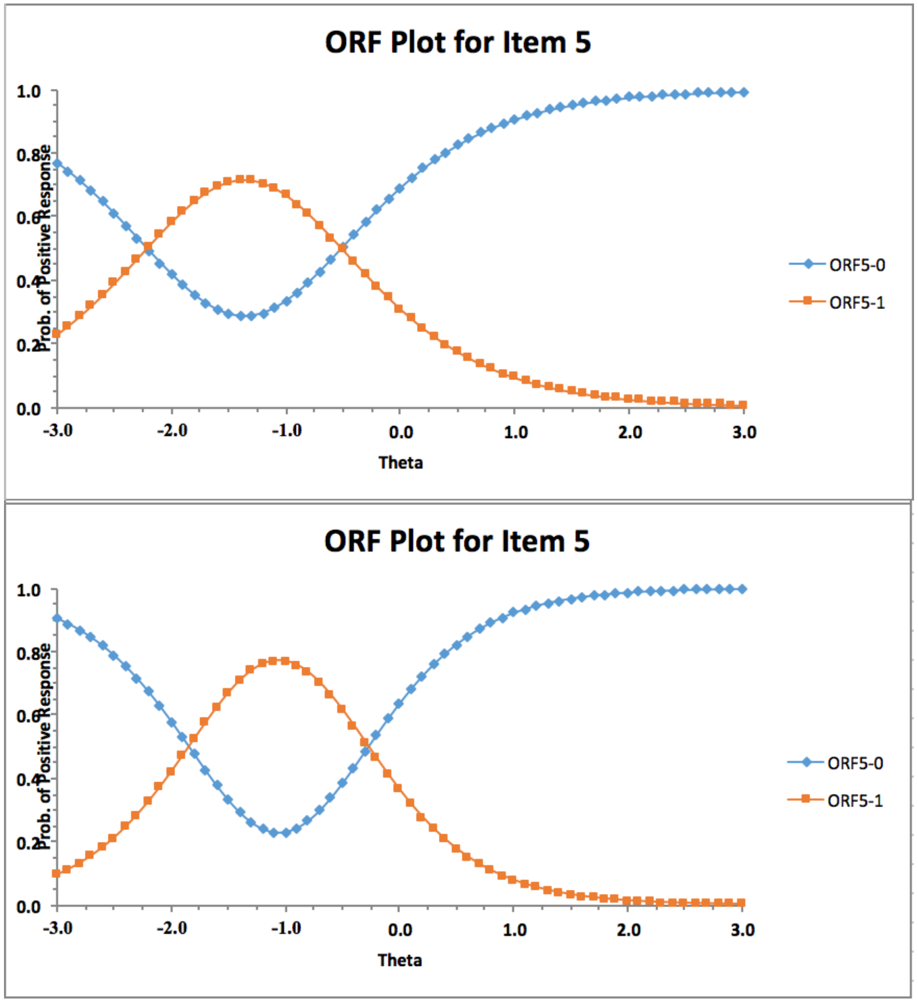


Figure 22. IRT ICCs for Item 5 ("***Keeping a careful record of things is not my strength***") of the CPS Self-control scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

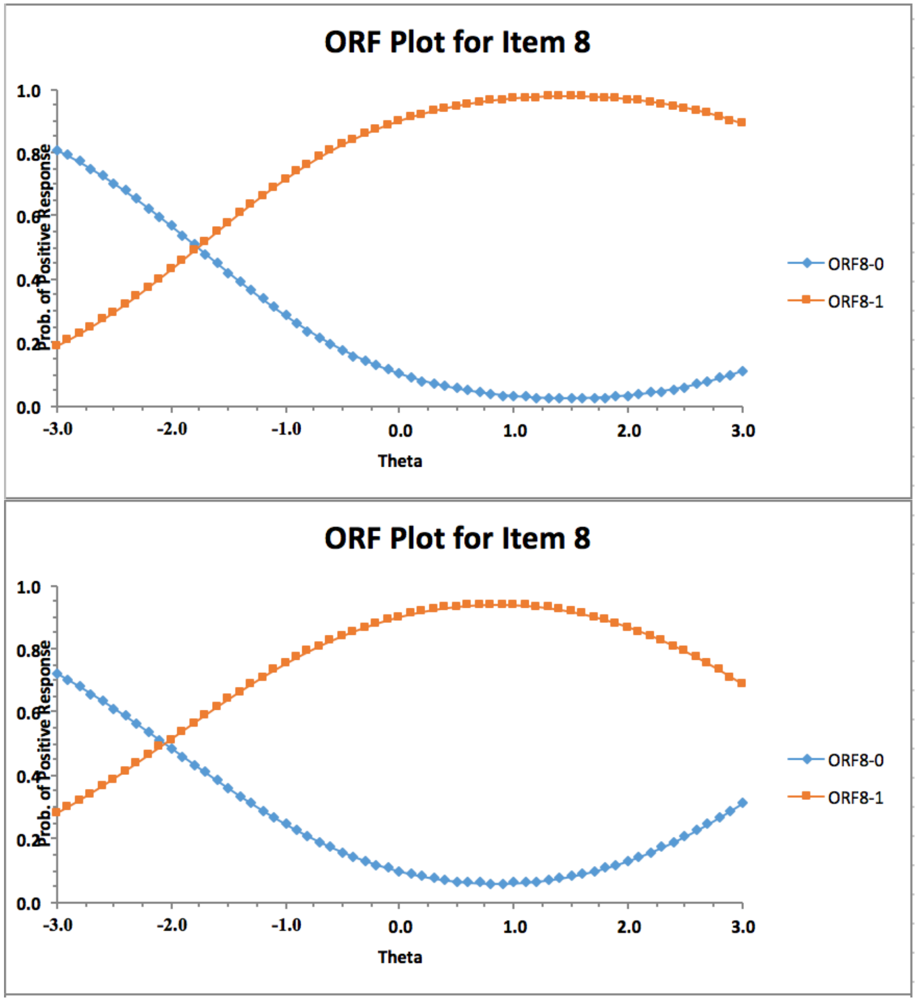


Figure 23. IRT ICCs for Item 8 ("***I am usually cautious***") of the CPS Self-control scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

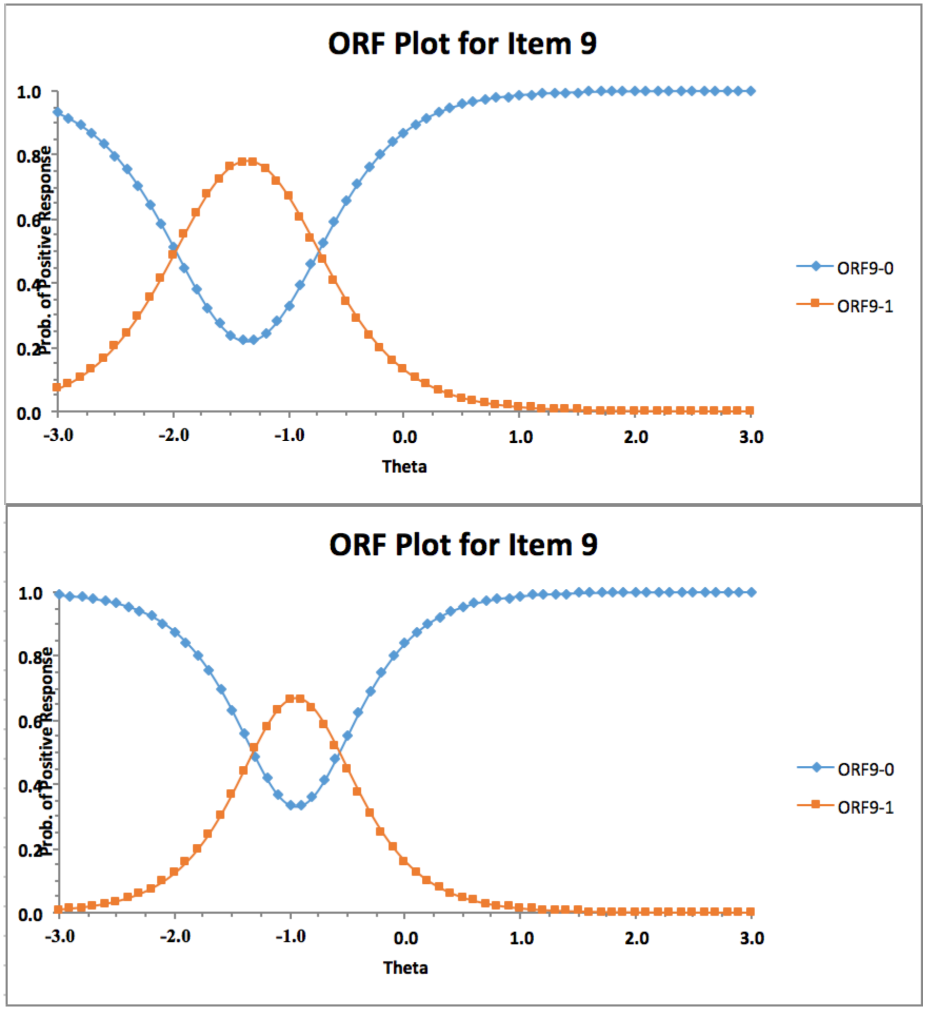


Figure 24. IRT ICCs for Item 9 ("***I often make careless mistakes***") of the CPS Self-control scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

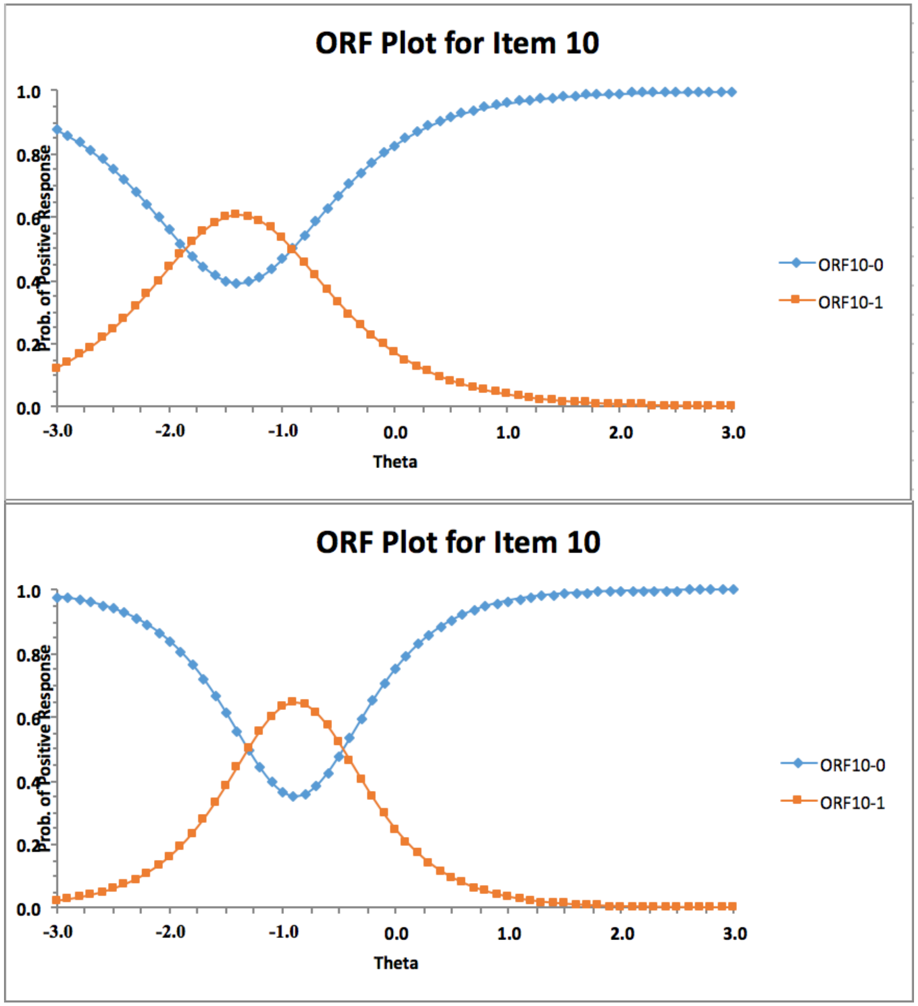


Figure 25. IRT ICCs for Item 10 ("***I can keep my concentration only on short tasks***") of the CPS Self-control scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

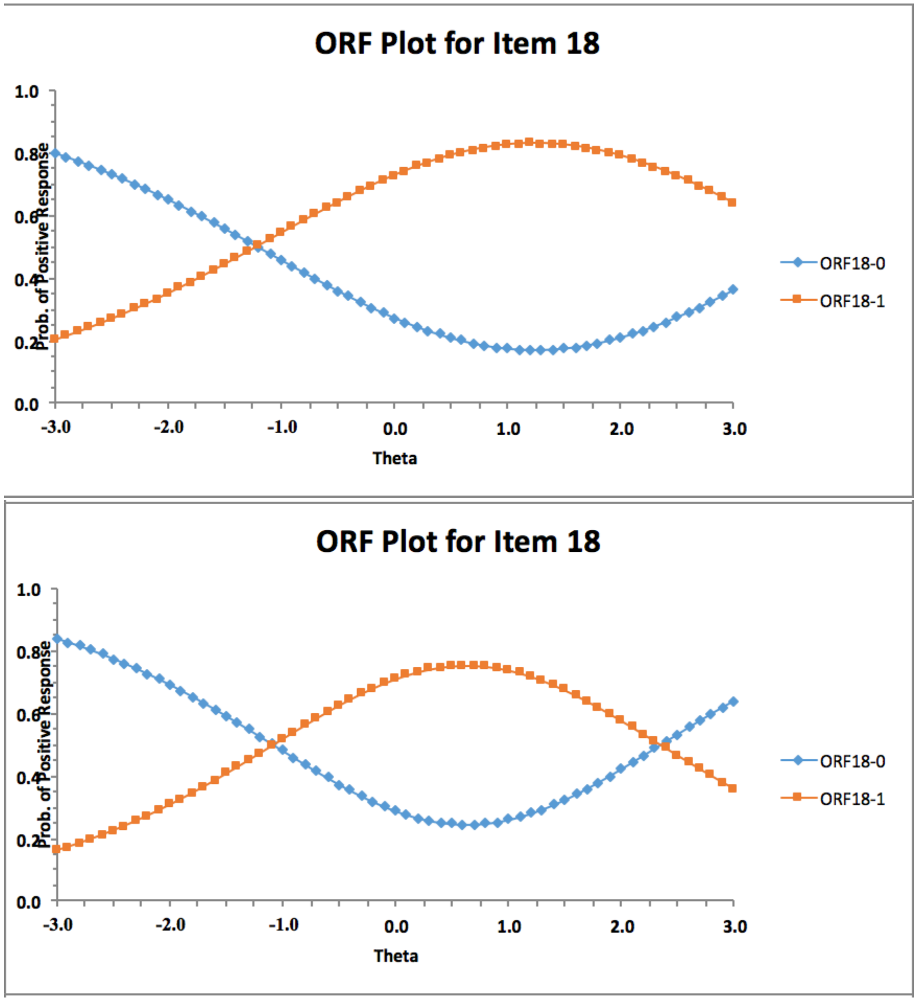
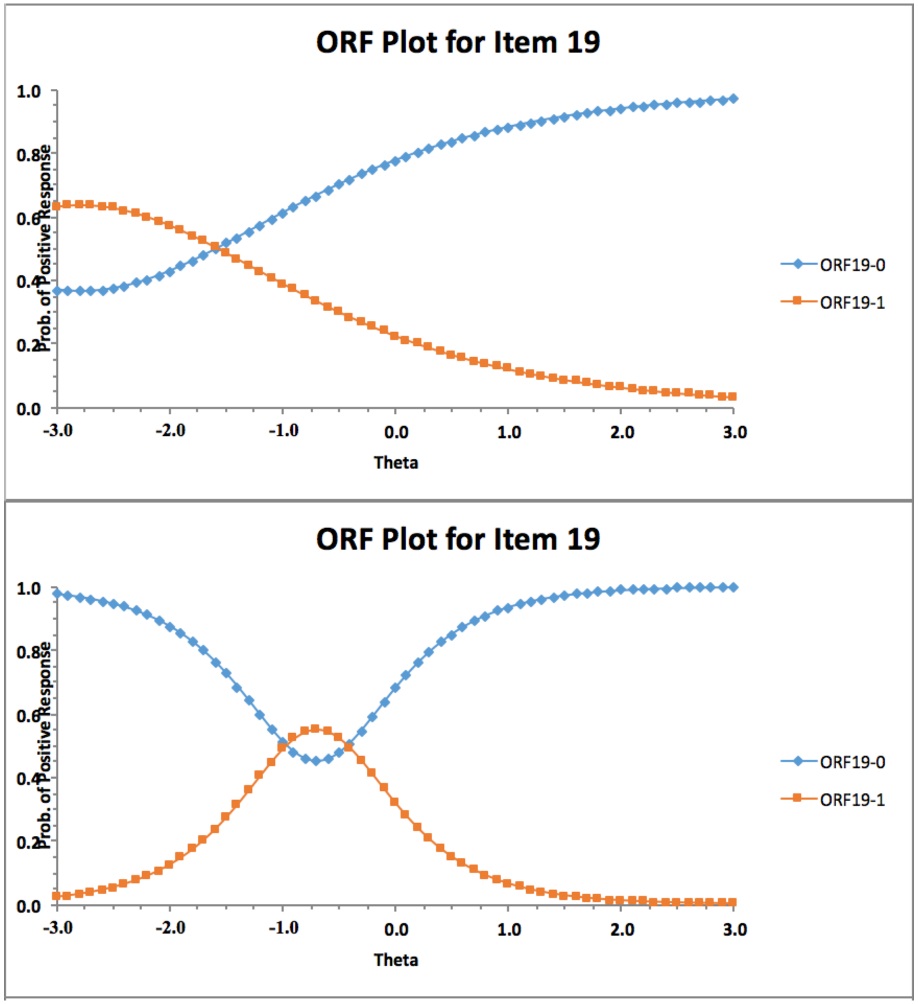


Figure 26. IRT ICCs for Item 18 ("***I believe people can never be too careful***") of the CPS Self-control scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.


ORF Plot for Item 19

ORF Plot for Item 19

Figure 27. IRT ICCs for Item 19 ("***I am only careful on tasks that are important to me***") of the CPS Self-control scale for the untrained (T) and trained (B) groups, respectively. This item is an intermediate item in the trained group only.

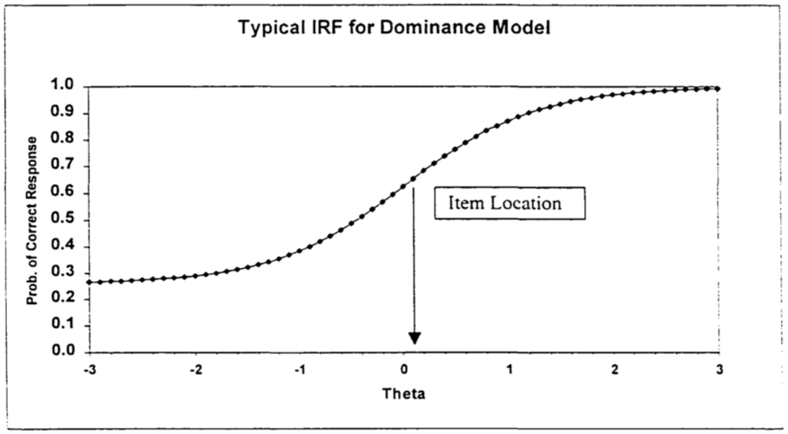


Figure 28. Typical Item Response Function (IRF) for a Dichotomous Dominance Model

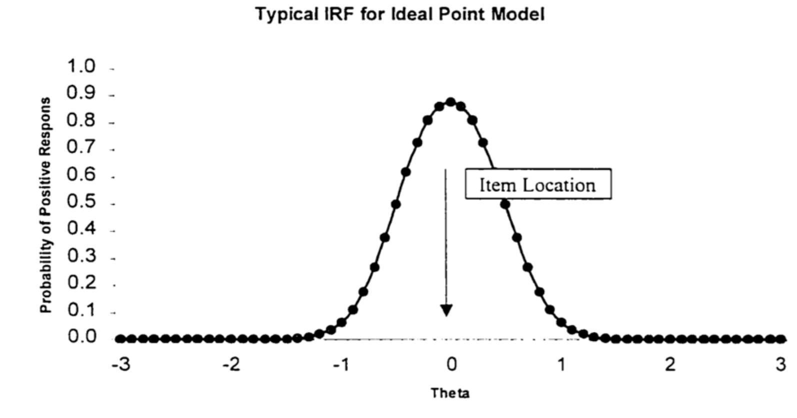


Figure 29. Ideal Point IRF for Item with Neutral Location

REFERENCES

Allen, T. D., Facteau, J. D., & Facteau, C. L. (2004). Structured interviewing for OCB: Construct validity, faking, and the effects of question type. *Human Performance, 17*(1), 1-24. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1207/S15327043HUP1701_1

Anderson, G., and Viswesvaran, C. (1998) '*An update of the validity of personality scales in personal selection: A meta analysis of studies published after 1992*'. Paper presented at the 13th Annual Conference of the Society of Industrial and Organizational Psychology, Dallas.

Ashton, M. C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior, 19*, 289–303.

Avdic, A. (2013). *Criterion-related validity of narrow-trait personality for predicting job performance, and the test of mediating mechanisms* (Order No. AAI3514364). Available from PsycINFO. (1399052876; 2013-99100-313). Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/1399052876?accountid=14553

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1-26. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*(1-2), 9-30. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/1468-2389.00160

Barrick, M. R., Mount, M. K., & Strauss, J. P. (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. *Journal of Applied Psychology, 78,* 715–722.

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology, 63*(3), 301-308. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.63.3.301

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*(1), 60-66. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.65.1.60

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 62*(1), 64-69. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.62.1.64

Bidjerano, T., & Dai, D. Y. (2007). The relationship between the Big-Five model of personality and self-regulated learning strategies. *Learning and Individual Differences, 17,* 69–81.

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin, 130*(6), 887-919. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0033-2909.130.6.887

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60*(5), 556-560. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.60.5.556

Broadfoot, A. A. (2008). *Comparing the dominance approach to the ideal-point approach in the measurement and predictability of personality* (Order No. AAI3316810). Available from PsycINFO. (621760437; 2008-99240-116). Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/621760437?accountid=14553

Brown, A., & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3,* 489–493.

Cao, M., Drasgow, F., & Cho, S. (2015). Developing ideal intermediate personality items for the ideal point model. *Organizational Research Methods, 18*(2), 252-275. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1177/1094428114555993

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI work satisfaction scale. *Personality and Individual Differences, 49*(7), 743-748. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1016/j.paid.2010.06.019

Carter, N. T., Dalal, D. K., Guan, L., LoPilato, A. C., & Withrow, S. A. (2017). Item response theory scoring and the detection of curvilinear relationships. *Psychological Methods, 22*(1), 191-203. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/met0000101

Carter, N. T., Dalal, D. K., Lake, C. J., Lin, B. C., & Zickar, M. J. (2011). Using mixed-model item response theory to analyze organizational survey responses: An illustration using the Job Descriptive Index. *Organizational Research Methods, 14,* 116–146.

Cattell, R. B., & Cattell, H. E. P. (1995). Personality structure and the new fifth edition of the 16PF. *Educational and Psychological Measurement, 55*(6), 926-937. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1177/0013164495055006002

Chernyshenko, O. S. (2003). *Applications of ideal point approaches to scale construction and scoring in personality measurement: The development of a six-faceted measure of conscientiousness* (Order No. AAI3070273). Available from PsycINFO. (620233080; 2003-95010-007). Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/620233080?accountid=14553

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*(4), 523-562. doi:http://dx.doi.org.proxy2.library.illinois.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. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/1040-3590.19.1.88

Conn, S., & Rieke, M. L. (Eds.). (1994). *The 16PF fifth edition technical manual.* Champaign, IL: Institute for Personality and Ability Testing.

Coombs, C. H. (1964). *A theory of data* Wiley, Oxford. Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/615430669?accountid=14553

Costa, P. T., Busch, C. M., Zonderman, A. B., & McCrae, R. R. (1986). Correlations of MMPI factor scales with measures of the five factor model of personality. *Journal of Personality Assessment, 50*(4), 640-650. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1207/s15327752jpa5004_10

Costa, P. T., Jr., & McCrae, R. R. (1988). From catalog to classification: Murray's needs and the five-factor model. *Journal of Personality and Social Psychology, 55*(2), 258-265. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0022-3514.55.2.258

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Crede´, M. (2010). Two caveats for the use of ideal point items: Discrepancies and bivariate constructs. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3,* 494–497.

Dalal, D. K., Carter, N. T., & Lake, C. J. (2014). Middle response scale options are inappropriate for ideal point scales. *Journal of Business and Psychology, 29*(3), 463-478. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1007/s10869-013-9326-5

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. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/j.1754-9434.2010.01279.x

De Raad, B., & Schouwenburg, H. C. (1996). Personality in learning and education: A review. *European Journal of Personality, 10,* 303–336.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. Journal of Personality Assessment, 49, 71-75.

Digman, J. M. (1989). Five robust trait dimensions: Development, stability, and utility. *Journal of Personality, 57*(2), 195-214. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/j.1467-6494.1989.tb00480.x

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. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/j.1754-9434.2010.01273.x

Drasgow, F., Levine, M. V., Tsien, S., Williams, B. A., & Mead, A. D. (1995). Fitting polytomous item response models to multiple-choice tests. *Applied Psychological Measurement, 19*, 145-165.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology*, *91*, 40-57.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the trier personality inventory (TPI). *Journal of Cross-Cultural*

Psychology, 24(2), 133-148. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1177/0022022193242001

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116*(3), 429-456. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0033-2909.116.3.429

Friedman, H. S., Tucker, J. S., Tomlinson-Keasey, C., Schwartz, J. E., Wingard, D. L., & Criqui, M. H. (1993). Does childhood personality predict longevity? *Journal of Personality and Social Psychology, 65*(1), 176-185. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0022-3514.65.1.176

Goldberg, L. R. (1981). Unconfounding situational attributions from uncertain, neutral, and ambiguous ones: A psychometric analysis of descriptions of oneself and various types of others. *Journal of Personality and Social Psychology, 41*(3), 517-552. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0022-3514.41.3.517

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216-1229. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment, 4*(1), 26-42. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/1040-3590.4.1.26

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*(1), 26-34. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0003-066X.48.1.26

Goldberg, L. R. (1997). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R. (1998). *International personality item pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences.* Retrieved Feb 25, 2018, from http:// ipip.ori.org/

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Green, J. A., O'Connor, D. B., Gartland, N., & Roberts, B. W. (2016). The Chernyshenko conscientiousness scales: A new facet measure of conscientiousness. *Assessment, 23*(3), 374-385. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1177/1073191115580639

Guion, R.M. & Gottier, R.F. (1965) Validity of Personality Measures in Personnel Selection. *Personnel Psychology, 18,* 135–164.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* Sage Publications, Inc, Thousand Oaks, CA. Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/618050327?accountid=14553

Hanisch, K. A. (1992). The Job Descriptive Index revisited: Questions about the question mark. *Journal of Applied Psychology, 77,* 377–382.

Harrison, S. H. (2009). *Curiosity in organizations* (Order No. AAI3357265). Available from PsycINFO. (622083650; 2009-99210-171). Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/622083650?accountid=14553

Harter, S. (1990). Causes, correlates, and the functional role of global self-worth: A life-span perspective. In R. J. Sternberg, & J. Kolligian Jr. (Eds.), *Competence considered; competence considered* (pp. 67-97, Chapter xv, 420 Pages) Yale University Press, New Haven, CT. Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/617861374?accountid=14553

Heath C., *Heath* D. (*2006*). *The Curse of Knowledge*, *Harvard Business Review.* Retrieved Feb 25, 2018 from https://hbr.org/2006/12/the-curse-of-knowledge

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology, 73*(1), 68-73. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.73.1.68

Hernández, A., Drasgow, F., & González-Romá, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology, 89,* 687–699.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*(1), 100-112. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.88.1.100

Hough, L. M. (1996). Personality at work: Issues and evidence. In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection*. Hillsdale, NJ: Erlbaum.

Hough, L. M. & Schneider, R. J. (1996). Personality traits, taxonomies, and applications in organizations. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations*. San Francisco, CA: Jossey Bass.

Huang, J., & Mead, A. D. (2014). Effect of personality item writing on psychometric properties of ideal-point and Likert scales. *Psychological Assessment, 26*(4), 1162-1172. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/a0037273

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology, 85*(6), 869-879. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.85.6.869

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology, 64*(5), 502-508. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.64.5.502

Jiao, H., Liu, J., Haynie, K., Woo, A., & Gorham, J. (2012). Comparison between dichotomous and polytomous scoring of innovative items in a large-scale computerized adaptive test. *Educational and Psychological Measurement, 72*(3), 493-509. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1177/0013164411422903

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2003). The core self-evaluations scale: Development of a measure. *Personnel Psychology, 56*(2), 303-331. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/j.1744-6570.2003.tb00152.x

Judge, T.A., Locke, E.A., & Durham, C.C. (1997). The dispositional causes of job satisfaction: A core evaluations approach. *Research in Organizational Behavior, 19,* 151-188.

Kalton, G., Roberts, J., & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *The Statistician, 29,* 65–78.

Katz, D., & Kahn, R. L. (1966). *The social psychology of organizations* Wiley, Oxford. Retrieved from

http://search.proquest.com.proxy2.library.illinois.edu/docview/615467105?accountid=14553

Kawamoto, T., Ura, M., & Hiraki, K. (2017). Curious people are less affected by social rejection. *Personality and Individual Differences, 105*, 264-267. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1016/j.paid.2016.10.006

Kosinski, M. (2009). Application of the dominance and ideal point IRT models to the extraversion scale from the IPIP Big Five Personality Questionnaire. (Mphil Dissertation) Cambridge University.

Kramer, U., de Roten, Y., & Drapeau, M. (2011). Training effects with the observer-rated cognitive errors and the coping action patterns scales. *Swiss Journal of Psychology / Schweizerische Zeitschrift Für Psychologie / Revue Suisse De Psychologie, 70*(1), 41-46. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1024/1421-0185/a000037

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business Psychology, 22,* 251–259.

LaPalme, M., Tay, L., & Wang, W. (2017). A within-person examination of the ideal-point response process. *Psychological Assessment,* doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/pas0000499

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60*(5), 550-555. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.60.5.550

Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Measurement Series No. 84–4). Arlington, VA: Office of Naval Re- search, Personnel and Training Research Programs.

Levine, M. V., & Williams, B. A. (1991, May). *An overview and evaluation of nonparametric IRF estimation strategies.* Paper presented at the Office of Naval Research Contractors' Meeting on Model-Based Measurement, Princeton, NJ.

Levine, M. V., & Williams, B. A. (1993). *Nonparametric models for polychotomously scored item responses: Analysis and integration.* Unpublished manuscript.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22,* 1–55.

Ling, Y., Zhang, M., Locke, K. D., Li, G., & Li, Z. (2016). Examining the process of responding to circumplex scales of interpersonal values items: Should ideal point scoring methods be considered? *Journal of Personality Assessment, 98*(3), 310-318. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1080/00223891.2015.1077852

Locke, K. D. (2000). Circumplex scales of interpersonal values: Reliability, validity, and applicability to interpersonal problems and personality disorders. *Journal of Personality Assessment, 75*(2), 249-267. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1207/S15327752JPA7502_6

Locke, E. A., & Hulin, C. L. (1962). A review and evaluation of the validity studies of activity vector analysis. *Personnel Psychology, 15*(2), 25-42. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/j.1744-6570.1962.tb01844.x

Locke, E.A., McClear, K., Knight, D. (1996). Self-esteem and work. *International Review of Industrial/Organizational Psychology, 11,* 1-32.

Luteijn, F., van Dijk, H., & Barelds, D.P.H. (2005). *NPV-J: Junior Nederlandse Persoonlijkheidsvragenlijst. Herziene handleiding 2005* [NPV-J: Dutch Personality Questionnaire-Junior: Professional manual (revised)]. Amsterdam: Harcourt Assessments B.V.

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences*, *19*, 451-458. doi:10.1016/j.lindif.2009.03.007

McCrae, R. R., & Costa, P. T. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology, 56*(4), 586-595. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0022-3514.56.4.586

McManus, M. A., & Kelly, M. L. (1999). Personality measures and biodata: Evidence regarding their incremental predictive value in the life insurance industry. *Personnel Psychology, 52*(1), 137-148. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619420262?accountid=14553

Miller, R. L., Griffin, M. A., & Hart, P. M. (1999). Personality and organizational health: The role of conscientiousness. *Work & Stress, 13*(1), 7-19. Retrieved from

http://search.proquest.com.proxy2.library.illinois.edu/docview/619409978?accountid=14553

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology,79*(4), 475-480. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.79.4.475

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1177/014662169201600206

Mussel, P. (2013). Intellect: A theoretical framework for personality traits related to intellectual achievements. *Journal of Personality and Social Psychology, 104,* 885–906. http://dx.doi.org/10.1037/a0031918

Mussel, P., Spengler, M., Litman, J. A., & Schuler, H. (2012). Development and validation of the German work-related curiosity scale. *European Journal of Psychological Assessment, 28*(2), 109-117. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1027/1015-5759/a000098

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*(4), 679-703. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.78.4.679

Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology, 48*(4), 775-802. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/618918996?accountid=14553

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,* 481–484.

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

Paulhus, D. L. (1999). *Manual for the Paulhus Deception Scales: BIDR Version 7*. Toronto, Ontario, Canada: Multi-Health Systems.

Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of behavior. *Journal of Personality and Social Psychology, 74,* 538–556.

Paunonen, S. V., & Ashton, M. C. (2001). Big Five predictors of academic achievement. *Journal of Research in Personality, 35,* 78–90.

Peabody, D., & De Raad, B. (2002). The substantive nature of psycholexical personality factors: A comparison across languages. *Journal of Personality and Social Psychology*, *83*, 983-997.

Perugini, M., & Gallucci, M. (1997). A hierarchical faceted model of the Big Five. *European Journal of Personality*, *11*, 279-301.

Presser, S., & Schuman, H. (1980). The measurement of a middle position in attitude surveys. *Public Opinion Quarterly, 44,* 70–85.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/a0014996

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69*(4), 581-588. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.69.4.581

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-230. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/63718816?accountid=14553

Reio, T. G., Jr., & Callahan, J. L. (2004). Affect, curiosity, and socialization-related learning: A path analysis of antecedents to job performance. *Journal of Business and Psychology, 19*(1), 3-22. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1023/B:JOBU.0000040269.727

Roberts, B. W., Bogg, T., Walton, K. E., Chernyshenko, O. S., & Stark, S. E. (2004). A lexical investigation of the lower-order structure of conscientiousness. *Journal of Research in Personality*, *38*, 164-178. doi:10.1016/S0092- 6566(03)00065-5

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology, 58*(1), 103-139. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/j.1744-6570.2005.00301.x

Roberts, B. W., Jackson, J. J., Fayard, J. V., Edmonds, G., & Meints, J. (2009). Conscientiousness. In M. R. Leary, & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior; handbook of individual differences in social behavior* (pp. 369-381, Chapter xv, 624 Pages) Guilford Press, New York, NY. Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/622105037?accountid=14553

Roberts, B. W., Lejuez, C., Krueger, R. F., Richards, J. M., & Hill, P. L. (2014). What is conscientiousness and how can it be assessed? *Developmental Psychology, 50*(5), 1315-1330. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/a0031109

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*(2), 211-233. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619419377?accountid=14553

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. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1177/01466216000241001

Roberts, J. S., & Shim, H. S. (2008). GGUM2004 Technical Reference Manual (v1.1). Atlanta: Georgia Polytechnic University.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80* (1, Whole No. 609)

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72*(2), 282-307. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1207/S15327752JP720212

Sackett, P. R., Burris, L. R., & Callahan, C. (1989). Integrity testing for personnel selection: An update. *Personnel Psychology, 42*(3), 491-529. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/j.1744-6570.1989.tb00666.x

Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology, 82*(1), 30-43. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.82.1.30

Salgado, J. F., Moscoso, S., & Berges, A. (2013). Conscientiousness, its facets, and the prediction of job performance ratings: Evidence against the narrow measures. *International Journal of Selection and Assessment, 21*(1), 74-84. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/ijsa.12018

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4), 100. Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/615782233?accountid=14553

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar big-five markers. *Journal of Personality Assessment, 63*(3), 506-516. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1207/s15327752jpa6303_8

Saucier, G., & Ostendorf, F. (1999). Hierarchical subcomponents of the Big Five personality factors: A cross-language replication. *Journal of Personality and Social Psychology*, *76*, 613-627.

Schmit, M. J., & Ryan, A. M. (1997, April). *Specificity of item content in personality tests: An IRT analysis.* Paper presented at the 12th Annual Society for Industrial and Organizational Psychology Conference, St. Louis, MO.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76*(1), 27-33. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.2307/329895

Spector, P. E., Bauer, J. A., & Fox, S. (2010). Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? *Journal of Applied Psychology, 95*(4), 781-790. doi: http://dx.doi.org/10.1037/a0019477

Stark, S. (2007). MODFIT: Plot theoretical item response functions and examine the fit of dichotomous or polytomous IRT models to response data. Champaign, IL.

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. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.91.1.25

Stewart, G. L. (1999). Trait bandwidth and stages of job performance: Assessing differential effects for conscientiousness and its subtraits. *Journal of Applied Psychology*, 84, 959– 968.

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model–data fit of ideal point and dominance models. *Applied Psychological Measurement, 35*(4), 280-295. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1177/0146621610390674

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*(4), 703-742. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/j.1744-6570.1991.tb00696.x

Thissen, D., Chen, W.-H., & Bock, R. D. (2003). MULTILOG 7 for Windows: Multiple-category item analysis and test scoring using item response theory [Computer software]. Skokie, IL: Scientific Software International.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273-286. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/h0070288

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1086/214483

Trautwein, U., Ludtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology, 98,* 438–456.

Tucker, J. S., Kressin, N. R., Spiro, A., & Ruscio, J. (1998). Intrapersonal characteristics and the timing of divorce: A prospective investigation. *Journal of Social and Personal Relationships, 15*(2), 211-225. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1177/0265407598152005

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76*(1), 31-48. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/h0026750

Valbuena, N. (2004). *An empirical comparison of measurement equivalence methods based on confirmatory factor analysis (with mean and covariance structures analysis) and item response theory.* Available from PsycINFO. (620630932; 2004-99020-128). Retrieved from

http://search.proquest.com.proxy2.library.illinois.edu/docview/620630932?accountid=14553

van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts and what model to use instead. *Applied Psychological Measurement, 18,* 97-110.

Van Scotter, J. R., & Motowidlo, S. J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology, 81*(5), 525-531. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.81.5.525

Vickers, R. R., Jr., Conway, T. L., & Hervig, L. K. (1990). Demonstration of replicable dimensions of health behaviors. *Preventive Medicine, 19*(4), 377-401.

Vispoel, W. P., & Kim, H. Y. (2014). Psychometric properties for the balanced inventory of desirable responding: Dichotomous versus polytomous conventional and IRT scoring. *Psychological Assessment, 26*(3), 878-891. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/a0036430

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *2nd ed.; computerized adaptive testing: A primer (2nd ed.)* (2nd ed. ed., pp. 61-100, Chapter xxiii, 335 Pages) Lawrence Erlbaum Associates Publishers, Mahwah, NJ. Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/619459885?accountid=14553

Wang, W. (2013). A Bayesian Markov chain Monte Carlo approach to the generalized graded unfolding model estimation: The future of non-cognitive measurement. Available from PsycINFO. (1676371094; 2015-99080-541). Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/1676371094?accountid=14553

Watson, D. (2000). *Mood and temperament* Guilford Press, New York, NY. Retrieved from https://search-proquest-com.proxy2.library.illinois.edu/docview/619496002?accountid=14553

Webb, E. (1915). *Character and intelligence: An attempt at an exact study of character.* Cambridge, England: Cambridge University Press.

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a Dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment, 24*(1), 65-77. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1027/1015-5759.24.1.65

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. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/j.2044-8325.1994.tb00562.x

APPENDIX A: TRAINING MATERIALS FOR THE DICHOTOMOUS GROUP

“I like bananas, but I don’t like oranges.”

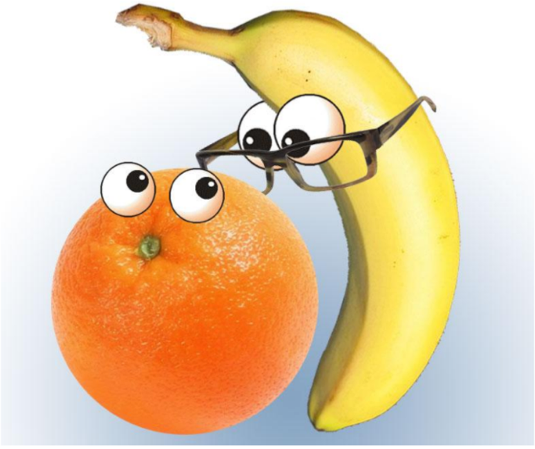
How would you respond??

It depends.

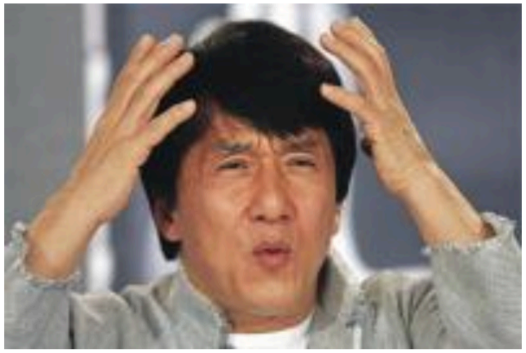
- If this statement fits you well, great, just go ahead and choose agree.

But,

(go to next page)



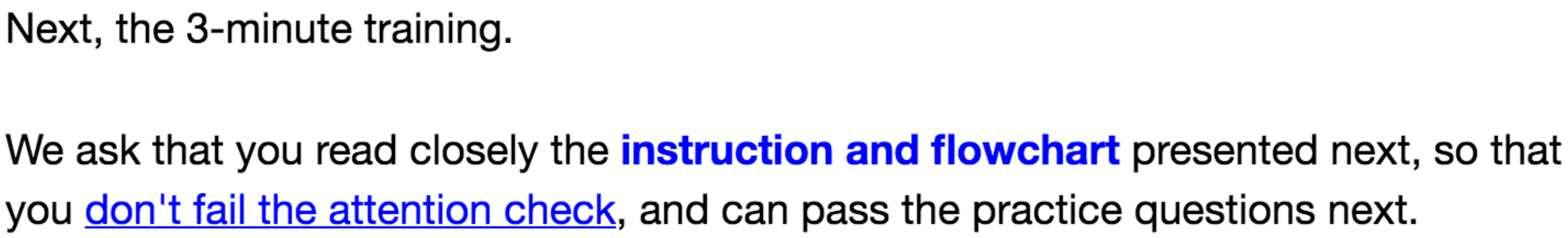
- If you love bananas **as well as** oranges, should you disagree or agree?
- If you **don't** like **either** bananas **or** oranges, how should you respond?



When asked to provide only **ONE** response for a statement they have **TWO** different attitudes toward, many participants have reported being confused.

In order to make you less confused, here's what we'll do:

(go to next page)

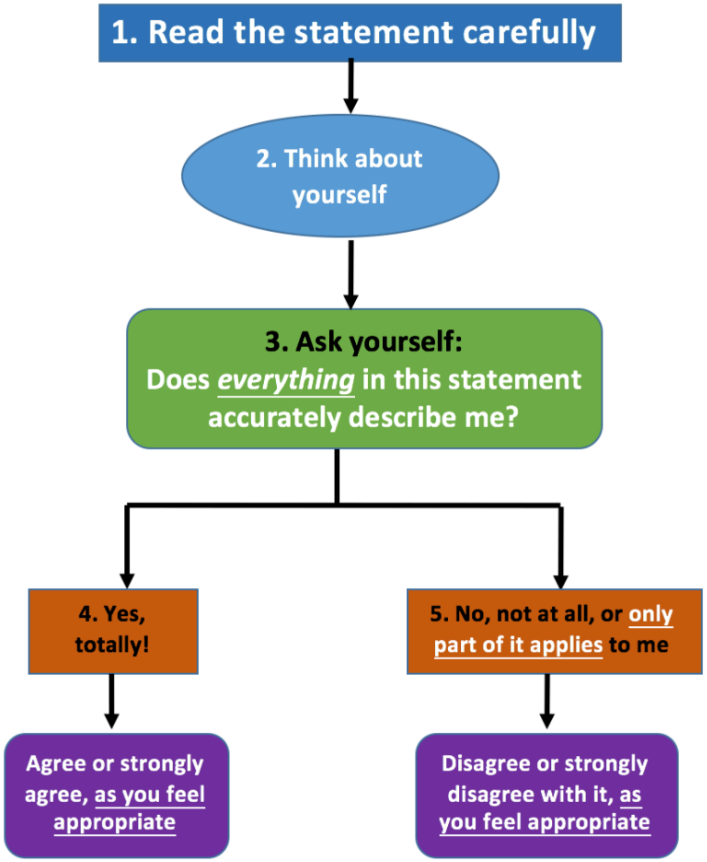


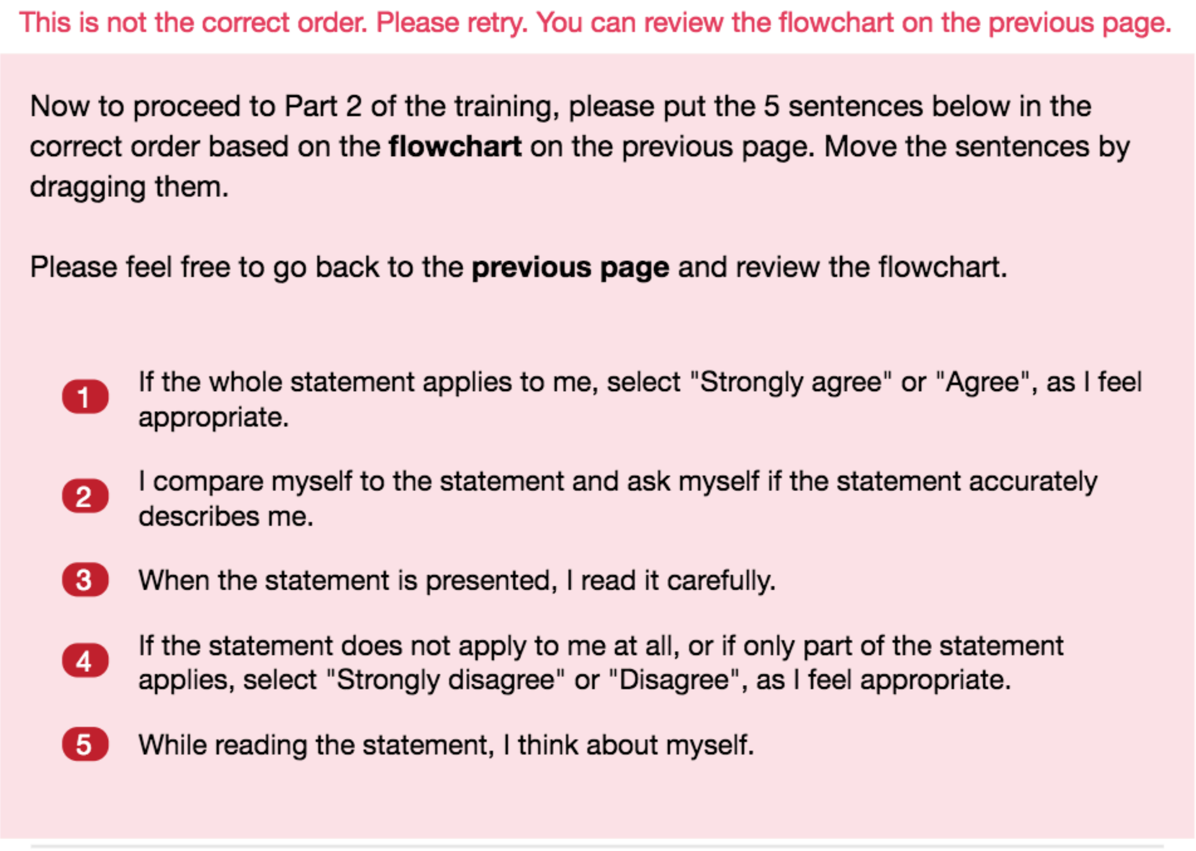
Next, the 3-minute training.

We ask that you read closely the **instruction and flowchart** presented next, so that you don't fail the attention check, and can pass the practice questions next.

1. **Strongly agree/Agree** only when **everything** in the statement accurately describes you.
2. If **nothing** or only **part** of it applies, **strongly disagree/disagree**!

3. As to what each of the response categories mean, please refer to the information below:
   "Strongly Disagree" = you are in total disagreement with the statement
   "Disagree" = you disagree with the statement for the most part
   "Agree" = you agree for the most part
   "Strongly Agree" = you are in total agreement with the statement


1. Read the statement carefully
2. Think about yourself
3. Ask yourself: Does *everything* in this statement accurately describe me?
4. Yes, totally!
5. No, not at all, or only part of it applies to me
Agree or strongly agree, as you feel appropriate
Disagree or strongly disagree with it, as you feel appropriate



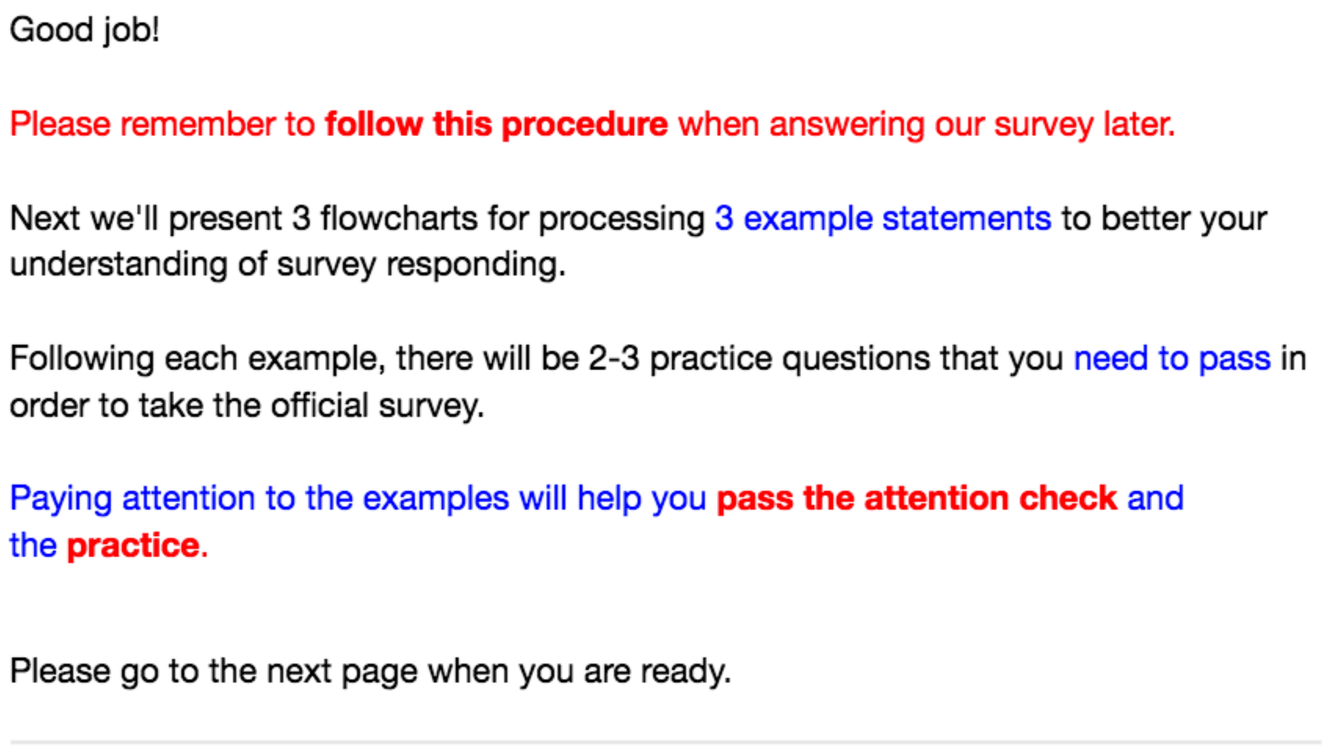
This is not the correct order. Please retry. You can review the flowchart on the previous page.

Now to proceed to Part 2 of the training, please put the 5 sentences below in the correct order based on the **flowchart** on the previous page. Move the sentences by dragging them.

Please feel free to go back to the **previous page** and review the flowchart.

1. If the whole statement applies to me, select "Strongly agree" or "Agree", as I feel appropriate.
2. I compare myself to the statement and ask myself if the statement accurately describes me.
3. When the statement is presented, I read it carefully.
4. If the statement does not apply to me at all, or if only part of the statement applies, select "Strongly disagree" or "Disagree", as I feel appropriate.
5. While reading the statement, I think about myself.

If participants got the order wrong, they would be instructed to review the flowchart and do it again. If they got it correct, they passed.



Good job!

Please remember to **follow this procedure** when answering our survey later.

Next we'll present 3 flowcharts for processing 3 example statements to better your understanding of survey responding.

Following each example, there will be 2-3 practice questions that you need to pass in order to take the official survey.

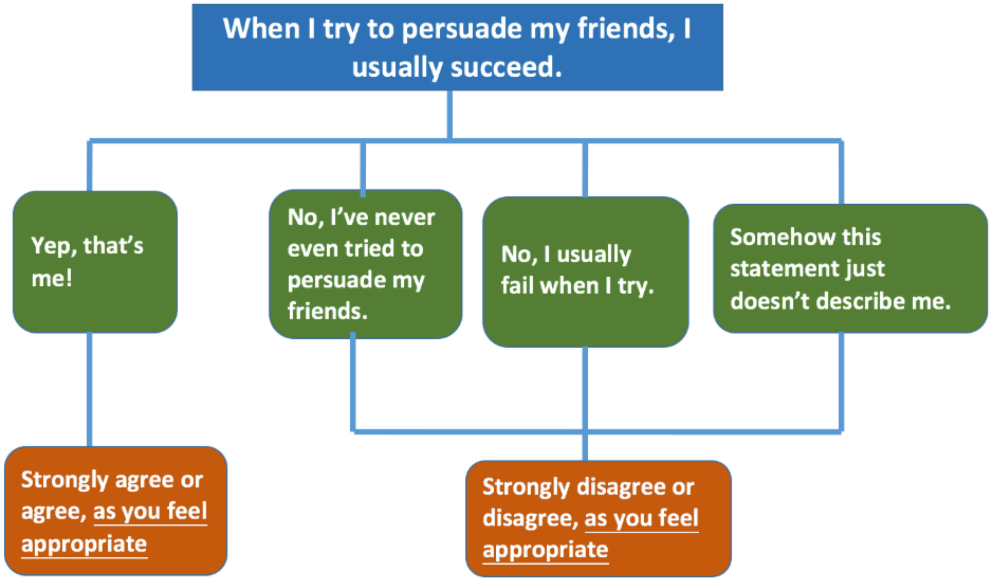
Paying attention to the examples will help you **pass the attention check** and the **practice**.

Please go to the next page when you are ready.

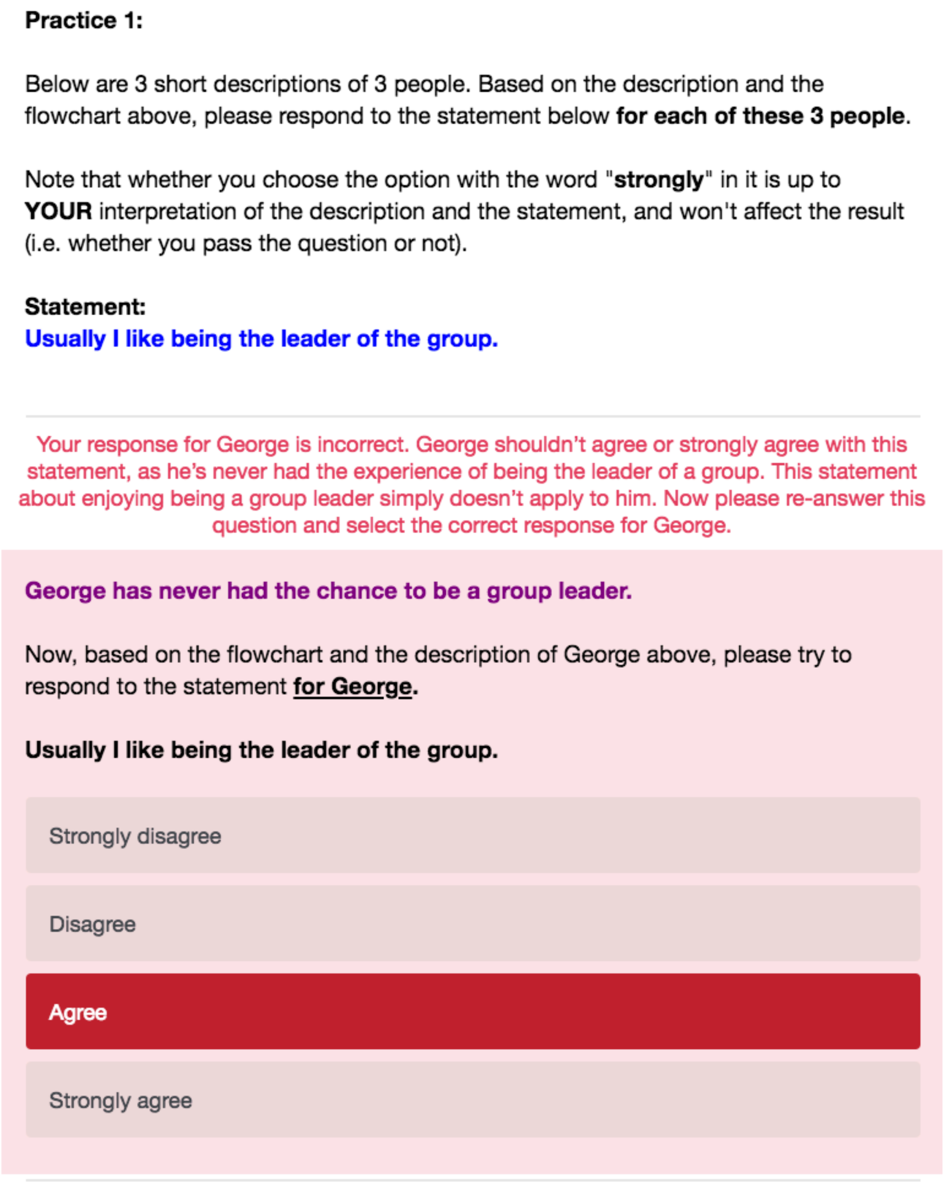
Example 1:

Statement: **When I try to persuade my friends, I usually succeed.**

The procedure you should follow:


When I try to persuade my friends, I usually succeed.
Yep, that's me!
No, I've never even tried to persuade my friends.
No, I usually fail when I try.
Somehow this statement just doesn't describe me.
Strongly agree or agree, as you feel appropriate
Strongly disagree or disagree, as you feel appropriate

1. **Strongly agree/agree** with the statement ***only*** when it's accurately describing you.
2. Otherwise, **strongly disagree/disagree** with it. It doesn't matter if it **overstates or understates** you.
3. You can also disagree with a statement for **other reasons**, as long as you feel that the statement is **not an accurate description** of you.



Practice 1:

Below are 3 short descriptions of 3 people. Based on the description and the flowchart above, please respond to the statement below **for each of these 3 people**.

Note that whether you choose the option with the word "**strongly**" in it is up to **YOUR** interpretation of the description and the statement, and won't affect the result (i.e. whether you pass the question or not).

Statement:
Usually I like being the leader of the group.

Your response for George is incorrect. George shouldn't agree or strongly agree with this statement, as he's never had the experience of being the leader of a group. This statement about enjoying being a group leader simply doesn't apply to him. Now please re-answer this question and select the correct response for George.

George has never had the chance to be a group leader.

Now, based on the flowchart and the description of George above, please try to respond to the statement **for George**.

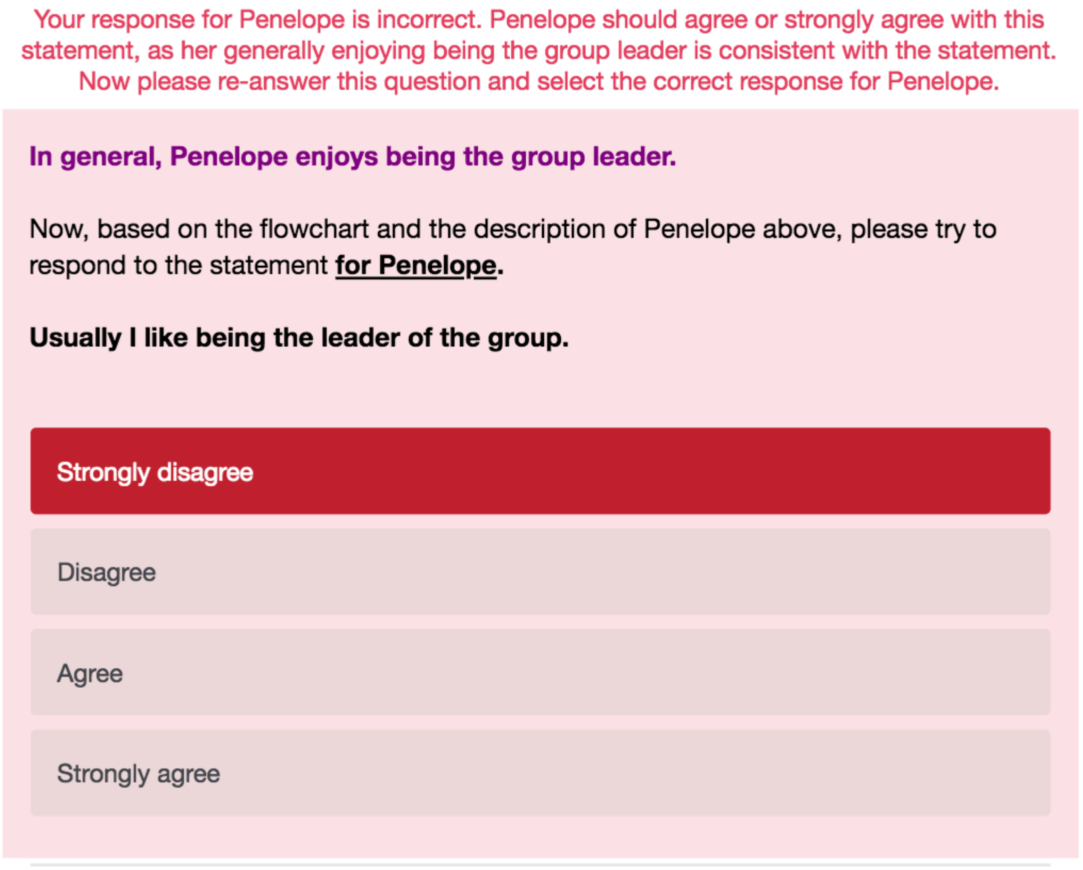
Usually I like being the leader of the group.

Strongly disagree

Disagree

[✓ Agree]

Strongly agree



Your response for Penelope is incorrect. Penelope should agree or strongly agree with this statement, as her generally enjoying being the group leader is consistent with the statement. Now please re-answer this question and select the correct response for Penelope.

In general, Penelope enjoys being the group leader.

Now, based on the flowchart and the description of Penelope above, please try to respond to the statement **for Penelope**.

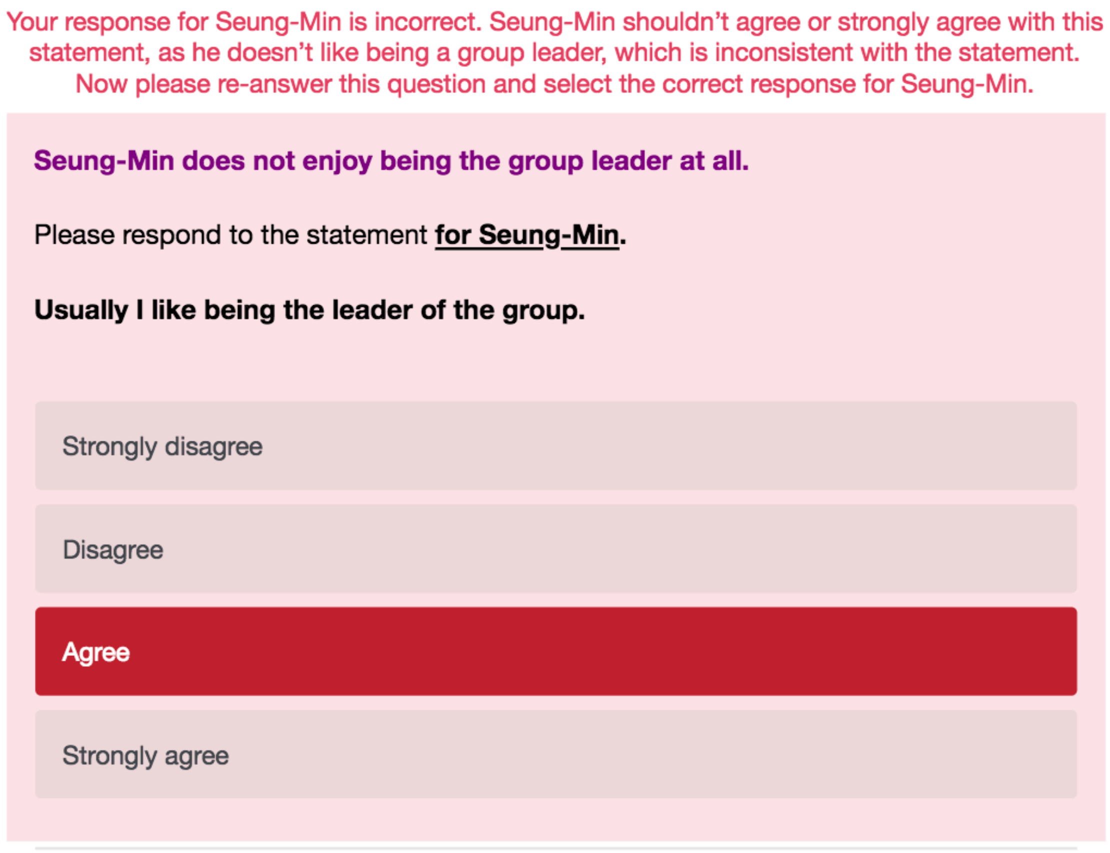
Usually I like being the leader of the group.

[✓ Strongly disagree]

Disagree

Agree

Strongly agree



Your response for Seung-Min is incorrect. Seung-Min shouldn't agree or strongly agree with this statement, as he doesn't like being a group leader, which is inconsistent with the statement. Now please re-answer this question and select the correct response for Seung-Min.

Seung-Min does not enjoy being the group leader at all.

Please respond to the statement **for Seung-Min**.

Usually I like being the leader of the group.

Strongly disagree

Disagree

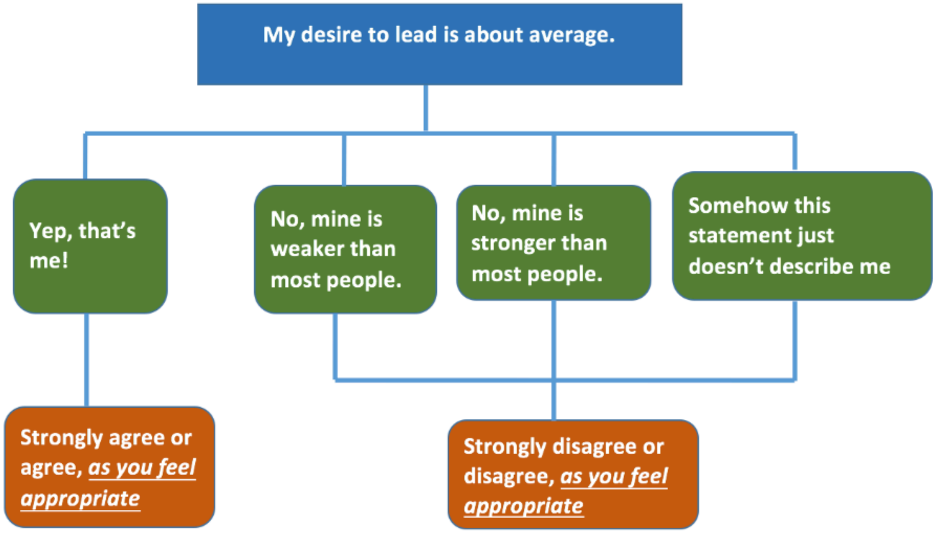
[✓ Agree]

Strongly agree

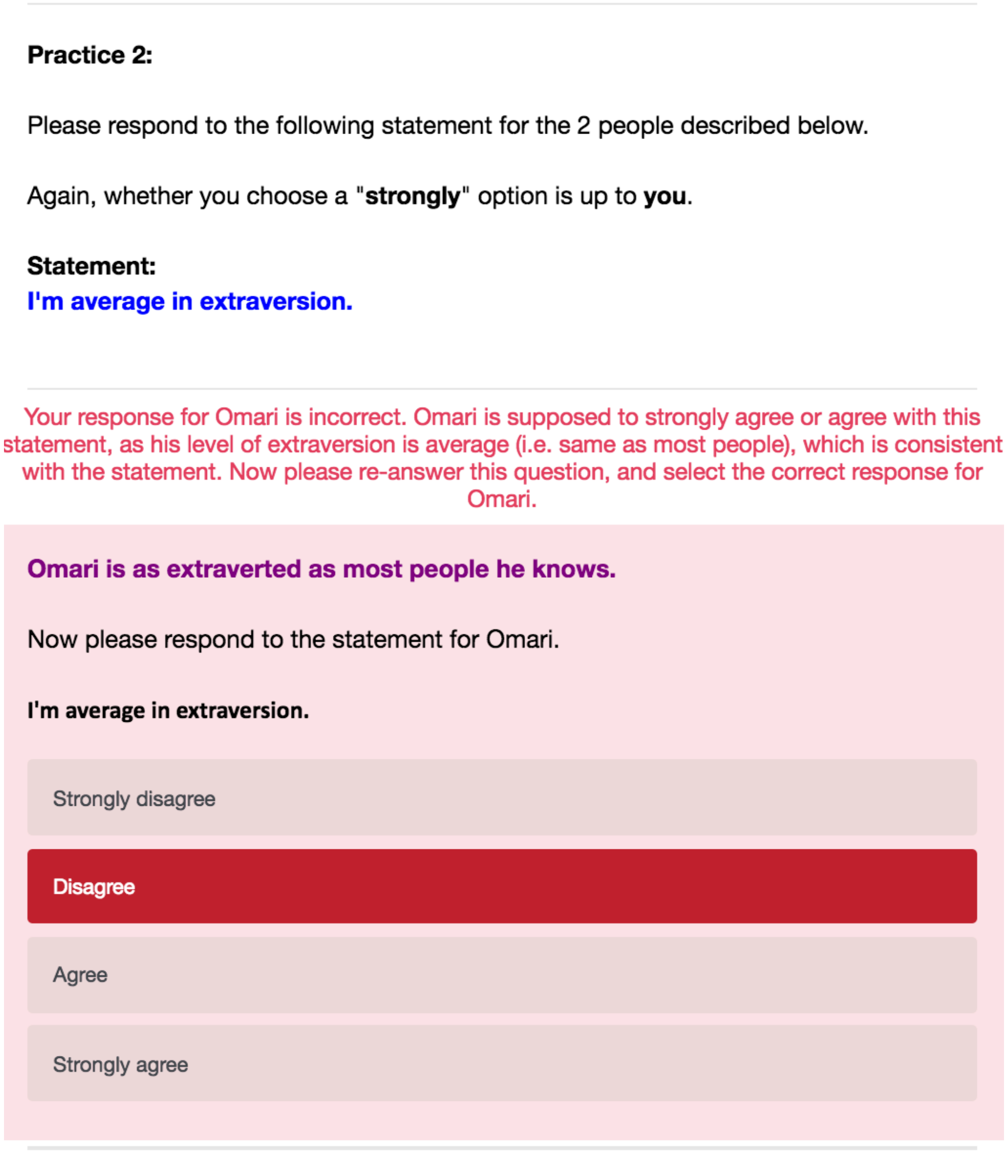
Example 2:

Statement: **My desire to lead a group is about average.**

The procedure you should follow:


My desire to lead is about average.
Yep, that's me!
No, mine is weaker than most people.
No, mine is stronger than most people.
Somehow this statement just doesn't describe me
Strongly agree or agree, *as you feel appropriate*
Strongly disagree or disagree, *as you feel appropriate*

1. **Strongly agree/agree** with the statement **only** when it's accurately describing you.
2. Otherwise, **strongly disagree/disagree** with it. It doesn't matter if it **overstates or understates** you.
3. You can also disagree with a statement for **other reasons**, as long as you feel that the statement is **not an accurate description** of you.



Practice 2:

Please respond to the following statement for the 2 people described below.

Again, whether you choose a "**strongly**" option is up to **you**.

Statement:
I'm average in extraversion.

Your response for Omari is incorrect. Omari is supposed to strongly agree or agree with this statement, as his level of extraversion is average (i.e. same as most people), which is consistent with the statement. Now please re-answer this question, and select the correct response for Omari.

Omari is as extraverted as most people he knows.

Now please respond to the statement for Omari.

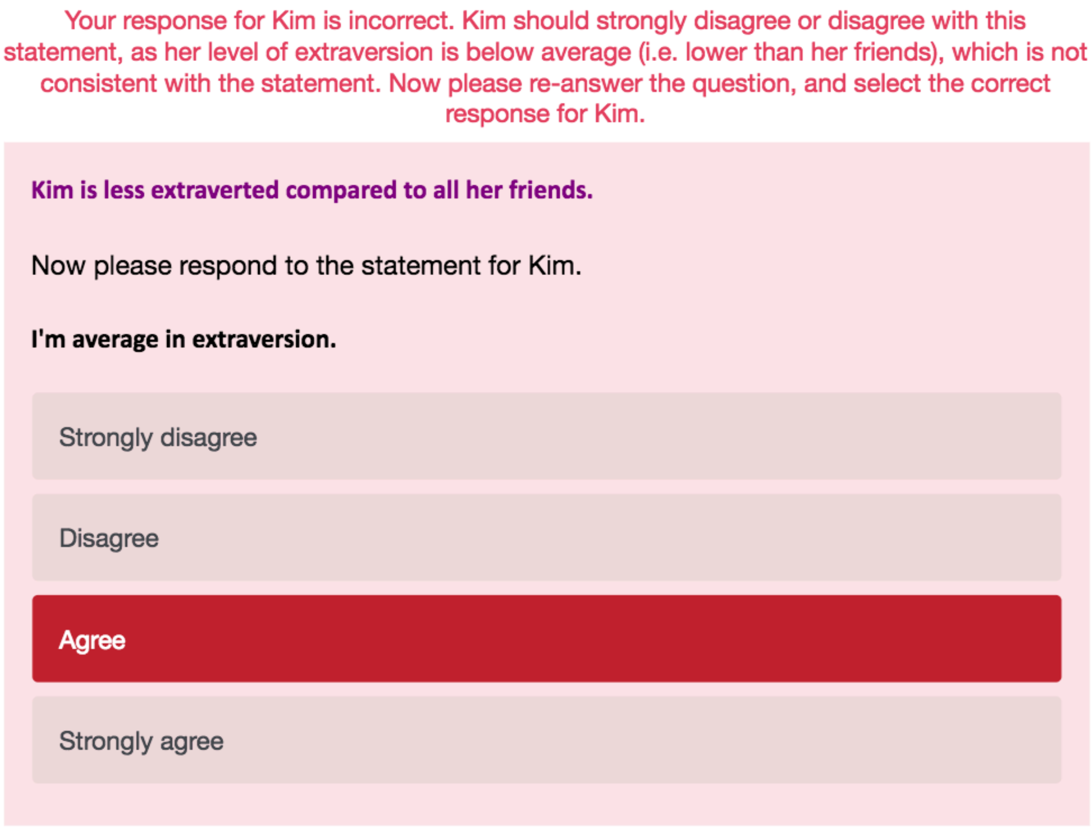
I'm average in extraversion.

Strongly disagree

[✓ Disagree]

Agree

Strongly agree



Your response for Kim is incorrect. Kim should strongly disagree or disagree with this statement, as her level of extraversion is below average (i.e. lower than her friends), which is not consistent with the statement. Now please re-answer the question, and select the correct response for Kim.

Kim is less extraverted compared to all her friends.

Now please respond to the statement for Kim.

I'm average in extraversion.

Strongly disagree

Disagree

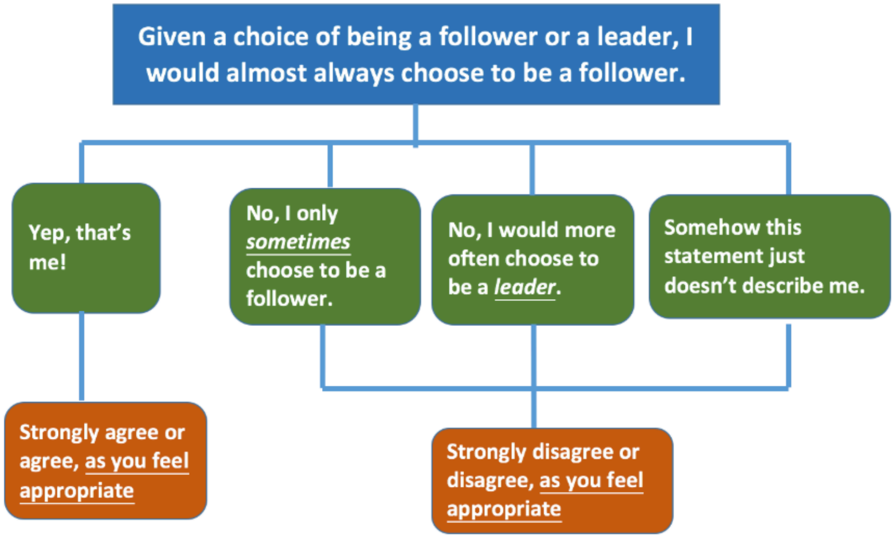
[✓ Agree]

Strongly agree

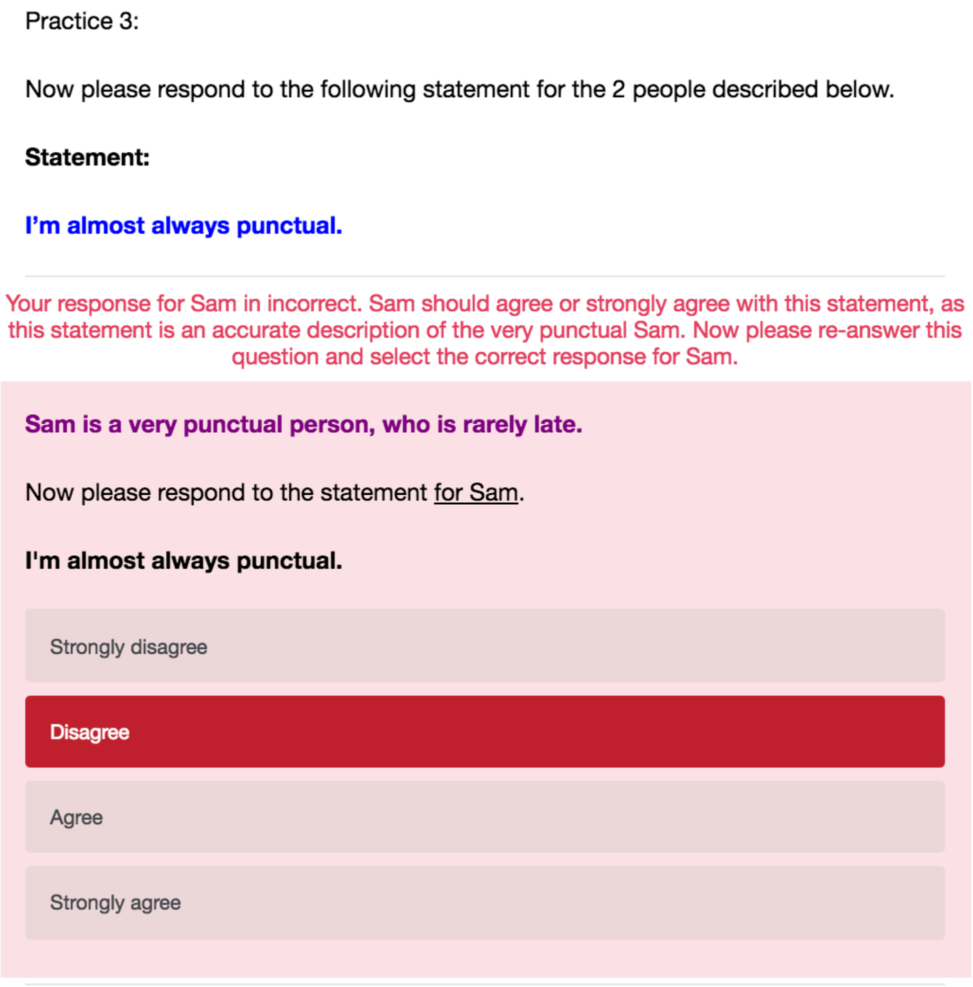
Example 3:

Statement: **Given a choice of being a follower or a leader, I would almost always choose to be a follower.**

The procedure you should follow:


Given a choice of being a follower or a leader, I would almost always choose to be a follower.
Yep, that's me!
No, I only *sometimes* choose to be a follower.
No, I would more often choose to be a *leader*.
Somehow this statement just doesn't describe me.
Strongly agree or agree, as you feel appropriate
Strongly disagree or disagree, as you feel appropriate

1. **Strongly agree/agree** with the statement **only** when it's accurately describing you.
2. Otherwise, **strongly disagree/disagree** with it. It doesn't matter if it **overstates or understates** you.
3. You can also disagree with a statement for **other reasons**, as long as you feel that the statement is **not an accurate description** of you.



Practice 3:

Now please respond to the following statement for the 2 people described below.

Statement:

I’m almost always punctual.

Your response for Sam in incorrect. Sam should agree or strongly agree with this statement, as this statement is an accurate description of the very punctual Sam. Now please re-answer this question and select the correct response for Sam.

Sam is a very punctual person, who is rarely late.

Now please respond to the statement for Sam.

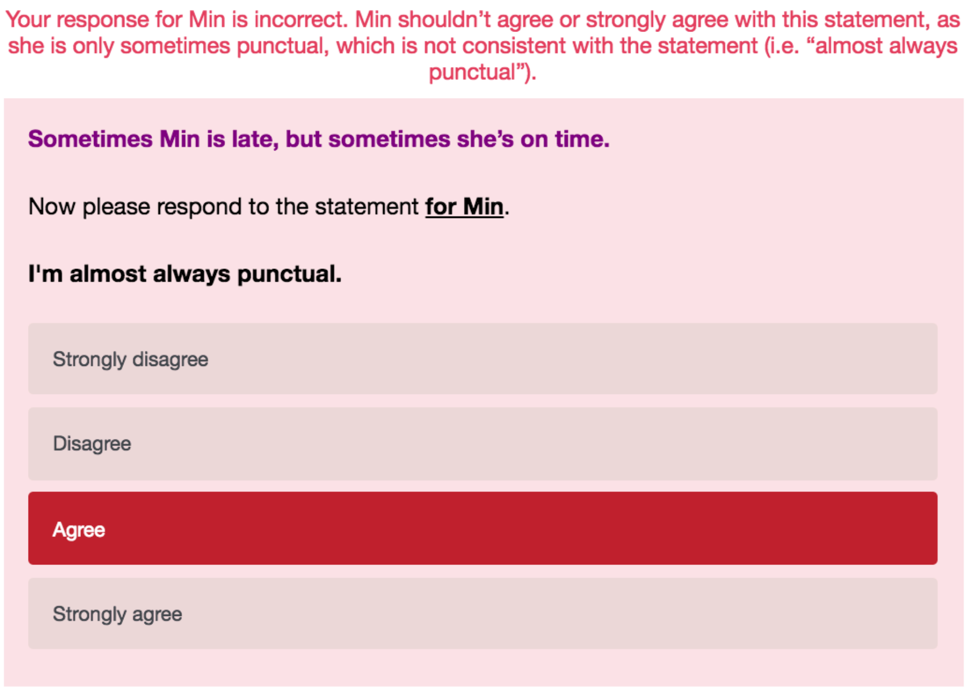
I'm almost always punctual.

Strongly disagree

[✓ Disagree]

Agree

Strongly agree



Your response for Min is incorrect. Min shouldn’t agree or strongly agree with this statement, as she is only sometimes punctual, which is not consistent with the statement (i.e. “almost always punctual”).

Sometimes Min is late, but sometimes she’s on time.

Now please respond to the statement **for Min**.

I'm almost always punctual.

Strongly disagree

Disagree

[✓ Agree]

Strongly agree

APPENDIX B: CRITERION MEASURES

Measure	ID	Content
SWLS	1	In most ways my life is close to my ideal
	2	The conditions of my life are excellent
	3	I am satisfied with my life
	4	So far I have gotten the important things I want in life
	5	If I could live my life over, I would change almost nothing
AP	1	How do/did you do in school?
CWB	1	Purposely wasted your employer's materials/supplies
	2	Complained about insignificant things at work
	3	Told people outside the job what a lousy place you work for
	4	Came to work late without permission
	5	Stayed home from work and said you were sick when you weren't
	6	Insulted someone about their job performance
	7	Made fun of someone's personal life
	8	Ignored someone at work
	9	Started an argument with someone at work
	10	Insulted or made fun of someone at work
HBCL_WME	1	I exercise to stay healthy.
	2	I gather information on things that affect my health by watching television and reading books, newspapers, and magazine articles.

	3	I see a doctor for regular checkups.
	4	I see a dentist for regular checkups.
	5	I discuss health with friends, neighbors, and relatives.
	6	I limit my intake of foods like coffee, sugar, fats, etc.
	7	I use dental floss regularly.
	8	I watch my weight.
	9	I take vitamins.
	10	I take health food supplements (e.g. protein additives, wheat germs, bran, lecithin).
HBCL_AC	1	I keep emergency numbers near the phone.
	2	I destroy old or unused medicines.
	3	I have a first aid kit in my home.
	4	I check the condition of electrical appliances, the cat, etc. to avoid accidents.
	5	I fix broken things around my home right away.
	6	I learn first aid techniques.
HBCL_TR	1	I cross busy streets in the middle of the block.
	2	I take more chances doing things than the average person.
	3	I speed while driving.
	4	I take chances when crossing the street.
	5	I carefully obey traffic rules so I won’t have accidents.
	6	I cross the street against the stop light.
	7	I engage in activities or hobbies where accidents are possible (e.g. motorcycle riding, skiing, using power tools, sky or skin diving, hang-gliding, etc.)

APPENDIX C: MEAN FOR THE IDEAL-POINT MEASURE ITEM RESPONSES

ID		G1		G3		G2		G4	
		N	Mean	N	Mean	N	Mean	N	Mean
CPS_Ord1	I can never find the things I want at home.	489	1.16	494	1.13	494	1.89	497	1.86
CPS_Ord2	I am an unorganized person.	490	1.27	493	1.22	494	1.94	496	1.92
CPS_Ord3	I try to keep track of my bills, but I'm not too accurate.	489	1.31	494	1.28	494	2.02	498	2.08
CPS_Ord4	Being in a clean room makes me feel uncomfortable.	490	1.07	494	1.08	495	1.53	496	1.5
CPS_Ord5	Organizing and arranging things is extremely fulfilling.	488	1.72	493	1.69	492	2.97	495	2.9
CPS_Ord6	It's hard for me to keep things in order.	490	1.3	493	1.28	494	2.08	497	2.08
CPS_Ord7	I can ignore a mess for a long time, but eventually I have to clean it up.	489	1.53	494	1.55	494	2.64	496	2.62
CPS_Ord8	I plan my time very carefully.	489	1.61	493	1.51	495	2.73	498	2.66
CPS_Ord9	I prefer not to plan ahead and instead take life as it comes.	488	1.28	494	1.26	494	2.12	497	2.03
CPS_Ord10	Every book on my bookshelf is in a specific order.	489	1.29	493	1.26	495	2.19	497	2.17
CPS_Ord11	I follow a strict daily schedule.	490	1.37	492	1.34	492	2.36	498	2.3
CPS_Ord12	I try to keep my room clean and tidy but I don't always have time to do so.	490	1.64	494	1.62	493	2.73	497	2.65
CPS_Ord13	When I have many things to do, I try to focus on the task with the highest priority first.	489	1.91	494	1.86	494	3.32	497	3.32
CPS_Ord14	Sometimes I wish that everyone was as organized as me.	490	1.51	493	1.49	494	2.63	496	2.59
CPS_Ord15	Being messy helps my creativity.	490	1.21	494	1.15	495	1.85	496	1.87
CPS_Ord16	Being clean helps me to focus.	489	1.8	494	1.8	495	3.17	497	3.1
CPS_Ord17	It bothers me a lot when my plans are disturbed.	490	1.68	494	1.66	493	2.85	497	2.91
CPS_Ord18	Organizing things is a waste of time.	490	1.08	493	1.05	494	1.64	497	1.6
CPS_Ord19	I am about average in regard to details.	488	1.48	494	1.46	495	2.47	498	2.37
CPS_Ord20	A little bit of disorganization is good for people.	489	1.52	494	1.48	495	2.42	497	2.38
Cao_Ord1	Occasionally I miss a deadline or two.	490	1.4	494	1.4	494	2.18	497	2.15

Cao_Ord2	Sometimes I do not put things in their proper place.	488	1.59	493	1.58	494	2.45	498	2.57
Cao_Ord3	Sometimes I can tolerate the messiness of my room.	490	1.6	492	1.6	495	2.61	497	2.66
Cao_Ord4	I spend time cleaning and organizing things when I am not busy.	490	1.67	493	1.65	494	2.78	497	2.79
Cao_Ord5	I deviate from my routines when needed.	489	1.88	493	1.84	493	3.01	497	3.03
Cao_Ord6	When my desk gets too messy, I will clean it up.	490	1.86	494	1.85	495	3.24	496	3.24
Cao_Ord7	I am about average in regard to details.	488	1.46	493	1.48	494	2.49	498	2.31
Cao_Ord8	My room neatness is about average.	490	1.59	492	1.61	495	2.67	496	2.55
Cao_Ord9	I consider myself as organized as most other people.	490	1.69	494	1.61	495	2.86	496	2.75
CPS_Cur1	I am open to new concepts but only if they are not hard to understand.	490	1.37	493	1.35	494	2.33	498	2.24
CPS_Cur2	I learn new things only when I have to.	490	1.18	493	1.18	494	1.96	498	1.83
CPS_Cur3	I am not really interested in new technology.	490	1.18	494	1.21	495	1.84	498	1.85
CPS_Cur4	I am usually intrigued by what I learn in classes.	490	1.88	494	1.8	494	3.18	498	3.18
CPS_Cur5	I only care about information that is relevant to me.	490	1.26	494	1.26	495	2.17	497	2.15
CPS_Cur6	I sometimes try new things just so I can learn more about them.	490	1.89	494	1.87	495	3.17	497	3.17
CPS_Cur7	I can be persuaded to try some new things, but most of the time I am reluctant to do so.	488	1.4	494	1.39	494	2.37	498	2.3
CPS_Cur8	I sometimes read non-fiction books to learn something new.	490	1.8	494	1.74	495	3.01	495	3.04
CPS_Cur9	I prefer to explore new concepts rather than apply them.	490	1.51	493	1.43	493	2.56	497	2.46
CPS_Cur10	I am interested in what is happening around the world.	487	1.92	493	1.86	495	3.22	498	3.21
CPS_Cur11	I am excited about new knowledge.	490	1.94	494	1.9	492	3.37	495	3.37
CPS_Cur12	I like to learn new things whenever I have time.	487	1.91	494	1.84	495	3.24	497	3.2
CPS_Cur13	I am as curious as anybody else I know.	490	1.81	492	1.64	489	3	497	2.85
CPS_Cur14	I am not curious about the things that I don't know.	489	1.17	494	1.11	494	1.79	495	1.71
CPS_Cur15	I would prefer a job where I don't have to learn anything new.	490	1.16	494	1.16	495	1.87	498	1.82
CPS_Cur16	I prefer to read fiction books rather than non-fiction.	490	1.51	494	1.5	494	2.59	495	2.57
CPS_Cur17	I am fascinated by science.	489	1.81	491	1.74	493	3.12	498	3.16

CPS_Cur18	I am not interested in learning new things.	490	1.08	494	1.08	494	1.61	498	1.56
CPS_Cur19	I like to experience new things, but find myself limited by my obligations.	489	1.67	494	1.62	495	2.68	497	2.73
CPS_Cur20	I try new restaurants only when other people recommend them.	490	1.33	494	1.28	494	2.29	498	2.18
Cao_Cur1	I like to experience new things, but seldom have time.	488	1.57	494	1.52	493	2.66	497	2.61
Cao_Cur2	I am not excited about new technology, but I become interested when others show me how to use it.	490	1.36	494	1.33	494	2.24	497	2.17
Cao_Cur3	At times I prefer to try new things rather than stick to old choices.	490	1.75	493	1.76	494	2.91	497	2.96
Cao_Cur4	Occasionally I find myself interested in information that I really don’t need.	490	1.85	494	1.85	493	3.13	498	3.12
Cao_Cur5	I do not mind trying new things when there are not many choices.	488	1.75	493	1.71	493	2.97	498	2.88
Cao_Cur6	I am about as curious as my friends.	489	1.74	493	1.58	492	2.89	496	2.73
Cao_Cur7	I am about average in curiosity about new knowledge.	490	1.54	494	1.47	495	2.51	498	2.32
Cao_Cur8	I have a moderate interest in learning new skills.	488	1.78	494	1.64	493	2.88	496	2.76
CPS_Ind1	I am competitive and play to win.	490	1.59	489	1.54	493	2.71	497	2.68
CPS_Ind2	I find it easy to stick to my plans.	490	1.79	494	1.7	494	2.97	498	2.95
CPS_Ind3	I am average at the things I do.	490	1.48	494	1.47	493	2.46	496	2.32
CPS_Ind4	I frequently make up believable excuses for not finishing my work.	490	1.14	493	1.17	494	1.81	497	1.78
CPS_Ind5	I finish my work on time but try not to work more than I have to.	489	1.42	494	1.44	494	2.44	497	2.45
CPS_Ind6	I work hard, but I know when it’s time to quit.	490	1.82	493	1.81	495	3.09	498	3.01
CPS_Ind7	I enjoy the process of doing things and don’t care much about the results.	490	1.23	494	1.2	495	2.03	498	1.94
CPS_Ind8	Being successful is more important than most other things in my life.	489	1.32	494	1.31	493	2.25	497	2.18
CPS_Ind9	I don’t care very much about the quality of my work.	490	1.04	494	1.06	495	1.51	497	1.46
CPS_Ind10	I hardly ever finish the tasks I start.	488	1.14	493	1.11	494	1.79	498	1.75
CPS_Ind11	I tend to do just what is expected of me when doing a job.	489	1.4	493	1.42	493	2.42	496	2.4

CPS_Ind12	I always want to be better than others in the things I do.	490	1.63	494	1.55	494	2.81	497	2.75
CPS_Ind13	There is too much to be done to waste time relaxing.	489	1.34	494	1.31	495	2.18	498	2.18
CPS_Ind14	When I set my mind on achieving a goal, I can always reach it.	489	1.72	491	1.58	494	3.02	497	2.9
CPS_Ind15	I always try to do my best work even when no one will know.	489	1.87	493	1.8	494	3.26	498	3.24
CPS_Ind16	If I am interested in something I don't mind working hard.	487	1.95	493	1.94	495	3.46	496	3.44
CPS_Ind17	To me, being moderately successful is enough.	489	1.63	493	1.63	494	2.67	497	2.68
CPS_Ind18	I don't really care about being successful.	489	1.23	493	1.21	494	1.96	496	1.95
CPS_Ind19	People should not sacrifice too much for work.	490	1.7	494	1.67	495	2.74	498	2.82
CPS_Ind20	I try to do the minimal amount of work possible to maintain my current status.	490	1.21	494	1.21	494	2	497	1.93
CPS_SC1	I try to consider all of the consequences of my actions, but sometimes can't help acting on impulse.	489	1.62	494	1.59	493	2.61	497	2.65
CPS_SC2	I have often missed important meetings because I forgot them.	490	1.12	494	1.11	494	1.65	497	1.6
CPS_SC3	I usually control my impulses.	490	1.81	494	1.82	494	3.04	498	3.03
CPS_SC4	It is hard to distract me when I am focused on a task.	489	1.67	491	1.61	495	2.76	498	2.77
CPS_SC5	Keeping a careful record of things is not my strength.	489	1.34	491	1.36	495	2.19	497	2.24
CPS_SC6	An impulsive decision isn't always bad.	490	1.81	494	1.77	493	2.87	498	2.88
CPS_SC7	I always think twice before saying something.	488	1.56	493	1.44	495	2.71	497	2.64
CPS_SC8	I am usually cautious.	488	1.85	494	1.85	495	3.14	496	3.11
CPS_SC9	I often make careless mistakes.	488	1.24	494	1.21	495	2.02	498	2.01
CPS_SC10	I can keep my concentration only on short tasks.	490	1.22	494	1.25	495	2.06	497	1.99
CPS_SC11	I don't think that being impulsive is a fault.	490	1.48	494	1.44	495	2.49	497	2.46
CPS_SC12	I am meticulous at most things I do.	490	1.68	494	1.63	494	2.84	497	2.86
CPS_SC13	I don't mind waiting for something better to come along.	489	1.78	493	1.72	493	2.89	498	2.86
CPS_SC14	My mind wanders a lot when I'm working on something.	490	1.45	494	1.45	495	2.44	498	2.41
CPS_SC15	I don't usually think before I talk.	489	1.17	493	1.19	495	1.91	498	1.87
CPS_SC16	I am more careful in places I am not familiar with.	490	1.91	494	1.92	493	3.33	497	3.3

CPS_SC17	I always have a detailed plan for my daily activities.	490	1.44	494	1.39	494	2.47	498	2.45
CPS_SC18	I believe people can never be too careful.	489	1.69	493	1.64	494	2.92	495	2.88
CPS_SC19	I am only careful on tasks that are important to me.	490	1.25	492	1.24	495	2.1	498	2.12
CPS_SC20	I make plans if I have enough time.	487	1.84	494	1.74	495	3.04	495	2.95

Note: G1: participants were not trained and the dichotomous response scale was used; G2: participants were not trained and a polytomous response scale was used; G3: participants were trained and a dichotomous response scale was used; G4: participants were trained and a polytomous response scale was used. CPS_Ord: the Order scale of the CPS; Cao_Ord: the intermediate items measuring order written by Cao and colleagues (2015); CPS_Cur: the Curiosity scale of the CPS; Cao_Cur: the intermediate items measuring curiosity written by Cao and colleagues (2015); CPS_Ind: the Industriousness scale of the CPS; CPS_SC: the Self-control scale of the CPS; N: the sample size based on which the mean was computed; Mean: the average of the responses.