2017-07-01

An Examination of the Psychometric Properties of the Trauma Inventory for Partners of Sex Addicts (TIPSA)

Steven Scott Stokes
Brigham Young University

An Examination of the Psychometric Properties of the Trauma Inventory for Partners of Sex Addicts (TIPSA)

Steven Scott Stokes

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R. Sudweeks, Chair
Lane Fischer
Joseph A. Olsen
Melissa A. Heath
Aaron P. Jackson

Educational Inquiry, Measurement, and Evaluation

Brigham Young University

ABSTRACT

An Examination of the Psychometric Properties of the Trauma Inventory for Partners of Sex Addicts (TIPSA)

Steven Scott Stokes
Educational Inquiry, Measurement, and Evaluation, BYU
Doctor of Philosophy

This study examined the psychometric properties of the Trauma Inventory for Partners of Sex Addicts (TIPSA). Using the Nominal Response Model (NRM), I examined several aspects of item and option functioning including discrimination, empirical category ordering, and information. Category Boundary Discrimination (CBD) parameters were calculated to determine the extent to which respondents distinguished between adjacent categories. Indistinguishable categories were collapsed through recoding. Empirically disordered response categories were also collapsed through recoding. Findings revealed that recoding solved some technical functioning issues in some items, and also revealed items (and perhaps option anchors) that were probably poorly conceived initially. In addition, nuisance or error variance was reduced only marginally by recoding, and the relative standing of respondents on the trait continuum remained largely unchanged. Items in need of modification or removal were identified, and issues of content validity were discussed.

Keywords: Nominal Response Model (NRM), Item Response Theory (IRT), trauma, sex addiction, Category Boundary Discrimination (CBD)

ACKNOWLEDGEMENTS

I would like to thank Dr. Richard Sudweeks, Dr. Lane Fischer, and all the members of my committee for taking the time to help me accomplish this project. I could not have done it without them.

I would also like to thank Dr. Sara Moulton for being my mentor and friend, and for being generous with her time and resources. I could not have done it without her either. Bob Bodily also helped me through several technical issues.

I am grateful to the faculty, staff, and students of Brigham Young University who inspire me. In particular, the EIME and IP&T departments have been very supportive.

My wife and my family have also supported me a great deal. Thank you very much.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

Addiction is a serious social problem that is estimated to impact approximately 50% of the adult population in the United States (Sussman, Lisha, & Griffiths, 2011). Although the term *addiction* is "often used without an attempt to define it," it is generally considered to be a pattern of behavior that is meant to produce pleasure or provide escape from internal strife, that an individual habitually fails to control because of obsession, and continues to employ despite significant negative consequences (Carnes, Murray, & Charpentier, 2004; Goodman, 1990, p. 1403; Kafka, 2010; Kraus, Voon, & Potenza, 2016).

For many years, addiction has been viewed and discussed almost exclusively in terms "substance-related disorders" (American Psychiatric Association, 2015; Kraus et al., 2016) and the spouses of addicts, and sometimes other close friends or family members, have been diagnosed and treated under the codependency paradigm (Minwalla, 2011; Stafford, 2001; Steffens & Rennie, 2006). This paradigm stems from the idea that individuals closest to the addict often modify or adapt their own behaviors in ways that facilitate or enable the addict to continue his or her substance use and abuse (Cermak, 1991). This paradigm assumes that spouses or partners carry their own "addictive or obsessive relationship with [the] addict," making them at least partly to blame for the addicts' prolonged negative behaviors (Steffens & Rennie, 2006, p. 261). Thus, those who espouse the codependency paradigm often refer to addiction as a "family disease" (Steffens & Rennie, 2006, p. 261).

Because addiction has traditionally been viewed in terms of substance abuse, so-called *behavioral addictions* have not technically been considered mental disorders (American Psychiatric Association, 2015; Kraus et al., 2016). In fact, Gambling Disorder is the only behavioral addiction that is formally recognized by the American Psychiatric Association (2015), and it was not included in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5) until the release of the fifth edition (2015). Other behavioral addictions such as Internet gaming disorder, exercise addiction, and shopping addiction have been described and proposed for inclusion in the DSM, but Gambling Disorder currently remains the only behavioral addiction formally recognized as such. According to the American Psychiatric Association (2015, p. 481), "at this time there is insufficient peer-reviewed evidence to establish the diagnostic criteria and course descriptions needed to identify these [compulsive] behaviors as mental disorders" (Kraus et al., 2016). Interestingly, there is considerable evidence that behavioral addictions and substance-related disorders often co-occur in ways that are negatively reinforcing (Carnes et al., 2004; Fearing, 2002; Freimuth et al., 2008; Schneider & Irons, 2001). Whether or not various behavioral addictions, including sex addiction, should be included in the DSM in an important question that deserves attention in future research (Ferree, 2001; Kafka, 2010; Kraus et al., 2016).

Despite the fact that sex addiction is not formally considered to be a mental disorder, the partners and spouses of sex addicts are often diagnosed and treated in much the same way as the partners of substance addicts—as codependent and coaddicted. More recently, however, mental health professionals have recognized that the discovery or revelation of sex addiction in a relationship can be extremely distressing to the point of inducing trauma, or Post Traumatic Stress Disorder (PTSD) symptoms (Skinner, 2015; Steffens & Rennie, 2006). The Trauma Inventory for Partners of Sex Addicts (TIPSA) is a groundbreaking scale designed to identify and measure the trauma experienced by the partners of sex addicts (Skinner, 2015). The precision of psychometric parameter estimates, discrimination, and ordering of TIPSA items and options, however, have not yet been sufficiently examined.

Purpose and Rationale

The purpose of this study was twofold. The first purpose was to estimate and examine important psychometric properties of the TIPSA including response option ordering and category boundary discrimination (CBD) parameters using the nominal response model (NRM) of item response theory (IRT). The second purpose was to make recommendations to the scale developer regarding which items and/or options are in need of improvement based on the findings.

As sex addictions and compulsions become increasingly recognized by counseling and therapeutic professionals as problematic, there is a need for research on the appropriate diagnosis and treatment methods not only for the addicts themselves, but also for their partners and spouses (Bancroft, 2008; Kraus et al., 2016). If the partners of sex addicts do, in fact, carry their own addictive and compulsive tendencies, then they ought to be diagnosed and treated accordingly. If, on the other hand, their customary negative behaviors are the result of trauma, then their diagnoses and treatment ought to reflect that fact. Therefore, this study has profound implications for the way in which mental health experts conceptualize and treat the problems faced by the partners of sex addicts.

The TIPSA represents an attempt to identify and measure trauma in the partners of sex addicts, but it's psychometric properties have not been sufficiently examined. This study will examine psychometric properties of TIPSA items using the NRM of IRT. Analysis will be performed using flexMIRT and R software. Only a handful of recent studies have used the NRM for the purposes of item and option analysis in scale development (Murray, Booth, & Molenaar, 2016; Preston, 2014a; Preston, Reise, Cai, & Hays, 2011). This research will not only improve the ability of mental health experts to identify trauma in the partners of sex addicts, but it will also contribute to the literature on the use of the NRM in psychometric scale development.

Research Questions

The following research questions and sub-questions were addressed in this study:

1. What is the discriminating power of the various TIPSA items and response options, and which items or options need to be revised for the purpose of improving their discriminating power?
2. To what extent does the scoring algorithm currently used to score responses to TIPSA items match the way the item options actually function, and how should the coding of the options be revised in order to minimize extraneous variance in the resulting scores?
   a. How does the discriminating power of the various options vary within individual items and across items?
   b. To what extent are boundaries between the options disordered?
   c. Which options, if any, should be collapsed and/or recoded?

CHAPTER 2: Review of Literature

Sex Addiction

Despite the rather myopic formal view of addiction as primarily a substance-related illness, many researchers and mental health professionals are willing to recognize the existence of a wide range of behavioral addictions, including sexual addiction (Bancroft, 2008; Carnes, 2013; Kafka, 2010; Kraus et al., 2016; MacLaren & Best, 2010; Schneider & Irons, 2001; Washton, 1989). MacLaren and Best (2010), for example, documented a wide variety of substance-related and compulsive behaviors in a sample of Canadian youth, including compulsive helping, work, relationships, shopping, food starving, food binging, prescription drugs, tobacco, alcohol, gambling, and sex. While many of these behaviors may be normal, enjoyable, necessary, and even healthy if done in moderation, the addict becomes obsessed to the point of losing control, and continues to engage in the activity despite severe negative personal and social consequences or disturbed levels of functioning.

In general terms, the negative consequences of addiction are well documented and often include (a) loss of self-control, (b) loss of pleasure or interest in hobbies and other activities that were once found enjoyable, (c) strained social relationships, (d) legal problems, (e) dangerous and / or criminal behavior, (f) physical harm or impairment, (g) financial problems, or (h) emotional trauma (Sussman et al., 2011). The negative consequences of compulsive sexual behavior are also well documented, often including intense shame, guilt, health risks (STDs), financial difficulties, work or educational role impairment, and damaged relationships (Bancroft, 2008; Black, 2009; Conner, 2013; Kafka, 2010). Thus, it is frequently necessary for addicts, and those closest to them, to seek professional help in order to overcome their compulsions, and repair the damage that has been done to self and others (Laaser, 2002; McCarthy, 2002).

Sex addiction, in particular, has the potential to be especially mentally, emotionally, and physically damaging for several reasons (Bancroft, 2008). While the term sex addiction is often thought of in terms of outlandish or extreme behaviors such as voyeurism, exploitation, or exhibitionism, Freimuth et al. (2008) point out that the majority of sex addictions occur within a range of activities that is generally socially acceptable (e.g., masturbation, pornography, anonymous sex). Indeed, sex addiction "is not defined by a specific frequency, quantity, intensity, or type of sexual behavior" (Freimuth et al., 2008, p. 142). Rather, sex addiction becomes evident when an individual "continually fantasizes about sex, pursues sexual stimuli, or acts out risky behaviors despite adverse social, physical, and psychological consequences" (Cooper, Delmonico, & Burg, 2000; Freimuth et al., 2008, p. 142).

First, sex addictions deserve special attention because they have often been found to coexist with other harmful addictions, compulsions, chemical dependencies, and mental or emotional problems (Carnes et al., 2004; Fearing, 2002; Freimuth et al., 2008; Schneider & Irons, 2001). For example, in an exhaustive 5-year study of 1,000 recovering sex addicts, less than 13% indicated that they had only one addiction (Carnes et al., 2004). The coexisting addictions, including drug and alcohol abuse in many cases, were found to interact with one another in ways that were cyclical and negatively reinforcing, making treatment and recovery even more challenging (Carnes et al., 2004). In a separate survey of 289 persons admitted to an inpatient facility for sex addicts, less than 17% reported sex as their only addiction (Carnes et al., 2004). Concurrent addictions included chemical dependency (42%), eating disorder (38%), compulsive working (28%), compulsive spending (26%), and compulsive gambling (5%) (Carnes et al., 2004). Sex addictions rarely occur in isolation, and often interact in reinforcing ways with other addictions and compulsive behaviors.

Second, sex addictions often put the addict, as well as his or her spouse or partner, at increased risk of serious mental, emotional, and physical health problems such as sexually transmitted diseases (STDs), abuse, rape, unwanted pregnancy, etc. (Freimuth et al., 2008; Hentsch-Cowles & Brock, 2013). The Centers for Disease Control and Prevention (CDC) has warned of epidemic levels of STDs in the United States, with more than 1,500,000 cases of chlamydia and about 800,000 cases of gonorrhea currently reported each year (Centers for Disease Control and Prevention (CDC), 2015, 2017a, 2017b). Individuals who engage in compulsive or unprotected sex with multiple partners are more likely to become infected (Centers for Disease Control and Prevention (CDC), 2015, 2017a, 2017b). In addition, sex addiction that culminates in abuse of others is not only illegal, but also leaves behind mental and emotional scars that are difficult to overcome.

Third, sexual relations are not only physical, but also intensely personal, vulnerable, trusting, and emotional experiences. Carnes (2013) describes human sexuality as simultaneously essential, powerful, and frightening. Healthy human sexual interactions are often built upon a significant foundation of trust, and represent an important and intimate part of healthy adult relationships. When that trust is breached by sex addiction, relationships are severely damaged, and the corresponding feelings of betrayal exacerbate the traumatic nature of the experience. According to betrayal trauma theory (Freyd, 1998), "trauma perpetuated by someone whom the victim trusts or on whom the victim depends (i.e., high-betrayal trauma; HBT) is more psychologically damaging than trauma perpetrated by someone with whom the victim is not close, or a noninterpersonal trauma" (Martin, Cromer, DePrince, & Freyd, 2013, p. 111). Martin and colleagues (2013) found, for example, that in a sample of college students, individuals that experienced High Betrayal Trauma (HBT) were more likely to experience deeper depression, dissociation, and PTSD-type symptoms. A growing body of empirical work shows that sex addiction has the potential to induce HBT experiences, and that this intense betrayal complicates the healing process. For example, 40% of sex addicts report losing a partner or spouse, and 70% report “severe marital or relationship problems” (Carnes, 2013, p. 15). In addition, feelings of betrayal are often discussed as being a fundamental to the trauma experienced by the partners of sex addicts (Hentsch-Cowles & Brock, 2013; Skinner, 2015; Steffens & Rennie, 2006). Indeed, “the spouse or partner of a sexual addict bears a great burden and experiences disruption in response to the out-of-control sexual behaviors of the addict” (Steffens & Rennie, 2006, p. 249).

Finally, sex addictions thrive on secrecy, and often remain unreported and untreated. The internet has made sexually explicit materials increasingly prevalent and accessible, so that pornography is now a multi-billion dollar industry worldwide with an estimated 700 – 800 million individual porn pages available on the internet (The Economist, 2015). Indeed, “the number of adult sites, and traffic to them, have exploded” in recent years (The Economist, 2015). A single popular porn site reported nearly 80 billion video viewings in 2014, and more than 18 billion site visits (The Economist, 2015). The proliferation of sexual material online has made it easier for sex addictions to remain concealed, unreported, and untreated. There is even some evidence that sex addictions like pornography impact micro and macroeconomic workplace performance, given that 25% of working adults admit to viewing porn while at work, and 70% of all online pornography access occurs between the normal working hours of 9 a.m. and 5 p.m. (Conner, 2013). One study found, for example, that 79% of sex addicts “talk of serious losses of job productivity, [and] 11 percent were demoted” as a result of their uncontrolled behavior (Carnes, 2013, p. 15). In sum, sex addictions come in many forms, and have a wide variety of negative consequences.

Because sex addictions have such a profound impact on the addict, as well as those closest to him or her, recovery often becomes a collaborative effort involving the addict and the partner (Hentsch-Cowles & Brock, 2013; Morgan, 1991). It is frequently necessary, however, to diagnose and treat the partners of sex addicts as distinct individuals, in order to address their unique needs and concerns. Theories on the proper diagnosis and treatment of partners of addicts can generally be divided into two broad camps: the addiction theory model (i.e., codependency and coaddiction) (Hentsch-Cowles & Brock, 2013; Schneider & Irons, 2001) and the trauma model (Skinner, 2015; Steffens & Rennie, 2006).

Addictions model. The addiction theory model originated from the view that addictions are primarily substance-related disorders, such as alcoholism or drug abuse. This model remains as an influential diagnosis and treatment paradigm because it recognizes sex addiction as a family disease, and emphasizes the role of the partner within that system (Black, 2009; Hentsch-Cowles & Brock, 2013; Steffens & Rennie, 2006). Those in favor of the addictions model rely heavily on the idea that the sex addict does not act out, or recover, in isolation, and that the coaddictive behavior of the spouse or partner is at least partly to blame (Hentsch-Cowles & Brock, 2013; Laaser, 2002; Schneider & Irons, 2001).

Steffens and Rennie (2006) summarized the addiction model well, stating that it

> centers upon the concept of the disease model, whereby individuals with illnesses or diseases such as addictions must engage in recovery behaviors to obtain remission and to prevent relapse. In addition, there is no cure. For WSAs [wives of sex addicts], this means that, although she is told that her husband's addictive or compulsive behaviors are not about her, she is also told that she carries her own disease that contributes to the continuation or deepening of the sexual addiction. Her attempts to "fix" the addict are

> therefore viewed as symptomatic of her own addictive illness, co-addiction. The co-addict in recovery demonstrates health by her ability to focus upon her own life and detach from the addict by reducing her obsession with her spouse's life and behaviors (pp. 261-262).

Thus, the addiction theory model treats and diagnoses the sex addict, and his or her partner, in much the same way as an alcoholic or chronic drug user (Laaser, 2002). The sex addict is treated for persisting in obsessive sexual behaviors despite severe negative consequences, and the co-sex addict partner is treated for obsessively persisting in negative behaviors such as

> raging, threatening, persistent anxiety, policing, snooping, attention/validation seeking, creating drama/chaos, self-blame, wokaholism, being obsessively busy, being a super-parent, rationalizing, manipulation, lying to friends and family about marriage/relationship, denial, perfectionism, self-harm, using the Internet, shopping, food, alcohol, other forms of substances to numb/act out, depression, sleeping extended hours, keeping up appearances, . . . being sexually available to the addict," etc. (Hentsch-Cowles & Brock, 2013, p. 329).

According to the addictions theory model, these and other similar behaviors are used as coping or regulating mechanisms by the co-addict, despite the fact that they are very often harmful and counterproductive. Thus, the addict and the co-addict follow "a similar path" to recovery, often involving addiction recovery groups, 12-step programs, networking with other addicts and co-addicts, etc. (Hentsch-Cowles & Brock, 2013, p. 329).

Trauma model. In contrast, proponents of the trauma model view the persistent negative behaviors of the partners of sex addicts as natural reactions to an intensely traumatic experience, as opposed to evidence of co-addiction (Steffens & Rennie, 2006). Viewed through the lens of trauma, “hypersensitive” or “obsessive” behaviors exhibited by partners of sex addicts are recognized as “attempts to survive and adapt to a dangerous situation . . . [and to] avoid painful stimuli” (Steffens & Rennie, 2006, p. 262). Unfortunately, supporting literature on the trauma model is relatively sparse, and the clinical trauma experienced by these individuals has not received adequate attention. For instance, in a 2002 article by McCarthy entitled “The Wife’s Role in Facilitating Recovery from Male Compulsive Sexual Behavior,” the word *trauma* appears only one time (McCarthy, 2002, p. 276).

From a trauma perspective, recovery for the partners of sex addicts is often described in terms similar to other forms of trauma, including stages such as (a) developing/pre-discovery, (b) crisis/decision/information gathering, (c) shock, (d) grief/ambivalence, (e) repair, and (f) growth (Hentsch-Cowles & Brock, 2013; Lebowitz, Harvey, & Herman, 1993). This treatment paradigm is much different from that of the addictions model, and more appropriate for individuals who are indeed suffering from trauma.

Thus we see that the addiction model asserts that partners of sex addicts ought to be treated for their own unique illness (co-addiction), while the trauma model asserts that WSAs are suffering from the effects of a terribly distressing experience (Steffens & Rennie, 2006). Unfortunately, the partners of sex addicts may or may not be receiving optimal care, depending on what they are actually experiencing, and which paradigm is used to diagnose and treat their unique problems. To the extent that the TIPSA can be validated as an appropriate instrument to detect trauma in the partners of sex addicts, it has the potential to help the partners of sex addicts receive the diagnoses and treatment that they need and deserve.

Cumulative trauma. Research on trauma more generally has shown that the degree of trauma experienced is influenced not only by the nature and severity of the traumatic event itself, but also by the level of interpersonal betrayal, and previous exposure to other (perhaps unrelated) traumatic experiences. This phenomenon is often referred to as *cumulative trauma* (Follette, Polusny, Bechtle, & Naugle, 1996; Martin et al., 2013; Schumm, Briggs-Phillips, & Hobfoll, 2006). For example, Follette et al. (1996) found in their sample that

> individuals who reported multiple types of victimization experiences [over time] showed increasingly higher levels of post-trauma symptomatology… such as anxiety, depression, and dissociation. The finding that the level of exposure to traumatic experiences has a cumulative effect is consistent with findings in combat related PTSD (p. 33).

These findings were corroborated by Martin et al. (2013) and Schumm et al. (2006), who independently found that the level of trauma experienced by individuals is cumulative over time, such that traumatic events early in life exacerbate the PTSD symptoms related to traumatic events later in life. In context, this implies that individuals who previously experienced sexual abuse, neglect, or other traumatic events are likely to be more intensely traumatized by the revelation of their partner’s sex addiction. The fact that such a revelation almost certainly represents a profound betrayal of trust only compounds the traumatic nature of the event.

Perhaps even more troubling, is the suggestion that “women who were sexually abused as children show an increased likelihood of being revictimized later in life” (Follette et al., 1996, p. 26; Schumm et al., 2006). That is, not only does trauma tend to compound over time, but individuals (particularly women) who were traumatized early in life are more likely to experience trauma again later in life. These are important issues that deserve consideration as we attempt to measure trauma in the spouses of sex addicts. The TISPA addresses the issue of cumulative trauma only indirectly, which may be insufficient for a scale explicitly designed to measure trauma in the partners of sex addicts (e.g., Item 11 under Exposure to Threat reads, "My partner's behaviors remind me of experiences I had earlier in my life."). In addition to the trauma items themselves, the TISPA also contains a brief set of demographic and mental health questions. These items revealed that 2.1% of the participants had a previous PTSD diagnosis, 10.1% had experienced a combination of PTSD and depression or anxiety, and 4.8% had experienced PTSD and some other mental disorder. Thus, a sizeable proportion of participants reported previous experience with PTSD, which may have exacerbated any trauma that resulted from their partner's sex addiction. Future work on the TIPSA should consider the impact and nature of cumulative trauma within the context of sex addiction, and how it might be more explicitly integrated into TIPSA items.

Scale Development and Psychometric Analysis

Like many other psychological constructs, trauma is a latent variable; it is not directly observable, and it lacks a concrete scale or units of measurement. Therefore, any attempt to measure trauma requires us to make inferences based on observable evidence such as responses to items on a questionnaire. Determining the extent to which these observable responses accurately measure what they are intended to measure requires thoughtful and technical analysis.

Classical test theory. The most traditional framework for modeling the relationship between total test scores and their targeted latent trait is Classical Test Theory (CTT) (de Ayala, 2009; Traub, 1997). CTT assumes that the instrument or test is unidimensional (i.e., it only measures a single latent trait), and that individual *i* can be located along the trait continuum as follows,

$$X_i = T_i + E_i \tag{1.0}$$

where X_i is an individual's observed score on the instrument, T_i is the individual's *true score* or expected score when the error is random, E_i is the error of measurement, and the *i* subscript denotes variation across persons (de Ayala, 2009). Therefore, a person's observed score on the instrument (X) is reflective of his or her *true* ability (T) plus random error (E) (de Ayala, 2009). Importantly, CTT was spawned out of a desire to understand and explain measurement error, or variations in observed scores around what was presumed to the fixed true score or actual ability level of a respondent (Traub, 1997).

While CTT is a useful starting point for modeling latent traits using total test scores, it does have some limitations. First, it should be obvious that a person's true score is an unknown quantity that is never directly observed. Consequently, the amount of error present in any given administration is also unknown, making it impossible to assess model-data fit. Indeed, "in CTT we do not examine model-data fit [but] simply assume the model to be true" (de Ayala, 2009, p. 7). Item response theory (IRT) overcomes this deficiency by allowing us to explicitly test how well a given model fits the data.

In addition, the unit of analysis in CTT is the composite score. Reliability, for example, is estimated only once for the entire instrument, and all individuals with equivalent raw scores are assumed to possess the same level of the targeted trait, regardless of which particular items were answered correctly (de Ayala, 2009). Furthermore, CCT defines item difficulty as simply the proportion of examinees who answered the item correctly which implies that item difficulty is constant across levels of the trait being measured. In other words, in CTT an item is assumed to be equally difficult for all members of the targeted population regardless of how much of the measured trait they possess.

Item response theory. In contrast to CTT which focuses on the total score that a respondent receives in response to a series of test or questionnaire items, IRT focuses at the item level. IRT includes a family of mathematical models which attempt to describe how the probability of choosing a particular answer in response to an item varies as a function of the person's trait level and specific characteristics of the item such as its overall difficulty and its discriminating power. All IRT models include an estimate of the respondent's trait level (i.e., how much of the trait or ability each person is estimated to possess). The various models differ in terms of which item characteristic they attempt to model, but their common purpose is to locate both persons and items along the latent trait continuum, thus improving our ability to make accurate inferences regarding possession of the trait in question. This study focuses on the use of two IRT models: (a) the generalized partial credit model (GPCM) and (b) the nominal response model (NRM).

IRT methods and models are well suited for analysis of psychometric scales because their very purpose is to define the relationship between a latent trait or ability (generally referred to as the Greek letter theta) and their manifestations (i.e., item responses), with emphasis on specific item and option characteristics and functioning (de Ayala, 2009; DeMars, 2010; Reckase, 2009). The targeted latent trait might be some skill or proficiency, or it could be some other psychological variable such as a particular attitude, aptitude, or belief (DeMars, 2010). Item responses might be dichotomous (e.g., yes/no, right/wrong) or polytomous (more than two categories on a Likert scale, for example) (DeMars, 2010). In this particular study, the targeted trait under consideration is trauma experienced by the partners of sex addicts, and the outward manifestations of that trait are responses to polytomous TIPSA items on a common 5-point scale.

Recent research illustrates the usefulness of IRT methods, not only within the context of educational assessments, but also in broader contexts such as the development of scales to measure psychological constructs (Murray et al., 2016; Preston et al., 2015; Preston & Reise, 2014a). In particular, the IRT concepts of item discrimination power and item and scale information are relevant to this discussion.

First, some IRT models include an item discrimination parameter. Models which include this parameter provide insight into how well a given item differentiates between individuals who have similar levels of the targeted trait. A poorly discriminating item provides little information about who does and who does not possess the trait over a given range of theta. In contrast, a well discriminating item will make more accurate distinctions between possessors of the trait over a narrow range of the trait scale. For example, some items may be more discriminating at lower levels of the targeted trait, while others are more discriminating at higher levels. The goal is to develop a coherent set of items that collectively discriminate across all relevant levels of the trait continuum. Items that poorly discriminate can be identified and flagged for review, modification, or removal.

In addition, IRT can be used to calculate a measure of precision, or *information*, for each item, across levels of targeted trait. IRT methods produce both an *item information function* (IIF) for each item in the scale and a *test information function* (TIF) for the scale as a whole. Each IIF graphically displays how the precision of the trauma estimates vary as a function of increasing levels of that trait. Similarly, the TIF will graphically display how the precision of the trait estimates of a specific factor vary as a function of increasing levels of that factor.

While both the GPCM and the NRM have been specifically designed for use with polytomous items (i.e., questionnaire items that have more than two response options), they contain important differences in terms of how they operationalize discrimination and option ordering.

Generalized partial credit model. The GPCM (Muraki, 1997) assumes that response options are empirically ordered, but estimates a unique set of thresholds for each item. In addition, the GPCM conceptualizes discrimination as a property of the item, and estimates category response functions (CRFs) using a single discrimination parameter (α) for each item, as follows,

$$P_{ix}\theta = \frac{exp \sum_{j=0}^{x} \alpha_i(\theta - \delta_{ij})}{\sum_{r=0}^{m}[exp \sum_{j=0}^{x} \alpha_i(\theta - \delta_{ij})]} \tag{2.0}$$

where $\sum_{j=0}^{x} \alpha_i\left(\theta - \delta_{ij}\right) = 0$ for model identification purposes. The GPCM does not provide any indication of respondents' ability to distinguish between adjacent response options, because slope parameters for response options within items are not estimated.

Nominal response model. In contrast, the NRM developed by Bock (1972) was initially intended to analyze polytomous items with unordered categories (e.g., multiple choice tests), and the slope of each CRF was assumed to vary across items (de Ayala, 2009). That is, the NRM makes no assumptions about category intersection ordering, and estimates unique a slope parameter for each response option. According to the NRM, the conditional probably of an individual with trait level θ responding in category x ($x = 0 \ldots m_i$) on item i can be written as follows,

$$P_{ix}(\theta) = \frac{\exp(\alpha_{ix}\theta + c_{ix})}{\sum_{x=1}^{m} \exp(\alpha_{ix}\theta + c_{ix})} \tag{3.0}$$

where $\sum \alpha_{ix} = \sum c_{ix} = 0$ for the purpose of model identification. When this model is used, a unique set of slope and intercept parameters (α and c respectively) are freely estimated for each response option within an item. In other words, each response option is assumed to have its own unique slope parameter. Thus, the NRM is the least restrictive IRT model and has traditionally been applied in settings where the response options have no natural order because it allows researchers to determine the effectiveness of distractors in multiple choice tests, for example.

Model comparisons. The equations for these two models reveal that the GPCM can be derived from the NRM by constraining the discrimination parameter α to be constant across all options within a given item. In other words, the GPCM models discrimination as a property of items, while the NRM models a slope parameter for each response category. The distinction between overall item discrimination versus option slopes is particularly evident in the subscripts that are used in each model. The NRM places two subscripts, i and x, on the discrimination parameter (α) indicating that discrimination is allowed to vary across items and across options. In contrast, the GPCM places only a single subscript, i, on α indicating that discrimination is allowed to vary across items, but not across options within an item. Hence, the GPCM is a constrained version of, and nested within, the NRM.

It should also be clear that the GPCM is more parsimonious than the NRM, because the GPCM estimates a single discrimination parameter for each item while the NRM estimates a unique slope parameter for each option within each item. Preston and Reise (2015) have pointed out that the discrimination parameter estimated by the GPCM for each item should be roughly equivalent to the weighted average of the CBD parameters from the NRM, although the defaults used in some software packages may produce slight discrepancies. In any event, equations 2.0 and 3.0 illustrate key differences between the NRM and the GPCM, and the way in which they operationalize discrimination at the option versus the item level.

Because the NRM makes fewer assumptions and its' parameter estimates are more free to vary relative to other more constrained models like the GPCM, it is less parsimonious but will generally improve model-data fit. Fit statistics can be used to determine the extent to which one model fits the data better than another.

Use of the NRM in scale development. As noted previously, the NRM was originally designed for use with unordered polytomous items, but in recent research it has been applied to ordered polytomous items to determine whether the response options are empirically ordered as intended, and the degree to which each successive option discriminates amongst individuals with increasingly higher levels of the trait (Preston, 2014a; Preston et al., 2015; Preston & Reise, 2014a, 2014b; Preston et al., 2011). One way to conceptualize discriminatory differences between options is to utilize category boundary discrimination (CBD) parameters (Preston & Reise, 2014a). Preston and Reise (2014a, p. 389) pointed out that, because the NRM estimates a separate discrimination parameter (α_x) for each response category, the "discrimination of the distinction between categories x and x-1" can be understood as the difference between α_x and α_{x-1}. They labeled this statistic a^* and called it the CBD parameter (Preston & Reise, 2014a). Furthermore, they argued that when category responses are assumed to be ordered, CBDs that are near zero indicate a lack of distinction between categories x and x-1; that is, "people (or raters) cannot differentiate between the response options" (Preston & Reise, 2014a, p. 389).

Utilizing CBD parameters to assess the discrimination of the distinction between categories is a relatively new approach that has only recently been applied to empirical data, but the original concept and formula were outlined much earlier by Thissen, Steinberg, and Fitzpatrick (1989). They pointed out that when the slopes for adjacent category response curves (i.e., α_x and α_{x-1}) are equal, "their trace lines are proportional at all values of θ; that, in turn, means that the information about θ is the same when either of these two response categories is chosen" (Thissen et al., 1989). Thus, Preston and Reise's (2014a) discussion of CBD parameters is simply an innovative extension and practical application of the concepts put forward by Thissen et al. (1989). The fact that the NRM estimates parameters for CRFs independently for each item makes CBD calculations possible, and allows us to observe empirical option functioning. This, in turn, allows us to determine the extent to which the item responses are ordered as intended by the scale developers.

Empirical option functioning estimated by the NRM is useful in the context of scale development, because it is common for psychological instruments to utilize response option scales that are ordered from least to greatest, best to worst, least frequent to most frequent, etc. The TIPSA presents respondents with a series of pertinent statements, and asks them to respond on a frequency continuum ranging from 1 (*never*), 2 (*occasionally/rarely*), 3 (*about half the time*), 4 (*more often than not*), to 5 (*always*). The assumption is that if the response options are inherently ordered, then higher item scores ought to reflect more of the target trait (Preston & Reise, 2014a). That is, as the level of the trait increases, so should the response options that are endorsed on an ordered scale or frequency continuum. Concerns arise, however, when empirical response data reveal that respondents did not respond in an ordered manner as expected (Nilsson, Sunnerhagen, & Grimby, 2005; Preston & Reise, 2014a).

In response to empirical category disordering, some scholars have utilized the approach taken here, namely, collapsing response categories to improve ordering (Ashley et al., 2013; Bee, Gibbons, Callaghan, Fraser, & Lovell, 2016; Bell et al., 1994; Bourke & Wallace, 2015; Brogårdh, Lexell, & Lundgren-Nilsson, 2013; das Nair, Moreton, & Lincoln, 2011; Dougherty, Nichols, & Nichols, 2011; Oluboyede & Smith, 2013). However, using simulated data, García-Pérez (2017) found that "collapsing categories only reduces the item information function and, hence, deteriorates the instrument" (p. 23). He argues that we ought to be more concerned with the information that an item provides at any given value of theta, than with the empirical ordering of the response options. In contrast, Preston and Reise (2014a) argue that "if a model that assumes that higher item scores reflect more of the trait is incorrectly applied, distortion of the scaling of individual differences may occur" due to nuisance or error variance (p. 396).

Appropriate measures of scale reliability may be useful in this context, to determine whether or not a rescoring procedure reduces "nuisance variation" in the resulting test scores (Preston & Reise, 2014a, p. 396). While reliability has been operationalized in several different ways, it is generally defined as "the extent to which the variance of an observed variable is explained by the true score that the variable is designed to measure. In other words, the reliability is defined as the ratio of the true score variance to the total variance of the observed measure" (Wang & Wang, 2012, p. 86). Thus, an increase in reliability would be an indication of a decrease in nuisance variation. While this may be one useful way to examine the impact of category response disordering and rescoring on nuisance variation, more empirical research in this area is warranted. Unfortunately, supporting literature on the NRM in scale development remains somewhat sparse, and research on its' practical application has been "lacking" (Preston & Reise, 2014a, p. 387).

In particular, the impact of nonnormal data on NRM parameter estimates is an issue that is important, yet "much understudied" (Preston & Reise, 2014b, p. 377). It is frequently noted that many psychological constructs, including trauma, are unlikely to be normally distributed in the population (Preston & Reise, 2014b). Yet it is common for researchers and scale developers to employ Marginal Maximum Likelihood Estimation (MML) or Expected Maximization (EM) algorithms that utilize a set of normal quadrature points that essentially ignore the distribution of the data, and assume that that the underlying construct is normally distributed in the population

(Cooper, Balsis, & Zimmerman, 2010; Kim, Kim, & Kamphaus, 2010). In other words, while it may be reasonable to assume that educational outcomes like test scores are normally distributed in the population, the assumption of normality is unlikely to hold in other fields such as personality or psychopathology (Woods, 2006). When response data are skewed and the construct being measured is likely to be nonnormally distributed in the population, traditional MML and EM estimation techniques may be inappropriate, and may result in “gross inaccuracies in the estimation of item parameters” (Preston & Reise, 2014b, p. 397).

Confirmatory factor analysis. Confirmatory factor analysis (CFA) is a commonly used statistical tool in scale development, because it allows the developer to formally and explicitly test the underlying structure of the unobserved latent factors using observed measures (i.e., indicators like item responses or observational ratings). Indeed, CFA is “almost always” used in the process of scale development to examine and test the underlying theoretical factor structure (Brown, 2015, p. 1). Factor analysis techniques like CFA shed light on the relationship(s) between items and between latent factors, and allow us to have confidence that a common set of items measure the same latent construct. In contrast, IRT is used to locate both persons and items along a latent trait continuum, given a person’s ability and the properties of the items, and provides us with additional information about item and response option functioning (Brown, 2015; de Ayala, 2009). Although, CFA and IRT share some analogous features and concepts (Brown, 2015; Reckase, 2009), they conceptualize scale development issues quite differently. While CFA focuses primarily on the factor structure without regard for the properties of the items themselves, IRT examines the latent factor with explicit emphasis on item characteristics (Reckase, 2009). For example, IRT methods can be used to determine which item response options are most likely to be endorsed at various levels of the targeted latent trait (de Ayala, 2009).

CFA is widely used in scale development, but the NRM of IRT, despite being developed more than forty years ago, is only recently gaining ground as important tool for that purpose (Bartholomew, Knott, & Moustaki, 2011; Murray et al., 2016; Preston et al., 2015; Preston & Reise, 2014a). Researchers and professionals in psychological measurement and scale development are beginning to recognize that, in addition to a sound understanding of the latent factor structure, properly functioning items and options are also important (Murray et al., 2016). The apparent interplay between CFA and IRT represents something of a "chicken or egg" conundrum that is undeniably important. Presently, the impact of IRT analysis on CFA outcomes, and vice versa, has not been sufficiently examined in the literature, and deserves future attention.

CHAPTER 3: Method

Participants and Sample

Data for this study came from a national administration of the TIPSA in 2015 via a link embedded in a blogpost by Skinner (2015) titled *The Lasting Effects of Sexual Betrayal* written for PsychologyToday.com. Using online SurveyMonkey™ software, responses were collected from 3,199 voluntary participants, who accessed the survey through the website. Prior to taking the survey, respondents were informed that their responses would be used for research purposes and that voluntary completion of the TIPSA implied consent.

Approximately 22% of the sample did not provide demographic information. Of the 78% who did provide demographics, 90.5% were between the ages of 21 and 55, 58.2% were currently married, 85.6% had at least some college education, and just over 93% identified as female. The most common age category was 36-40 years old.

Table 1 describes the demographics (age, gender, relationship status, educational background, self-reported mental health history, and religious preference) for the full sample.

Participant Inclusion

Because the data contained only 6% male respondents, I determined to focus exclusively on the responses of females in this study. Hence, those who identified as males and any respondents who did not specify their gender were excluded from the analysis leaving a sample size of 2,339. The reduced sample is described in Table 2. It is fairly common to assume that the partners of sex addicts are female, and sex addiction is primarily a male problem (McCarthy, 2002; Milrad, 1999).

Table 1

Demographics of Respondents in Full Sample (N = 3,199)

Category	n	%	Category	n	%
Gender			Mental health history		
Male	164	5.1	Anxiety only	156	4.9
Female	2,339	73.1	ADHD only	50	1.6
Missing	696	21.8	Bipolar Disorder only	6	0.2
Age range			BPD only	5	0.2
18-20	49	1.5	Depression only	269	8.4
21-25	234	7.3	Narcissism only	2	0.1
26-30	352	11.0	PTSD only	52	1.6
31-35	421	13.2	Substance related only	6	0.2
36-40	389	12.2	Depression + anxiety	399	12.5
41-45	392	12.3	PTSD + Dep and/or Anx	251	7.9
46-50	256	8.0	PTSD + any other combo	115	3.6
51-55	226	7.1	Any other combination	201	6.3
56-60	120	3.8	None of the above	896	28.0
Over 60	68	2.1	Missing	791	24.7
Missing	692	21.6	Religious preference		
Relationship status			Agnostic	128	4.0
Single	224	7.0	Atheist	170	5.3
Committed relationship	613	19.2	Born Again Christian	165	5.2
Married	1,192	37.3	Baptist	67	2.1
Married currently separated	265	8.3	Buddhist	38	1.2
Divorced	210	6.6	Catholic	290	9.1
Missing	695	21.7	Christian	545	17.0
Education			Hindu	18	0.6
GED	79	2.5	Jewish	22	0.7
High School Graduate	281	8.8	Lutheran	27	0.8
Some College	738	23.1	LDS (Mormon)	308	9.6
Associate Degree	284	8.9	Methodist	32	1.0
Bachelors Degree	681	21.3	Muslim	33	1.0
Graduate Degree	442	13.8	Protestant	48	1.5
Missing	694	21.7	Wicca	18	0.6
			I am not religious	518	16.2
			Spiritual	31	1.0
			Other	35	1.1
			Missing	706	22.1

Table 2

Demographics of Respondents in Reduced Sample (N = 2,339)

Category	n	%	Category	n	%
Age range			Mental health history		
18-20	45	1.9	Anxiety only	146	6.2
21-25	211	9.0	ADHD only	40	1.7
26-30	338	14.5	Bipolar Disorder only	3	0.1
31-35	391	16.7	BPD only	5	0.2
36-40	357	15.3	Depression only	254	10.9
41-45	366	15.6	Narcissism only	1	0.0
46-50	240	10.3	PTSD only	49	2.1
51-55	215	9.2	Substance related only	5	0.2
56-60	111	4.7	Depression + anxiety	376	16.1
Over 60	61	2.6	PTSD + Dep and/or Anx	237	10.1
Missing	4	0.2	PTSD + any other combo	112	4.8
Relationship status			Any other combination	191	8.2
Single	191	8.2	None of the above	829	35.4
Committed relationship	572	24.5	Missing	91	3.9
Married	1130	48.4	Religious preference		
Married currently separated	250	10.7	Agnostic	112	4.8
Divorced	190	8.1	Atheist	149	6.4
Missing	6	0.3	Born Again Christian	158	6.8
Education			Baptist	64	2.7
GED	72	3.1	Buddhist	35	1.5
High School Graduate	257	11.0	Catholic	268	11.5
Some College	695	29.7	Christian	514	22.0
Associate Degree	270	11.5	Hindu	15	0.6
Bachelors Degree	638	27.3	Jewish	20	0.9
Graduate Degree	402	17.2	Lutheran	26	1.1
Missing	5	0.2	LDS (Mormon)	301	12.9
			Methodist	31	1.3
			Muslim	29	1.2
			Protestant	45	1.9
			Wicca	16	0.7
			I am not religious	478	20.4
			Spiritual	30	1.3
			Other	31	1.3
			Missing	17	0.7

Description of Sample

Supplementary information gathered in the TIPSA included (a) how long the respondent had known of her partner's behavior (Table 3), (b) how long she had been experiencing the symptoms described in the survey (Table 3), (c) how frequently it was difficult for her to fulfill important life roles (Table 4), and (d) the sexual behaviors in which her partner had been involved (Table 5). Over 73% of respondents had known about their partner's behavior for a year or more, and 32% had known for more than 5 years. Seventy percent had been experiencing the trauma symptoms described in the survey a year or more, and over 25% had experienced symptoms for more than 5 years. Spearman's rho coefficient revealed that the reported length of time of having known about their partner's behavior and the length of time experiencing trauma symptoms were significantly correlated $r_s = .813$, $p < .01$. The TIPSA did not, however, ask participants to report the severity of their symptoms, so we do not know if symptoms improved or became more severe over time.

Table 3

Length of Time Respondents Have Known About Their Partner's Behavior, and How Long They Have Reportedly Experienced Trauma-Related Symptoms

	How long known of partner's behavior		Experiencing symptoms	
Length of time	n	%	n	%
Less than one month	81	3.5	94	4.0
2 – 3 months	135	5.8	150	6.4
4 – 6 months	173	7.4	190	8.1
7 – 12 months	240	10.3	267	11.4
More than one year but less than two	343	14.7	362	15.5
More than two years but less than five	618	26.4	671	28.7
More than five years	748	32.0	597	25.5
Missing	1	0.0	8	0.3
Total	2,339	100.0	2,339	100.0

Table 4

It Has Become Difficult for Me to Fulfill Important Roles (That of Employee, Parent, etc.) Since Discovering My Partner's Sexual Behaviors

Response category	n	%
Never	228	9.7
Occasionally / Rarely	628	26.8
About half the time	618	26.4
More often than not	582	24.9
Always	275	11.8
Missing	8	0.3
Total (reduced sample)	2,339	100.0

The TIPSA also asked respondents how frequently it was difficult for them to fulfil important life roles since learning of their partner's behavior. Nearly 12% of respondents selected *always*, and over 63% reported that fulfilling important roles was difficult *about half the time* or more (Table 4). Less than 10% indicated that it was never difficult for them to fulfil important life roles. The ability to fulfill life roles was uncorrelated with length of time that respondents had known about their partner's behaviors, and only marginally correlated with experiencing trauma symptoms $r_s = .06$, $p < .01$.

Finally, the TIPSA asked respondents to identify the compulsive sexual behaviors exhibited by their partners. A full summary of this item is found in Table 5. Pornography was the most commonly reported sex addiction at nearly 80%, followed by "other" (75.4%) and masturbation (71.1%; Table 5). Over 1,700 (75.4%) respondents selected the *other* category and left an open-ended response describing their partner's compulsive sexual behaviors. A detailed qualitative analysis of these responses is beyond the scope of the current study, but it should be noted that they included a wide variety of socially acceptable (e.g., homosexuality, various types of pornography), and unacceptable (e.g., rape, molestation, bestiality), sexual activities and behaviors.

Table 5

Please Indicate Which Behaviors Your Partner Has Been Involved In (Select All That Apply; Reduced Sample, n = 2,339)

Behavior	n	%
Pornography	1,870	79.9
Other	1,764	75.4
Masturbation	1,664	71.1
Infidelity – multiple affairs	1,039	44.4
Hooked up with a random person for a sexual encounter	829	35.4
Infidelity – one time	505	21.6
Visiting topless bars	456	19.5
Illegal sexual activity	267	11.4
Going to massage parlors	235	10.0
Watched someone in their home or apartment	207	8.8

Instrument

The TIPSA was designed to identify and measure trauma in the spouses or partners of sex addicts (Skinner, 2015). The TIPSA is intended to be used as a psychological measurement and screening instrument rather than as a tool to make diagnostic distinctions. While measuring trauma may facilitate correct diagnosis, the TIPSA alone is not intended to diagnose any mental disorder.

The TIPSA was, however, modeled after the five diagnostic criteria for Posttraumatic Stress Disorder (PTSD) outlined by the American Psychiatric Association (2015). For clarity, Skinner (2015) termed these criteria as (a) exposed to threat, (b) reliving the event, (c) avoidance, (d) cognition and negative mood, and (e) emotional arousal. Guided by these five factors, 53 items were developed and made available via an online survey system (

).

Table 6

Number of Items by Factor (Diagnostic Criteria)

Factor	Criterion name	Number of items
1	Exposed to threat	11
2	Reliving the event	10
3	Avoidance	6
4	Cognition and negative mood	13
5	Emotional arousal	13
Total		53

All items utilized a common set of response options that defined a frequency continuum ranging from 1 (*never*), 2 (*occasionally/rarely*), 3 (*about half the time*), 4 (*more often than not*), to 5 (*always*). In addition to the TIPSA items, the instrument gathered general demographic information on gender, age range, current relationship status, educational background, religious preference, and self-reported mental health history (see Tables 1 and 2).

Analysis

Data were downloaded from SurveyMonkey™ in an SPSS (Version 24.0) file. SPSS was used to calculate basic descriptive statistics and to clean and otherwise prepare the data. Because flexMIRT requires a 0 starting point, all items designed to measure trauma that were used in the IRT analysis were recoded from 1 through 5, to 0 through 4. IRT analyses were performed using flexMIRT software (Cai, 2013). The output from flexMIRT was read into R 3.3.1 (R Core Team, 2014) and R Studio 0.99.903 (RStudio Team, 2015) to calculate CBD parameters and category intersection points, and to plot the corresponding graphs. Standard IRT models assume construct unidimensionality, so I analyzed each of the five factors separately. Additional CFA is needed to confirm the five-factor structure of the TIPSA and to illustrate that each diagnostic criterion is in fact unidimensional.

To begin, the items in each of the five subscales were analyzed using the GPCM and again using the NRM. The NRM and GPCM are both divide-by-total models. When used to analyze data with ordered categories, the primary difference between the NRM and the GPCM is that the NRM estimates a slope parameter for each response option within an item, but the GPCM estimates only a single, global slope parameter for each item. The GPCM can be derived from the NRM by constraining the slope parameters for all the options within each item to be equal. Consequently, the GPCM is a special case of the NRM and is considered to be nested within the NRM. Since the NRM freely estimates more parameters than the GPCM, it will always fit the data equally well or better than the GPCM. In other words, the NRM will almost always reduce model-data misfit. However, this reduction may not be statistically significant (de Ayala, 2009; Preston & Reise, 2014a). In this case, chi-square difference tests and AIC/BIC fit statistics were used to compare the relative model-data fit of the GPCM and the NRM, and to determine whether or not modeling each category response option independently using the NRM significantly reduced model model-data misfit. To visualize the response option functioning, CRFs were plotted using R (Preston, 2014c).

CBD parameters were calculated using an R script that simply subtracts each option discrimination parameter from its' adjacent neighbor (Preston & Reise, 2014a). Very low (approaching 0) CBD parameters indicated that moving from one category to the next does not meaningfully contribute to overall item discrimination, while high CBD parameters indicate that the neighboring response option discriminates meaningfully over the next-higher range of theta (i.e., trauma) estimates (Preston & Reise, 2014a).

The Wald test (Wald, 1945) was also calculated using an R script, to determine whether there was a significant difference between CBD parameters (Preston, 2014b). Items with statistically significantly different CBDs based on the Wald test were flagged, and the response options that comprised the least discriminating CBD were collapsed through recoding.

Category intersection points were also calculated to determine whether or not the response options were empirically ordered over increasing levels of theta, as intended by the scale developer. When a response option set was found to be empirically disordered, the least discriminating option was collapsed into its' nearest neighbor in an attempt to improve discrimination, information, and option ordering.

After all problematic options were identified and collapsed as described, the process of parameter estimation and Wald testing was repeated until the item either became dichotomous (i.e., only two response options remained), or the remaining options were empirically ordered and sufficiently discriminating.

In addition to using the Wald test and intersection parameters to identify faulty item functioning, I also performed a visual inspection of all CRFs to identify poorly performing options. Such options were similarly recoded, and the changes were integrated into subsequent iterations. After all problematic options were identified and collapsed as described, I compared the pre- and post-analysis NRM model fit statistics to determine if the proposed changes reduced model-data misfit.

At the conclusion of rescoring and parameter estimate procedures, I addressed the issue of nuisance variation and reliability by calculating three sets of reliability coefficients for each factor: (a) Cronbach's alpha (Cronbach, 1951), (b) McDonald's omega (McDonald, 2013), and (c) Raykov's rho (Raykov, 2009). Each of these were calculated from CFA factor loadings generated using MPlus software (Muthen & Muthen, 2015). Because the factor loadings were not equal, and Chronbach's alpha assumes that the measures are tau-equivalent (i.e., that differences in the factor loadings are not statistically significantly different), I also calculated McDonald's omega, which only assumes that the measures are congeneric (i.e., that the factor loading are not equal and that error variances are not equal). However, omega and rho also assume that the CFA model fits the data. Therefore, the largest error covariances between item pairs, as revealed by the modification indices, were correlated to improve model-data fit. Raykov's rho was also calculated, because it explicitly accounts for these correlated error variances. Examining all three of these estimates of reliability provided us with an indication of the reliability of responses from each factor, and the extent to which the rescoring procedures reduced extraneous nuisance variation in test scores for each factor.

CHAPTER 4: Results

Analyses were performed as outlined in the method section, and all model estimation procedures converged normally. Factors 2 and 5 required an increase in the number of estimation cycles in order to converge, but ultimately reached normal termination and convergence.

Model Comparisons

Across each factor, the NRM generated smaller AIC statistics compared to the GPCM indicating better model-data fit, but larger BIC statistics indicating worse model-data fit (Table 7). This result is not necessarily surprising, given that BIC more strongly penalizes a lack of parsimony, and the NRM is a less parsimonious model. In general, the NRM can be expected to fit the data at least somewhat better than more constrained models which are nested within the NRM because it imposes fewer constraints and leaves more parameters free to vary. In this particular case, chi-square difference tests revealed that the NRM produced a statistically significant reduction in model-data misfit across each of the five factors relative to the GPCM (Table 7). However, the change in R^2, which can be interpreted as a kind of effect size, was relatively small across all factors. For example, in Factor 1 (Exposure to Threat), the NRM explained only one-sixth of one percent (0.17%) more of the variance than the GPCM (Table 7).

Despite the larger BIC statistics and the rather small effect sizes, I determined to proceed with the analysis using the NRM because it allowed me to estimate CBD parameters and observe empirical option functioning (i.e., ordering and discrimination).

Table 7

Comparison of Fit Statistics for the Generalized Partial Credit Model (GPCM) and Nominal Response Model (NRM) by Factor.

Model	Parameters	-2LL	DF	ΔG^2	P	ΔR^2	AIC	BIC
Factor 1: Exposed to threat								
GPCM	55	62464.06	1992				62574.06	62890.72
NRM	88	62356.89	1959	107.17*	0.000	0.0017	62532.89	63039.55
Difference	-33	107.17	33					-148.83
Factor 2: Reliving the event								
GPCM	50	61412.42	973				61512.42	61800.30
NRM	80	61325.76	943	86.66*	0.000	0.0014	61485.76	61485.76
Difference	-30	86.66	30				26.66	-146.06
Factor 3: Avoidance								
GPCM	40	40744.01	33				40804.01	40976.73
NRM	64	40695.65	15	48.36*	0.000	0.0001	40791.65	41068.01
Difference	-18	48.36	18				12.36	-91.28
Factor 4: Cognition and negative mood								
GPCM	65	80884.63	8126				81014.63	81388.87
NRM	104	80726.08	8087	158.55*	0.000	0.0019	80934.08	81532.86
Difference	-39	158.55	39					-143.99
Factor 5: Emotional arousal								
GPCM	65	72655.06	8126				72785.06	73159.29
NRM	104	72395.34	8070	259.72*	0.000		72603.34	73202.12
Difference	-39	259.72	39				181.72	-42.83

* Significantly improved model-data fit from using the NRM ($p < .001$)

Analyses revealed that all five factors contained items with problematic options including within-item CBD parameter estimates that varied significantly according to the Wald test ($\alpha < .05$; Table 8) and/or empirically disordered intersection parameters (Table 9). Because larger CBDs indicate more informative category functioning, when CBD parameters were found to be significantly different within a particular item, the two adjacent response options comprising the lowest CBD were collapsed into a single category by recoding (Preston & Reise, 2014a). When category intersections were disordered, the two response options that comprised the lowest CBD parameter were similarly collapsed through recoding in order to improve functioning (Ashley et al., 2013; Oluboyede & Smith, 2013).

Parameter estimation, checking for problematic CBDs and intersections, and recoding was an iterative process that continued until all category intersections were ordered and CBD parameters were not significant, or until there were only two response options remaining. Parameters from the initial estimation of CBDs (Table 8) and intersections (Table 9) are reported in this paper, but parameter estimates from intermediate iterations are not included. A description of the final rescoring structures, and a summary of the number of response options per item is found in Table 10.

After the initial round of parameter estimation, the following items did not contain significant CBD parameters or disordered intersections:

- Factor 1: items 1, 2, 5, 10;
- Factor 2: items 12, 13, 15, 16, 17, 18, 20;
- Factor 3: items 22, 23, 28, 29;
- Factor 4: items 25, 26, 30, 31, 34, 35, 37, 38, 39; and
- Factor 5: items 41, 42, 43, 45, 46, 47, 48, 49, 50.

Table 8

Category Boundary Discrimination (CBD) Parameters by Item by Factor and by Model.

	NRM				
Item	α_1^* (*SE*)	α_2^* (*SE*)	α_3^* (*SE*)	α_4^* (*SE*)	α_{GPCM}(*SE*)
Factor 1					
1	0.73 (0.19)	0.71 (0.10)	1.00 (0.10)	0.76 (0.10)	0.85 (0.05)
2	0.95 (0.29)	0.85 (0.13)	1.02 (0.10)	0.82 (0.09)	0.92 (0.05)
3	**1.31 (0.12)**	**0.31 (0.22)**	**0.32 (0.27)**	**0.39 (0.32)**	0.75 (0.07)
4	0.47 (0.47)	0.69 (0.09)	0.20 (0.11)	0.36 (0.10)	0.44 (0.03)
5	0.76 (0.76)	0.66 (0.10)	0.81 (0.10)	0.75 (0.10)	0.74 (0.04)
6	0.01 (-0.14)	0.05 (0.08)	-0.39 (0.11)	-0.78 (0.18)	-0.17 (0.02)
7	0.41 (0.39)	0.49 (0.16)	0.69 (0.13)	1.02 (0.09)	0.80 (0.06)
8	0.25 (0.25)	0.82 (0.12)	0.46 (0.09)	0.77 (0.08)	0.63 (0.04)
9	0.74 (0.08)	0.98 (0.15)	0.60 (0.18)	1.08 (0.26)	0.78 (0.06)
10	0.83 (0.09)	1.21 (0.14)	1.05 (0.19)	1.44 (0.26)	1.02 (0.08)
11	0.19 (0.06)	0.20 (0.08)	0.20 (0.09)	0.28 (0.10)	0.21 (0.02)
Mean	0.61	0.63	0.54	0.63	0.63
Factor 2					
12	0.05 (0.25)	1.18 (0.13)	0.80 (0.10)	1.05 (0.10)	1.09 (0.05)
13	0.02 (0.17)	0.83 (0.12)	0.76 (0.10)	1.00 (0.12)	0.99 (0.05)
14	**0.60 (0.09)**	**0.53 (0.09)**	**0.67 (0.10)**	**1.21 (0.13)**	0.72 (0.04)
15	0.31 (0.18)	0.69 (0.11)	0.61 (0.10)	1.25 (0.10)	0.87 (0.04)
16	0.74 (0.16)	1.30 (0.13)	1.16 (0.12)	1.67 (0.14)	1.35 (0.06)
17	0.82 (0.16)	0.85 (0.11)	1.04 (0.11)	1.33 (0.11)	1.06 (0.05)
18	0.54 (0.12)	0.75 (0.10)	0.82 (0.11)	1.07 (0.10)	0.85 (0.04)
19	0.46 (0.11)	0.55 (0.10)	0.36 (0.10)	1.08 (0.09)	0.64 (0.04)
20	0.85 (0.14)	0.88 (0.11)	0.67 (0.10)	1.04 (0.10)	0.89 (0.05)
21	**1.29 (0.20)**	**1.56 (0.13)**	**1.11 (0.11)**	**1.13 (0.12)**	1.29 (0.06)
Mean	0.57	0.91	0.80	1.18	0.98
Factor 3					
22	0.58 (0.17)	0.56 (0.09)	0.41 (0.07)	0.29 (0.08)	0.42 (0.03)
23	0.61 (0.16)	0.65 (0.10)	0.60 (0.08)	0.34 (0.08)	0.54 (0.04)
24	0.74 (0.14)	0.78 (0.13)	1.10 (0.13)	1.11 (0.13)	0.97 (0.06)
27	0.31 (0.08)	0.20 (0.09)	0.03 (0.08)	0.31 (0.08)	0.19 (0.02)
28	1.53 (0.18)	2.06 (0.30)	1.93 (0.39)	1.59 (0.39)	1.73 (0.14)
29	0.79 (0.11)	0.79 (0.10)	0.40 (0.10)	0.46 (0.11)	0.60 (0.04)
Mean	0.76	0.84	0.75	0.68	0.74

(table continued)

Table 8 (continued)

Item	NRM				
	α_1^*(*SE*)	α_2^*(*SE*)	α_3^*(*SE*)	α_4^*(*SE*)	α_{GPCM}(*SE*)
Factor 4					
25	0.74 (0.14)	0.91 (0.10)	0.64 (0.08)	0.64 (0.09)	0.73 (0.04)
26	0.99 (0.16)	1.09 (0.12)	1.05 (0.10)	1.16 (0.11)	1.09 (0.05)
30	0.06 (0.11)	0.75 (0.10)	0.51 (0.09)	0.90 (0.09)	0.61 (0.03)
31	0.22 (0.08)	0.54 (0.08)	0.47 (0.09)	0.95 (0.11)	0.52 (0.03)
32	**1.93 (0.36)**	**0.75 (0.15)**	**0.86 (0.11)**	**1.05 (0.08)**	0.96 (0.05)
33	0.37 (0.14)	0.57 (0.11)	0.60 (0.10)	1.00 (0.08)	0.71 (0.04)
34	1.09 (0.15)	1.15 (0.12)	0.80 (0.10)	1.17 (0.11)	1.03 (0.05)
35	0.65 (0.15)	1.04 (0.09)	0.74 (0.07)	0.73 (0.15)	0.84 (0.04)
36	0.77 (0.08)	0.68 (0.10)	0.65 (0.12)	1.04 (0.16)	0.74 (0.04)
37	1.19 (0.15)	1.60 (0.15)	1.09 (0.13)	1.79 (0.16)	1.43 (0.07)
38	0.89 (0.15)	1.43 (0.14)	1.16 (0.13)	1.61 (0.14)	1.32 (0.07)
39	0.24 (0.14)	0.65 (0.08)	0.36 (0.06)	0.48 (0.08)	0.46 (0.03)
40	0.06 (0.06)	0.17 (0.07)	-0.04 (0.08)	0.28 (0.10)	0.10 (0.02)
Mean	0.71	0.87	0.68	0.99	0.81
Factor 5					
41	0.77 (0.11)	0.62 (0.09)	0.63 (0.09)	0.98 (0.10)	0.75 (0.04)
42	0.52 (0.22)	0.83 (0.11)	0.86 (0.10)	1.56 (0.11)	1.07 (0.06)
43	0.29 (0.21)	0.91 (0.12)	0.74 (0.10)	1.83 (0.13)	1.12 (0.06)
44	**1.51 (0.42)**	**1.42 (0.23)**	**1.92 (0.19)**	**2.38 (0.16)**	2.06 (0.11)
45	0.02 (0.18)	0.30 (0.09)	0.26 (0.07)	0.33 (0.06)	0.34 (0.03)
46	1.01 (0.14)	0.74 (0.09)	0.73 (0.09)	1.20 (0.12)	0.89 (0.04)
47	0.68 (0.08)	0.48 (0.07)	0.40 (0.09)	0.75 (0.12)	0.54 (0.03)
48	0.01 (0.16)	0.45 (0.10)	0.49 (0.09)	0.69 (0.09)	0.64 (0.04)
49	0.01 (0.17)	0.41 (0.12)	0.57 (0.09)	1.04 (0.09)	0.74 (0.04)
50	0.04 (0.17)	0.33 (0.11)	0.50 (0.09)	1.18 (0.09)	0.70 (0.04)
51	**2.11 (0.59)**	**1.23 (0.25)**	**1.86 (0.21)**	**2.18 (0.15)**	2.00 (0.09)
52	0.51 (0.06)	0.54 (0.09)	0.28 (0.11)	0.56 (0.16)	0.48 (0.03)
53	0.58 (0.07)	0.28 (0.11)	0.34 (0.15)	0.95 (0.25)	0.48 (0.04)
Mean	0.62	0.66	0.74	1.20	0.91

Note. Boldface type indicates significant CBD parameter estimates according to the Wald test ($p < .05$).

Table 9

Intersection Parameters Produced by the NRM for the Items in Factors 1-5

Item	Intersection Parameters 1 (*SE*)	2 (*SE*)	3 (*SE*)	4 (*SE*)
Factor 1				
1	-3.77 (0.68)	-0.72 (0.10)	-0.39 (0.07)	1.58 (0.14)
2	-3.65 (0.67)	-1.61 (0.15)	-0.64 (0.07)	0.95 (0.09)
3	**2.18 (0.16)**	**4.90 (2.80)**	**1.13 (0.62)**	**2.85 (1.26)**
4	**-1.06 (0.16)**	**1.35 (0.20)**	**0.20 (0.41)**	**0.36 (0.22)**
5	-1.75 (0.18)	-0.14 (0.10)	-0.07 (0.09)	1.04 (0.10)
6	**41.00 (1.23)**	**17.20 (35.09)**	**-1.82 (0.42)**	**-2.41 (0.60)**
7	**-4.83 (2.48)**	**-1.73 (0.33)**	**-2.06 (0.28)**	**-0.94 (0.09)**
8	**-7.40 (4.97)**	**-1.15 (0.12)**	**-1.30 (0.26)**	**-0.12 (0.08)**
9	**1.26 (0.16)**	**1.53 (0.18)**	**0.97 (0.19)**	**2.01 (0.20)**
10	-0.12 (0.07)	0.88 (0.10)	0.94 (0.10)	1.53 (0.10)
11	**0.42 (0.36)**	**3.20 (1.26)**	**0.00 (0.38)**	**1.86 (0.57)**
Factor 2				
12	-45.60 (33.65)	-0.96 (0.08)	-0.17 (0.07)	0.86 (0.05)
13	-55.50 (2.97)	-1.05 (0.13)	-0.51 (0.08)	0.82 (0.04)
14	**-1.92 (0.22)**	**1.02 (0.19)**	**0.37 (0.10)**	**1.26 (0.07)**
15	-6.19 (3.97)	-1.14 (0.14)	-0.87 (0.15)	0.10 (0.05)
16	-3.19 (0.47)	-0.57 (0.06)	0.02 (0.05)	1.07 (0.05)
17	-2.55 (0.31)	-0.65 (0.09)	-0.28 (0.07)	0.66 (0.05)
18	-2.44 (0.41)	-0.36 (0.09)	-0.07 (0.09)	0.38 (0.06)
19	**-2.02 (0.39)**	**-0.20 (0.14)**	**-1.22 (0.39)**	**0.14 (0.06)**
20	-2.06 (0.21)	-0.40 (0.08)	-0.25 (0.11)	0.62 (0.06)
21	-2.47 (0.21)	-0.53 (0.05)	0.32 (0.05)	1.48 (0.08)
Factor 3				
22	-3.48 (0.75)	-1.79 (0.25)	0.37 (0.14)	2.10 (0.50)
23	-3.05 (0.55)	-0.85 (0.12)	-0.47 (0.12)	1.76 (0.36)
24	**-1.85 (0.21)**	**-0.04 (0.10)**	**-0.11 (0.07)**	**0.86 (0.07)**
27	**0.48 (0.27)**	**1.45 (0.79)**	**-11.67 (21.92)**	**0.48 (0.22)**
28	-0.54 (0.05)	0.47 (0.06)	0.83 (0.06)	1.45 (0.09)
29	-0.97 (0.10)	0.27 (0.09)	0.38 (.17)	1.48 (0.26)

(table continued)

Table 9 (continued)

Item	Intersection Parameters 1 (*SE*)	2 (*SE*)	3 (*SE*)	4 (*SE*)
Factor 4				
25	-2.76 (0.34)	-0.73 (0.08)	-0.28 (0.10)	1.25 (0.14)
26	-2.24 (0.20)	-0.77 (0.07)	-0.44 (0.07)	1.09 (0.07)
30	-11.00 (16.61)	-0.81 (0.11)	-0.67 (0.16)	0.20 (0.07)
31	-0.91 (0.34)	-0.30 (0.12)	0.40 (0.15)	0.67 (0.08)
32	-2.33 (0.15)	-1.51 (0.18)	-1.35 (0.14)	-0.32 (0.06)
33	**-3.43 (0.97)**	**-0.93 (0.16)**	**-1.30 (0.20)**	**-0.39 (0.07)**
34	-1.83 (0.13)	-0.57 (0.07)	-0.24 (0.09)	0.91 (0.07)
35	-3.31 (0.52)	-1.12 (0.09)	0.35 (0.07)	3.82 (0.06)
36	**-0.14 (0.08)**	**0.93 (0.15)**	**0.97 (0.15)**	**1.27 (0.11)**
37	-1.65 (0.11)	-0.43 (0.05)	-0.06 (0.06)	0.78 (0.05)
38	-1.99 (0.18)	-0.60 (0.06)	-0.19 (0.06)	0.71 (0.05)
39	-7.75 (4.06)	-1.38 (0.16)	-0.31 (0.15)	2.69 (0.41)
40	**2.67 (2.54)**	**1.88 (0.83)**	**-10.75 (18.28)**	**2.14 (0.76)**
Factor 5				
41	-2.60 (0.287)	-0.48 (0.105)	-0.08 (0.101)	0.96 (0.076)
42	-5.54 (1.779)	-1.27 (0.125)	-0.93 (0.110)	0.63 (0.044)
43	-7.45 (4.881)	-1.44 (0.133)	-1.31 (0.158)	0.52 (0.390)
44	-3.30 (0.341)	-1.73 (0.116)	-1.13 (0.070)	0.04 (0.034)
45	-76.50 (4.977)	-2.77 (0.626)	-2.38 (0.670)	0.91 (0.112)
46	-2.41 (0.213)	-0.80 (0.106)	0.05 (0.081)	1.16 (0.073)
47	-1.66 (0.172)	0.65 (0.150)	1.12 (0.217)	1.67 (0.181)
48	-129.00 (6.109)	-1.18 (0.323)	-1.27 (0.207)	-0.09 (0.060)
49	-103.00 (561.41)	-2.15 (0.485)	-1.65 (0.212)	-0.17 (0.055)
50	-31.50 (391.83)	-2.06 (0.475)	-1.58 (0.237)	-0.19 (0.054)
51	-2.93 (0.218)	-1.89 (0.150)	-1.28 (0.076)	-0.24 (0.037)
52	**0.45 (0.119)**	**2.07 (0.308)**	**1.36 (0.436)**	**2.32 (0.436)**
53	**1.84 (0.207)**	**3.89 (1.393)**	**2.47 (0.800)**	**1.92 (0.250)**

Note. Boldface type indicates items with empirically disordered response category intersections.

Table 10

Number of Recoded Response Options and Proposed Rescoring Configurations

Item and Factor #	Number of Recoded Response Options				Proposed Rescoring Structure
	5	4	3	2	
Factor 1					
1	X				NA
2	X				NA
3				X	(0,1,1,1,1)
4		X			(0,1,2,2,3)
5		X			(0,1,2,2,3)
6				X	(0,1,1,1,1)
7			X		(0,0,0,1,2)
8			X		(0,0,1,1,2)
9			X		(0,0,1,1,2)
10			X		(0,0,1,1,2)
11			X		(0,0,1,1,2)
Subtotal	2	2	5	2	
Factor 2					
12		X			(0,0,1,2,3)
13		X			(0,0,1,2,3)
14				X	(0,0,0,0,1)
15			X		(0,0,1,1,2)
16	X				NA
17	X				NA
18	X				NA
19		X			(0,1,2,2,3)
20		X			(0,1,2,2,3)
21				X	(0,0,1,1,1)
Subtotal	3	4	1	2	
Factor 3					
22	X				NA
23	X				NA
24			X		(0,0,0,1,2)
27			X		(0,1,1,1,2)
28	X				NA
29	X				NA
Subtotal	4	0	2	0	

(table continued)

Table 10 (continued)

Item and Factor #	Number of Recoded Response Options: 5	4	3	2	Proposed Rescoring Structure
Factor 4					
25	X				NA
26	X				NA
30			X		(0,0,1,1,2)
31			X		(0,0,1,1,2)
32				X	(0,1,1,1,1)
33			X		(0,0,1,1,2)
34			X		(0,0,0,1,2)
35	X				NA
36			X		(0,0,1,1,2)
37			X		(0,0,1,1,2)
38			X		(0,0,1,1,2)
39		X			(0,0,1,2,3)
40				X	(0,0,0,0,1)
Subtotal	3	1	7	2	
Factor 5					
41	X				NA
42				X	(0,0,0,0,1)
43				X	(0,0,0,0,1)
44				X	(0,0,0,0,1)
45				X	(0,0,0,0,1)
46		X			(0,1,1,2,3)
47	X				NA
48			X		(0,0,0,1,2)
49			X		(0,0,0,1,2)
50			X		(0,0,0,1,2)
51				X	(0,0,0,0,1)
52		X			(0,1,2,2,3)
53			X		(0,1,1,1,2)
Subtotal	2	2	4	5	
Total	14	9	19	11	

Note. The original five-category scoring structure was (0,1,2,3,4). NA indicates items that were not rescored and retained their original scoring structure. X indicates the final number of response options retained by each item at the conclusion of the analysis.

However, in addition to examining CBD and intersection parameters, all CRFs were visually inspected for evidence of poorly functioning response options. These visual inspections revealed that, for some items, the lowest category, 0 (*Never*), did not function adequately over a practical range of theta. In particular, several items in Factor 5 (Emotional Arousal) did not seem to support the lowest response option. When this occurred, the bottom two categories were collapsed to improve functioning at the low end of the theta scale. Collapsing of categories continued until CBD parameters were non-significant and intersections were ordered, or the item became dichotomous. Item 48 (Figure 1) is illustrative of this phenomenon, as evidenced by the fact that the 0 (*Never*) remains below the adjacent category 1 (*Occasionally/rarely*) until they intersect at -129.00 on the theta scale, or 129 standard deviations below the mean (see Table 9). A total of 11 items exhibited similar problems with the lowest response option, and were subsequently rescored, including:

- Factor 2: items 12, 13, 15;
- Factor 4: items 30 and 39; and
- Factor 5: items 42, 43, 45, 48, 49, and 50.

Graphs for the remaining items that exhibited this phenomenon, but improved through rescoring, are found in Appendix D.

In addition, the visual inspection of CRFs revealed three items with a poorly functioning middle option. For example, CRFs for Item 5 (Figure 2) showed that the middle option was relatively unutilized, and never had the highest likelihood of being endorsed. These items were rescored by collapsing the underperforming middle option into the adjacent option with the lowest category information, as revealed by the category information curves. The other two items that had underperforming middle options included Factor 1 item 10 and Factor 2 item 20. Graphs for these items are found in Appendix E.

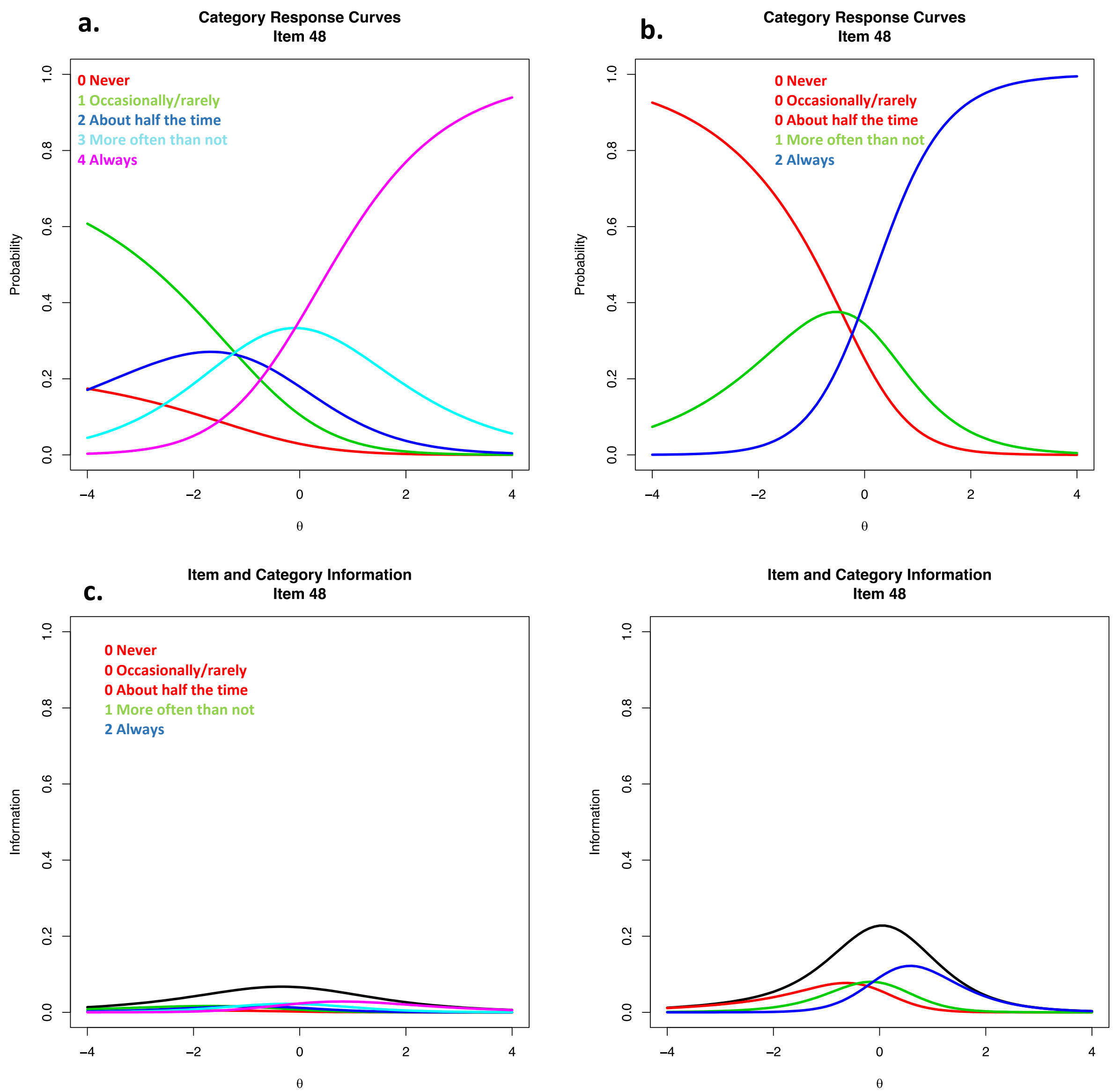


Figure 1. Factor 5 Item 48: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

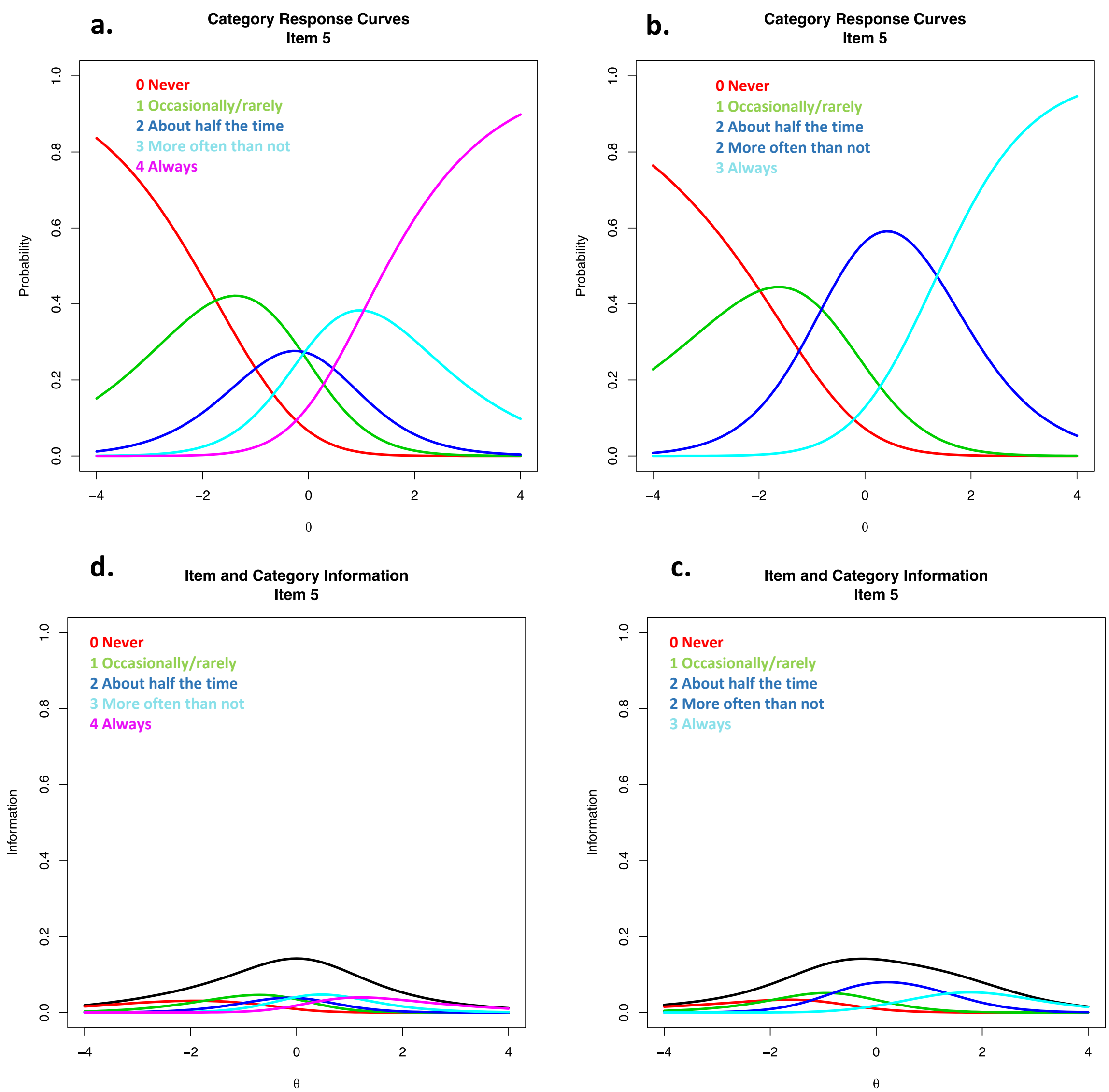


Figure 2. Factor 1 Item 5: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

Because IRT parameters are estimated in the presence of all other items in the model, the parameter estimates for some items changed as the other items in the model were rescored. For example, there were 5 items that initially passed all tests, but experienced problems (i.e., significant CBD parameters, or intersection disordering) in subsequent iterations of parameter estimation as other items were rescored. Figures for these items are found in Appendix F. The majority were in Factor 4 and included:

- Factor 4 items 31, 34, 37, 38; and
- Factor 5 item 46.

Finally, my analysis also revealed 8 poorly functioning items whose discriminating power and information did not improve after rescoring, including:

- Factor 1 items 6 and 11;
- Factor 3 items 22, 23, and 27;
- Factor 4 items 39 and 40; and
- Factor 5 item 45.

These items individually exhibited significant CBD parameters, disordered intersections, or underperforming response options in their original format and generally contributed very little psychometric information. Post hoc rescoring methods did not meaningfully improve their functioning. Preston and Reise (2014a) suggest that when item functioning cannot be improved using post hoc rescoring methods, “a researcher may want to consider revising the item or rethinking the content used to anchor the response category” (p. 392). That is, such items are likely to have been inadequately conceived or poorly written to begin with and should not be expected to improve through simply combining some response categories. These items need to be thoughtfully checked against the definition of the construct that they are intended to measure to ensure that they are indeed capturing the correct essence. Any of these particular items that are deemed to be necessary from the standpoint of content validity ought to be examined in order to improve the substantive meaning of the item. Such problematic items may even be candidates for removal unless the item makes a unique contribution to the construct validity of the subscale in the event that the factor already has a sufficient number of items. In this regard, Factor 1 item 6 (Figure 3) is an example. It functioned very poorly initially, and did not improve through

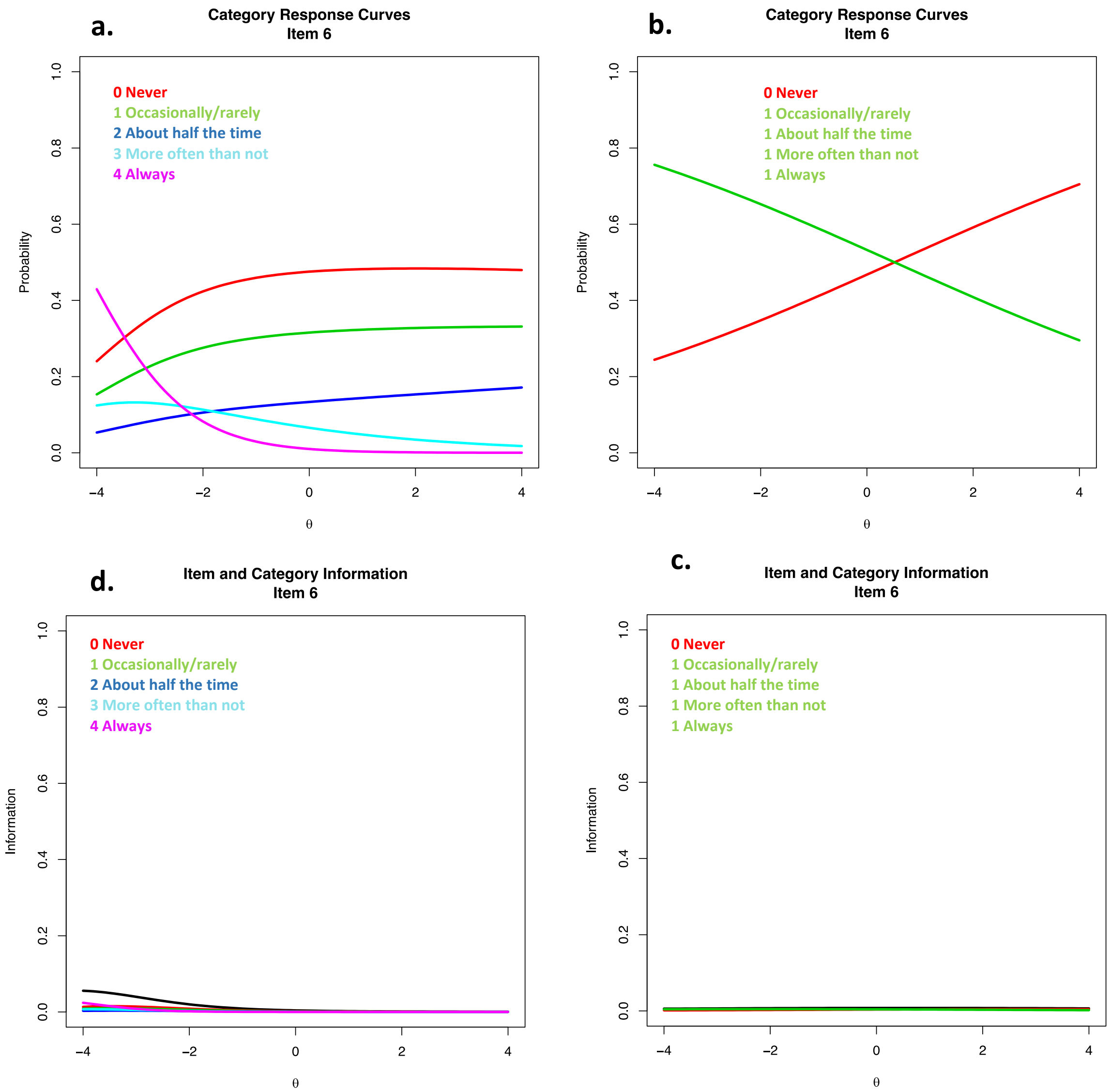


Figure 3. Factor 1 Item 6: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

rescoring. Information was also virtually nonexistent, before and after rescoring. Graphs for the remainder of these items are found in Appendix G.

After completing the iterative process of estimating, rescoring, re-estimating, and performing visual inspections, I found that 12 (23%) of the original 53 items functioned properly with the original five response categories and did not need to be rescored. However, while these items functioned properly in terms of discriminating power and category ordering, nine (75%) of them (items 1, 2, 22, 23, 25, 29, 35, 41, and 47) contributed relatively little in terms of information. Item 22 (Figure 4) is illustrative of this point, because it functions acceptably in terms of discrimination and category ordering, but has an item information curve that peaks well below 0.2. Unfortunately, scholars have not reached a consensus regarding what constitutes a sufficient amount of psychometric item information. Some have suggested that 0.2 is a relevant cutoff for item information, and that items that peak below 0.2 are candidates for modification or removal (Preston et al., 2015; Ura, Preston, & Mearns, 2015). In contrast, García-Pérez (2017) seems to suggest, somewhat indirectly, that any information is valuable and should be maintained. This debate is revisited briefly in the discussion section. Figures for the remaining items that functioned appropriately are found in Appendix H. These included:

- Factor 1 items 1, 2;
- Factor 2 items 16, 17, 18;
- Factor 3 items 28, 29;
- Factor 4 items 25, 26, 35; and
- Factor 5 items 41, 47.

To assess the overall impact of rescoring, several additional analyses were performed. First, I estimated and plotted test information functions (TIFs) for each of the five factors before and after rescoring, and found that overall test information increased, at least somewhat, in each case (Figures 5 through 9). Factor 3 (Avoidance; Figure 7) experienced the smallest increase in relative test information across levels of the trait. Factor 5 (Emotional arousal; Figure 9) experienced a slight decrease in test information at lower levels of theta, but an increase in test information at higher levels of the trait as a result of the proposed scoring modifications.

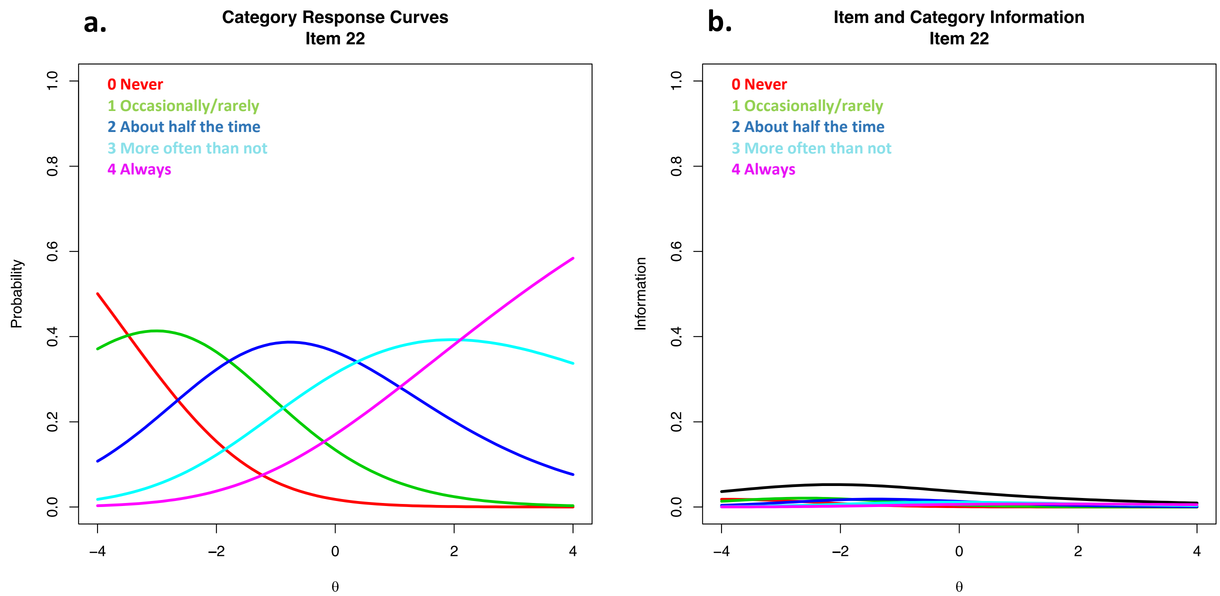


Figure 4. Factor 3 Item 22: Category Response Curves (a) and Item Category Information Functions (b).

Overall, the rescoring procedure appears to have increased relative test information across each of the five factors.

In addition to examining TIFs for each factor, item and category information curves (IICs and CICs) were estimated for each item before and after rescoring. These graphs appear within the corresponding figures for each individual item. In most cases, item information did improve at least somewhat.

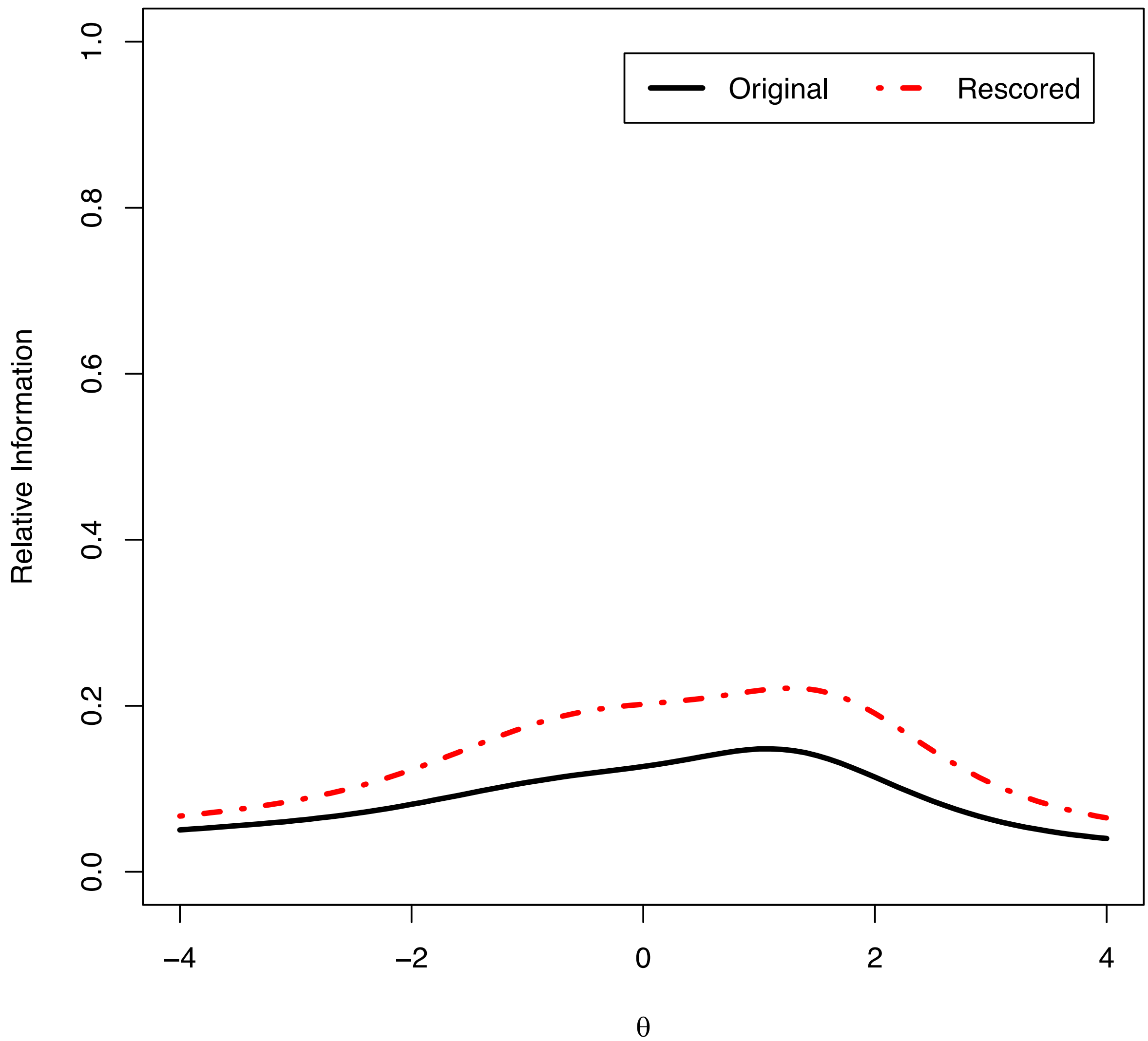


Figure 5. Original and rescored Test Information Functions for Factor 1.

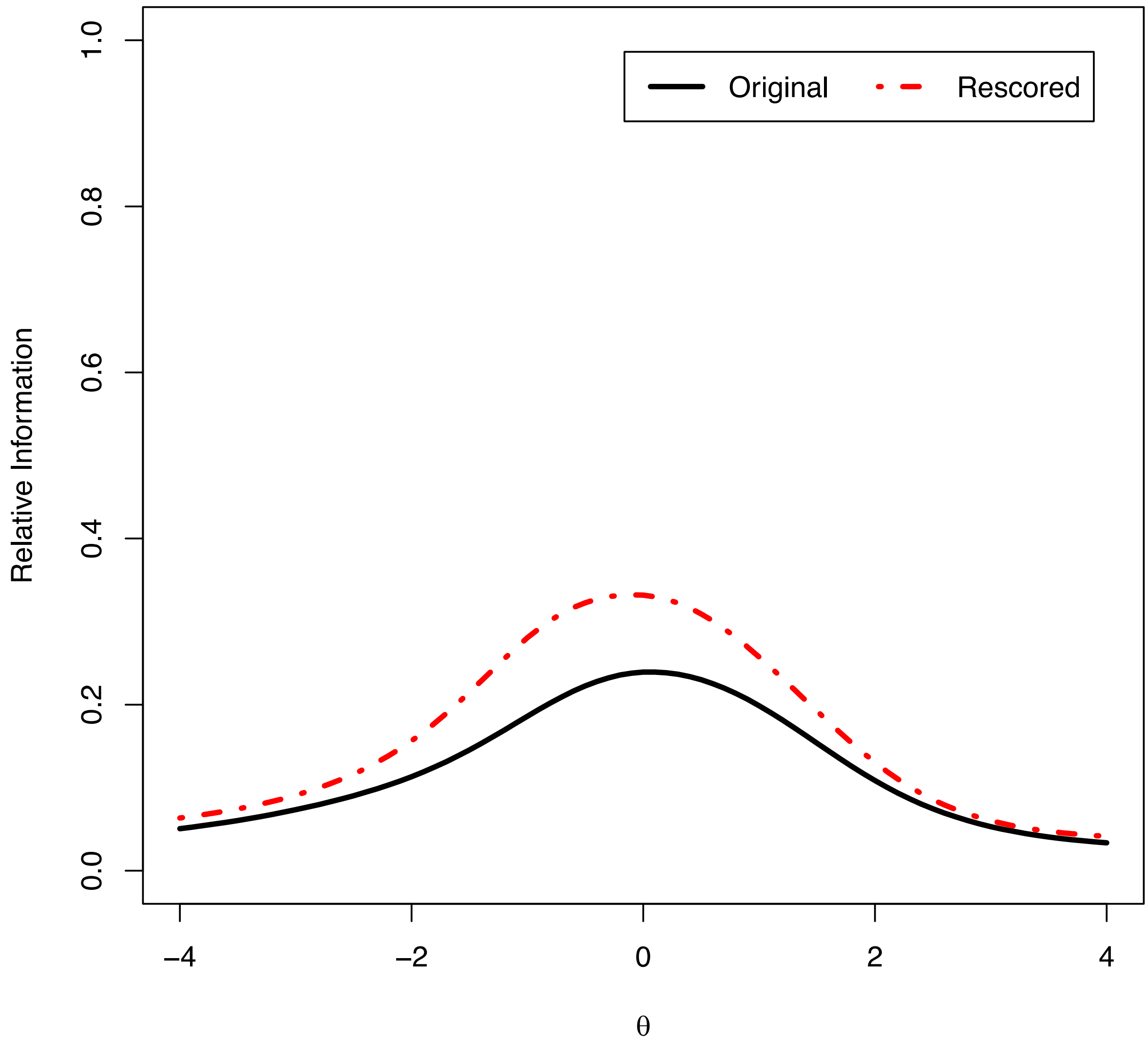


Figure 6. Original and rescored Test Information Functions for Factor 2.

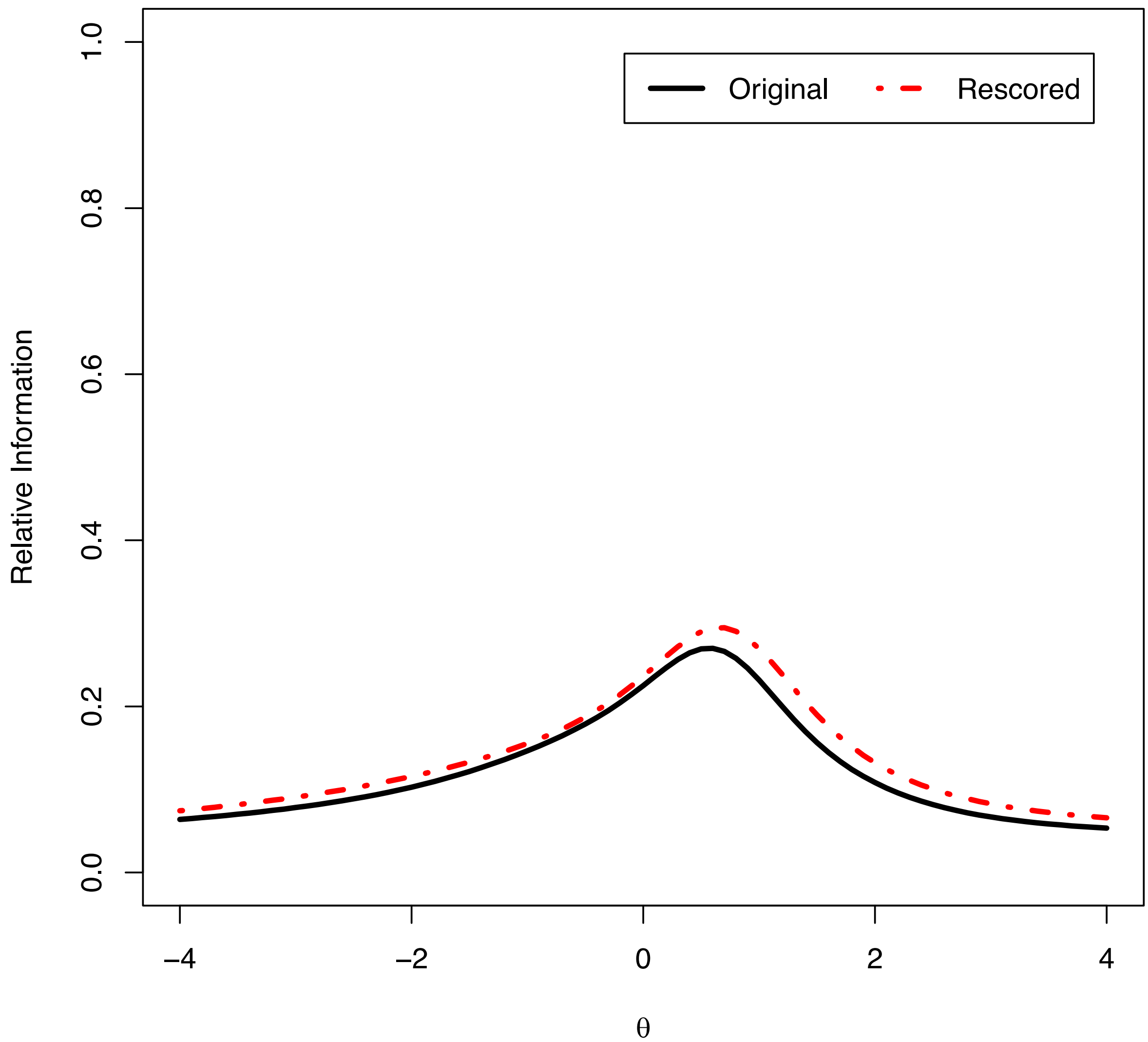


Figure 7. Original and rescored Test Information Functions for Factor 3.

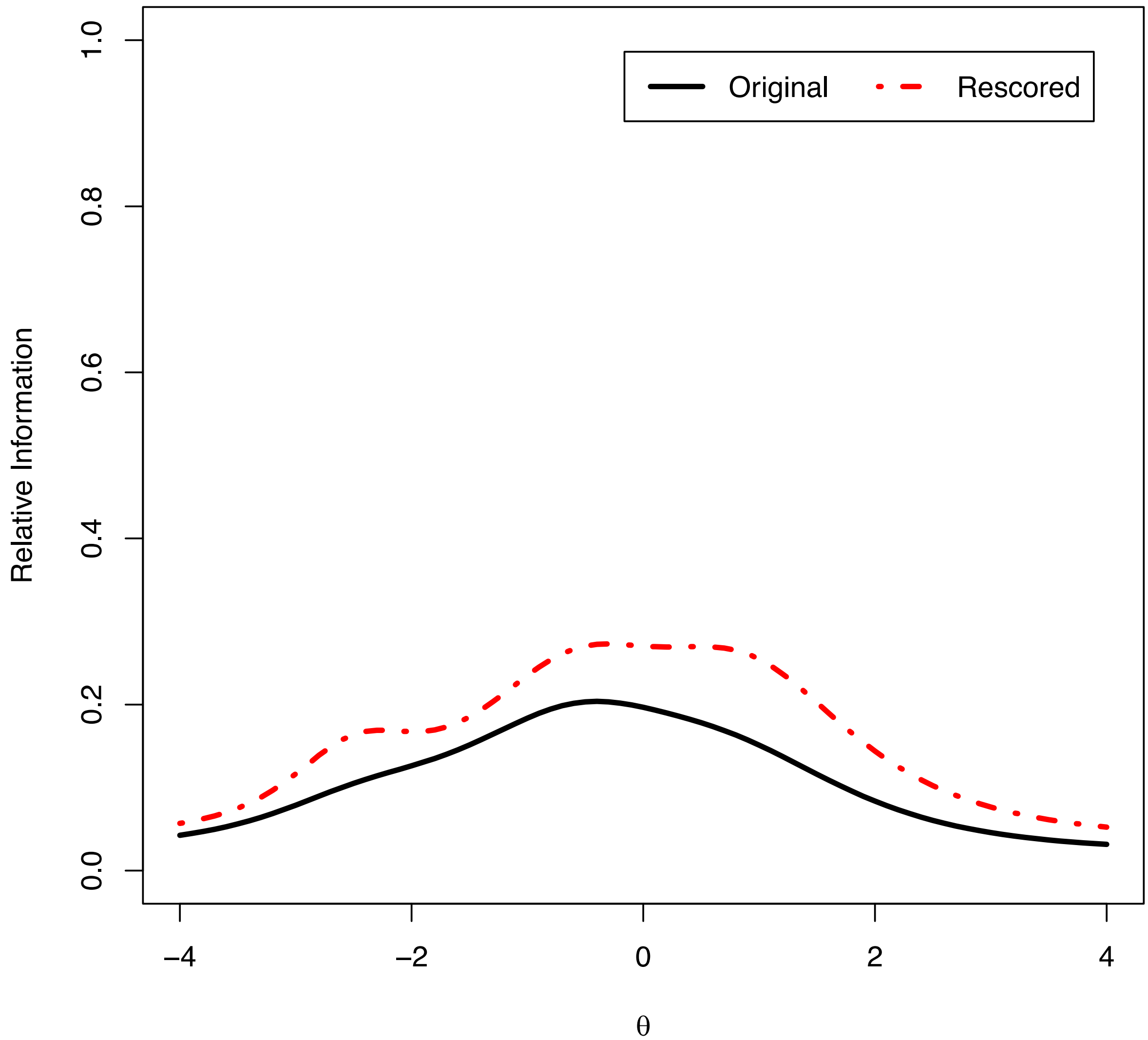


Figure 8. Original and rescored Test Information Functions for Factor 4.

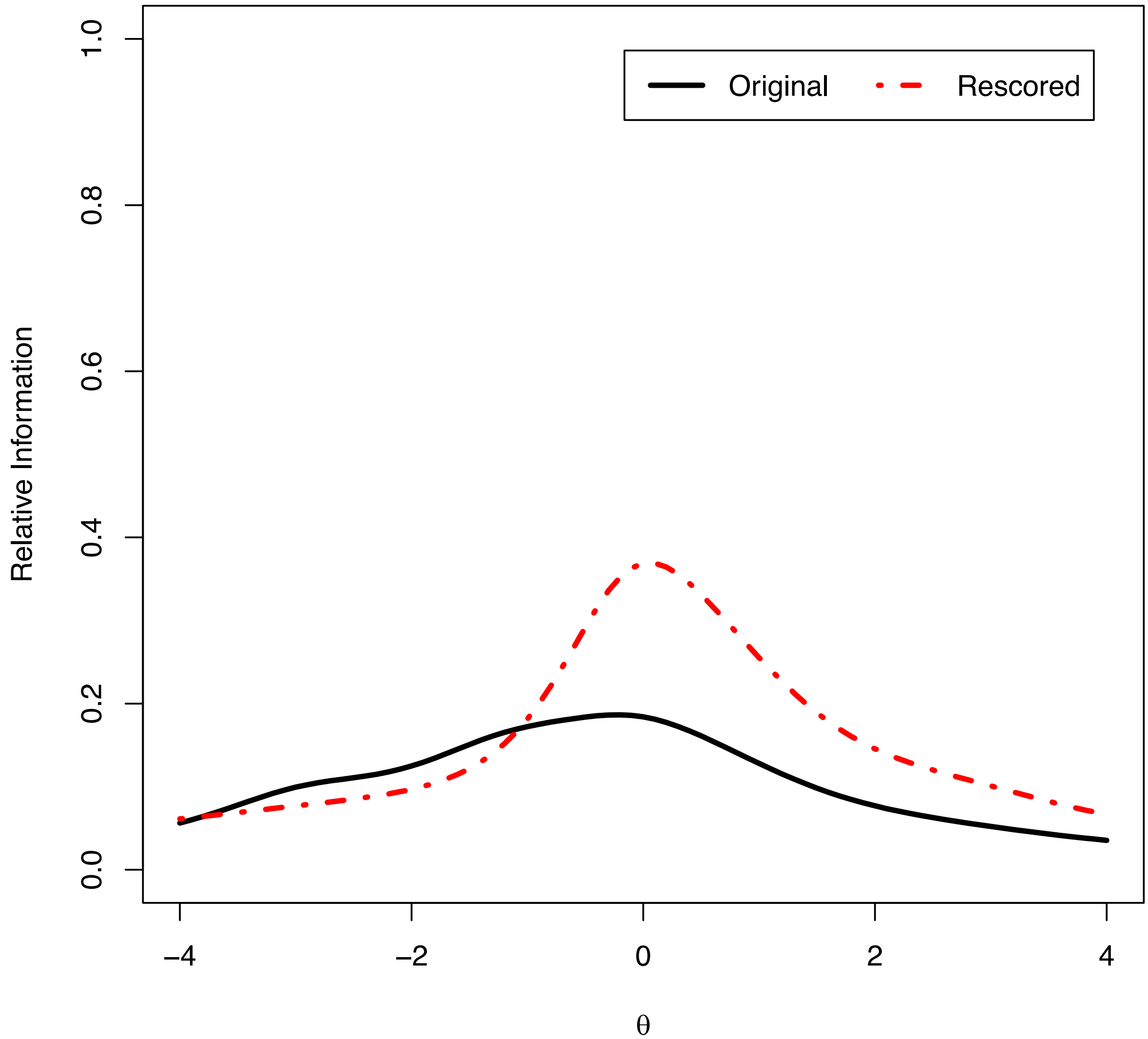


Figure 9. Original and rescored Test Information Functions for Factor 5.

Third, in an effort to assess the impact of rescoring on so-called nuisance or error variance, I used CFA factor loadings to estimate three separate reliability coefficients for each factor before and after rescoring (Table 11).

Because reliability is conceptualized as the ratio of true score variance to total observed variance, higher reliability estimates indicate lower levels of nuisance or error variance. With the exception of Raykov's rho estimates for Factor 5, the rescoring procedure appears to have decreased nuisance variance and the increased reliability of total test scores. The rho estimates are likely to be the most appropriate, given the fact that the assumption of tau-equivalence (i.e., equal factor loadings) required by coefficient alpha is violated in this case as discussed previously.

Table 11

Three Separate Measures of Reliability Before and After Rescoring

		Estimated Reliability Coefficients			Fit Statistics			Number of Items	Number of Correlated Error Pairs
Factor	Scale	Alpha	Omega	Rho	RMSEA	CFI	TLI		
1	Original	.762	.781	.707	.064	.961	.946	11	4
1	Rescored	**.786**	**.803**	**.746**	.056	.965	.953	11	3
2	Original	.888	.889	.839	.082	.973	.961	10	4
2	Rescored	**.900**	**.901**	**.862**	.075	.975	.965	10	3
3	Original	.687	.695	.627	.066	.988	.974	6	2
3	Rescored	**.691**	**.698**	**.631**	.063	.988	.974	6	2
4	Original	.867	.873	.806	.081	.964	.952	13	6
4	Rescored	**.880**	**.887**	**.828**	.066	.972	.964	13	5
5	Original	.866	.871	**.799**	.061	.984	.979	13	7
5	Rescored	**.873**	**.879**	.788	.064	.978	.971	13	6

Note. Boldface type indicates a relatively larger estimate of reliability.

Omega does not assume tau-equivalence, but it does assume that the error variances are uncorrelated, which was also violated. Raykov's rho does not assume tau-equivalence and it explicitly accounts for correlated error terms, making it the most appropriate estimate of reliability in this particular case. All three reliability coefficients, taken together with the appropriate caveats, give us a more complete picture of the reliability of each factor and whether or not rescoring reduced nuisance variation.

Going one step further, it is possible to estimate the amount of error variance as a percentage of total variance, before and after rescoring, using the formula

$$\sigma_e^2 = \sigma_x^2(1 - \rho) \tag{4.0}$$

where σ_e^2 is the estimated error variance, σ_x^2 is the total observed variance, and ρ is an estimate of reliability. This equation assumes that the error terms for each item are uncorrelated, once again making Raykov's rho the most appropriate measure of reliability. In addition, I consulted the Mplus CFA modification indices for evidence of correlated errors, and included "with" statements to account for those correlations. Correlated errors with the largest modification index were correlated in succession up until the point that sufficient CFA model fit was achieved.

After estimating the error variances for each factor before and after rescoring in this manner, I compared the results and found that the rescoring procedure produced a small decrease in error variance in Factors 1 through 4 and increased it only slightly in Factor 5, which is consistent with the findings already discussed (Table 12). The apparent reduction in estimated error variance in Factors 1 through 4 as a result of rescoring is akin to a reduction in the "nuisance variation" referred to by Preston and Reise (2014a, p. 396).

Fourth, test characteristic curves (TCCs) were estimated for each factor before and after rescoring to determine what impact, if any, the rescoring process had on relative score outcomes. These figures are found in Appendix J along with scatterplots of original versus rescored outcomes. Following the pattern outlined by de Ayala (2009), I estimated TCCs using the proportion of total points possible, as opposed to a simple raw sum score, so that the original and rescored TCCs could be represented on the same metric.

Table 12

Estimated Reliability and Error Variance by Factor and Scale

Factor	Scale	Mean	Total Variance	Raykov's Reliability Coefficient	Estimated Error Variance	Error Variance as a Percent of Total Variance	Reduction in Error Variance
1	Original	19.69	41.41	.707	12.13	29.3%	
	Rescored	12.71	19.23	.746	4.89	25.4%	3.9%
2	Original	25.14	66.20	.839	10.66	16.1%	
	Rescored	17.12	37.69	.862	5.20	13.8%	2.3%
3	Original	12.17	23.03	.627	8.59	37.3%	
	Rescored	9.68	16.45	.631	6.07	36.9%	0.4%
4	Original	29.70	87.86	.806	17.05	19.4%	
	Rescored	15.66	34.08	.828	5.86	17.2%	2.2%
5	Original	32.57	67.27	.799	13.52	20.1%	
	Rescored	12.64	30.72	.788	6.51	21.2%	-1.1%

TCC curves can be interpreted as the expected test score (or in this case, the expected proportion correct) for persons having various levels of trauma. Ideally, rescoring would reduce the expected total score by the same amount at each level of the trait, and the proportion of total possible would be roughly equal before and after rescoring. In addition, scatterplots of original and rescored sum scores should be tightly clustered in a straight and positively sloped linear fashion. This would indicate that the rescoring procedure did not have a differential impact on test score outcomes for individuals with high versus low levels of the trait, and would provide evidence that rescoring did not impact the relative standing of individuals on the trait continuum.

I found that rescoring did reduce the expected total test score by roughly the same amount over increasing levels of theta, and that the expected proportions of total correct for each factor were approximately equal. Factor 5 (Figure 10), however, experienced the largest discrepancy between original and rescored items as an expected proportion of total possible, particularly at low levels of trauma. The scatterplot of original and rescored sum scores is also slightly curved, showing that the relationship between them is not entirely linear. This seems to indicate that, at least for Factor 5, rescoring caused individuals at lower trait levels to receive relatively lower scores. Overall, the fact that TCCs appear to be roughly equivalent before and after rescoring for Factors 1 through 4 indicates that rescoring had only a minor impact on the relative ranking of individuals' final test scores in each factor. Scatterplots of rescored test scores on original test scores corroborated these findings, and clearly illustrated that the two sets of scores were highly linearly related.

Finally, because the data are categorical rather than continuous, I calculated parametric and non-parametric correlation coefficients to determine the extent to which rescoring impacted the relative ranking of test score outcomes (Table 13). All three sets of correlation coefficients, including Kendall's tau which accounts for tied ranks, revealed original and rescored sum scores to be significantly and highly correlated, indicating that rescoring had only minor and insignificant impact on the relative ranking of individuals on each factor.

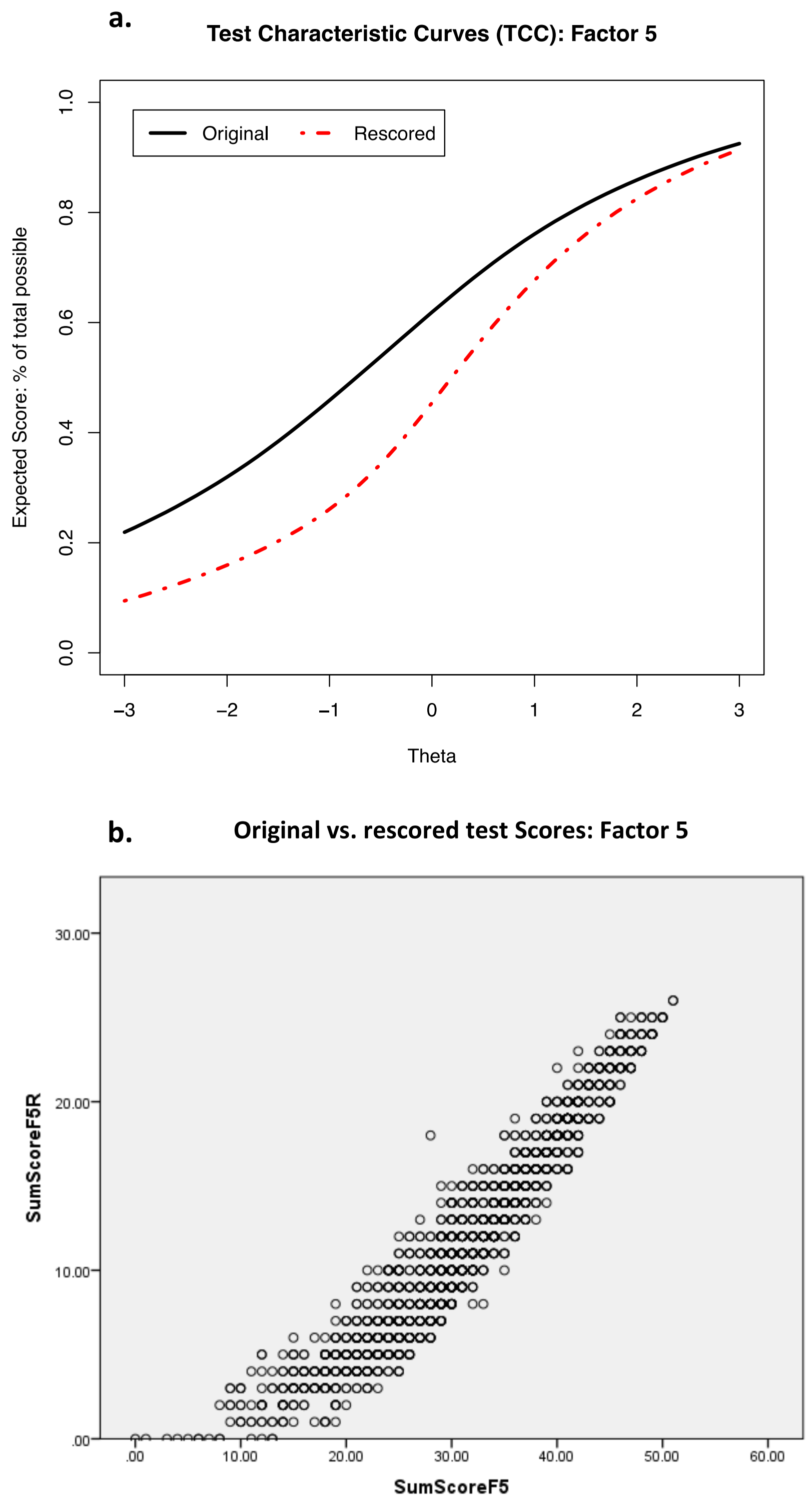


Figure 10. Rescored vs. original Test Characteristic Curves (a), and scatterplot of original vs. rescored test scores (b) for Factor 5.

Table 13

Correlations Between Original and Rescored Scores by Factor

	Correlations Coefficients		
Factor	Pearson	Kendall's tau	Spearman's rho
1	.969**	.884**	.968**
2	.986**	.923**	.984**
3	.976**	.904**	.973**
4	.968**	.873**	.967**
5	.957**	.887**	.970**

Note. ** Significant at the .01 level (2-tailed).

Assessing Data Normality

Finally, it is important to point out that the distribution of responses to several items was highly skewed (Table 14). Some were positively skewed (e.g., Items 3, 6, 9, 52, and 53) while others were negatively skewed (e.g., Items 7, 32, 44, 49, and 51). Some of this skewness may be due to the overly strong, or overly vague, wording of the item stems. For example, all of the items identified as highly positively skewed deal with the threat or probability of physical violence in one form or another (see Appendix A), which is likely to exist primarily in extreme cases. In contrast, highly negatively skewed items are generally vague and are perhaps only tangentially related to their associated factors (see Appendix A). In any event, several items do exhibit evidence of skewness.

In addition, there is reason to believe that trauma, like many other psychological constructs, is not normally distributed in the population. Thus, the fact that traditional IRT methods were used, including standard MML and EM estimation algorithms, perhaps introduced bias into the estimates (Preston & Reise, 2014b). The implications of skewness for this study and for future research are covered in the discussion section.

Table 14

Response Option Frequency, Mean Item Scores, and Skewness by Item

		Response options							
Factor	Item	Never	Occasionally / Rarely	About half the time	More often than not	Always	Mean	S.D.	Skewness
1	1	2.8	20.8	23.7	35.9	16.8	2.43	1.079	-0.250
	2	1.3	11.2	23.6	38.7	25.2	2.75	0.996	-0.486
	3	**86.2**	**7.9**	**2.3**	**2.3**	**1.3**	**0.25**	**0.729**	**3.440**
	4	24.3	33.1	13.4	14.0	15.2	1.63	1.382	0.483
	5	11.3	23.6	21.2	25.3	18.5	2.16	1.287	-0.098
	6[R]	**46.9**	**31.0**	**13.1**	**6.9**	**2.0**	**0.86**	**1.020**	**1.129**
	7	**1.4**	**5.8**	**8.3**	**23.2**	**61.3**	**3.37**	**0.955**	**-1.579**
	8	2.4	11.7	17.6	29.4	38.9	2.91	1.114	-0.743
	9	**59.9**	**21.9**	**7.5**	**7.2**	**3.5**	**0.72**	**1.095**	**1.527**
	10	37.9	31.3	13.5	10.5	6.8	1.17	1.233	0.883
	11	31.5	28.2	14.7	15.3	10.2	1.45	1.340	0.551
2	12	1.5	14.1	24.6	31.6	28.2	2.71	1.069	-0.387
	13	3.9	11.9	19.9	32.5	31.7	2.76	1.136	-0.658
	14	13.8	33.8	19.5	19.5	13.3	1.85	1.263	0.276
	15	2.4	11.9	17.5	27.2	41.0	2.93	1.126	-0.751
	16	4.0	20.3	25.2	30.1	20.4	2.42	1.140	-0.221
	17	4.8	17.4	20.5	28.8	28.6	2.59	1.203	-0.431
	18	7.8	19.4	19.8	22.8	30.3	2.48	1.308	-0.341
	19	8.7	16.3	15.4	24.2	35.2	2.61	1.340	-0.543
	20	6.8	18.8	20.4	26.7	27.3	2.49	1.257	-0.361
	21	3.8	21.6	29.7	29.5	15.4	2.31	1.087	-0.096

(table continued)

Table 14 (continued)

Factor	Item	Never	Occasionally / Rarely	About half the time	More often than not	Always	Mean	S.D.	Skewness
		Response options							
3	22	3.3	15.1	33.8	29.8	18.1	2.44	1.051	-0.194
	23	4.8	17.9	24	32.7	20.6	2.47	1.143	-0.341
	24	11.4	24.8	19.2	24.5	20.0	2.17	1.313	-0.085
	27	24.8	19.8	14.7	21.1	19.5	1.91	1.475	0.062
	28	30.5	31.9	16.2	12.6	8.8	1.37	1.275	0.665
	29	19.3	26.3	20.5	20.2	13.7	1.83	1.324	0.176
4	25	5.1	19.0	25.1	31.4	19.5	2.41	1.148	-0.280
	26	5.7	17.1	22.2	35.2	19.8	2.46	1.152	-0.414
	30	8.3	15.3	19.4	26.1	30.9	2.56	1.292	-0.495
	31	19.5	21.4	22.0	19.4	17.8	1.95	1.377	0.058
	32	**1.9**	**7.8**	**12.8**	**28.9**	**48.6**	**3.15**	**1.038**	**-1.124**
	33	4.8	11.6	13.5	24.8	45.3	2.94	1.216	-0.913
	34	7.9	19.2	23.0	29.6	20.3	2.35	1.222	-0.286
	35^R	4.1	18.2	38.8	35.1	3.7	2.16	0.905	-0.353
	36	33.9	28.8	16.4	12.0	8.8	1.33	1.292	0.688
	37	9.9	21.0	22.0	26.0	21.2	2.28	1.279	-0.202
	38	8.2	18.8	21.7	27.9	23.5	2.4	1.256	-0.316
	39^R	3.0	16.3	31.9	36.4	12.3	2.39	0.996	-0.268
	40	32.3	27.4	19.9	12.9	7.5	1.36	1.259	0.593
5	41	5.6	21.7	24.4	28.0	20.3	2.36	1.186	-0.187
	42	1.3	12.0	19.3	37.3	30.1	2.83	1.030	-0.597
	43	1.5	9.4	17.5	38.5	33.2	2.92	1.007	-0.760
	44	**0.8**	**5.6**	**11.2**	**32.2**	**50.2**	**3.25**	**0.921**	**-1.205**
	45^R	1.9	8.9	19.6	37.9	31.8	2.89	1.013	-0.723
	46	4.7	19.8	28.2	30.0	17.4	2.36	1.119	-0.188
	47	14.4	33.2	24.6	18.3	9.5	1.75	1.187	0.310
	48	2.9	10.6	15.1	28.2	43.2	2.98	1.127	-0.898
	49	**2.7**	**7.4**	**13.4**	**29.1**	**47.5**	**3.11**	**1.063**	**-1.127**
	50	2.7	9.3	14.6	27.5	46.0	3.05	1.103	-0.991
	51	**0.9**	**4.5**	**9.1**	**27.1**	**58.4**	**3.38**	**0.890**	**-1.491**
	52	**41.8**	**31.5**	**12.1**	**10.0**	**4.6**	**1.04**	**1.164**	**1.004**
	53	**62.1**	**22.5**	**8.5**	**4.7**	**2.3**	**0.63**	**0.981**	**1.688**

Note. The superscript R indicates items that were reverse coded. Boldface type indicates items with a skewness statistic greater than an absolute value of 1.

CHAPTER 5: Discussion

Summary of Findings

These analyses revealed that several TIPSA items contained response options that were technically problematic in that they were either empirically disordered or they were indistinguishable across varying levels of trauma causing them to be underutilized. Some of these problems were mitigated through post hoc modifications of the response options (i.e., by collapsing two or more adjacent categories into one). These modifications produced better model fit and improved overall test information across each of the five factors. That is, modifying the scoring structure improved the functioning of many of the items and options, as evidenced by non-significant differences between CBD parameters after rescoring, improved empirical ordering, and increases in the item information curves.

Some items, however, did not improve as a result of these revisions and ought to be evaluated for other problems such as vague or ambiguous wording of the item stem or clear wording describing an extremely infrequent behavior or situation. These items were 6, 11, 22, 23, 27, 39, 40, and 45. In addition, items including response categories that were collapsed to the point of becoming dichotomous ought to be carefully evaluated and perhaps modified, since the need for such scoring modifications could be evidence of content or wording problems. The resulting dichotomous items are likely to be evidence of situations that are too extreme (i.e., the threat of physical violence) or too infrequent (i.e., disturbing dreams) to meaningfully capture trauma in the partners of sex addicts. In other words, the scenarios described in these item stems are not sufficiently sensitive as indicators of trauma. For example, Items 14, 20, 42, 43, 44, 45, 51 (the majority of which are in Factor 5: Emotional Arousal) were rescored as follows:

- 0 (never) to 0
- 1 (occasionally/rarely) to 0
- 2 (*about half the time*) to 0
- 3 (*more often than not*) to 0
- to 4 (*always*) to 1

This modified scoring structure is problematic because the new zero category has no clear meaning other than any occurrence that is less frequent than always. The rescoring of these items renders a response of 3 (*more often than not*) equivalent to a response of 0 (*never*). In other words, these items register trauma only for people with extreme (i.e., *always*) responses and treat all other respondents as if they have no trauma at all. Items that are thus extreme, or that result in such questionable revised scoring structures, likely need to have the item stem revised or reworded. Alternatively, there could be a mismatch between the frequency continuum utilized in the option anchors, and the scenario described in the stem. For example, perhaps the threat of violence is a powerful predictor of trauma, but a frequency continuum from 1 (never) to 5 (always) is inappropriate.

In contrast, after rescoring, Items 3, 6, and 32 registered trauma for every respondent who selected anything above 0 (*never*). These items were recoded as follows:

- 0 (never) to 0
- 1 (occasionally/rarely) to 1
- 2 (*about half the time*) to 1
- 3 (*more often than not*) to 1
- to 4 (*always*) to 1

Thus, the revised scoring structure for items 3, 6, and 32 treats a response of 1 (*occasionally/rarely*) the same as a response of 4 (*always*). While this rescoring did improve technical item functioning, the validity problems associated with treating a response of 1 (*occasionally/rarely*) on the revised scale the same as a response of 4 (*always*) on the original scale should be obvious. This is probably evidence that these items are in need of rewording or other semantic modifications, aside from simply post hoc rescoring.

This study suggests that solving technical psychometric problems through post hoc scoring modifications has little practical impact on score outcomes such as the relative standing of individuals or perturbations due to nuisance variance. This is evidenced by the high correlations between respondents' scores on the original items and their scores on the rescored items for each factor as well as the small increases in test score reliability. Changing the scoring structure in an effort to solve psychometric problems did not seem to have a meaningful impact on relative standing of respondents on the trait continuum, nor does it seem to substantially decrease nuisance variance as evidenced by the reliability statistics. Even with the potential problems discovered in the originally scored items, the initial scores and modified scores of respondents were very highly correlated. This indicates that the initial problems had little impact on final score outcomes, that the modifications done in this study may not be helpful, and that it may be prudent to focus on revising the stems of problematic items rather than wholesale collapsing of response categories. In particular, item stems that are overly extreme, or overly vague, deserve attention.

Regarding technical item functioning, I have suggested that items be rescored, modified, or removed based primarily on a set of psychometric criteria, namely significant CBD parameters and/or category response disordering. Several (e.g., Items 6 and 11) items with these technical problems contributed very little psychometric information, both before and after rescoring. However, there was also a set of items that functioned properly in terms of discrimination and category ordering, but still contributed relatively small amounts of psychometric information (e.g., Items 35, 41, 47; see Appendix H).

While there does not appear to be a consensus in the literature on what constitutes a sufficient amount of item information, some scholars have suggested that any item whose information curve does not peak above 0.2 is a candidate for revision or removal (Preston et al., 2015; Ura et al., 2015). In contrast, García-Pérez (2017) implies that all item information ought to be jealously guarded, writing that there is “no reason to give up the discriminative information that each category provides [by collapsing categories when they are disordered]. Even when some categories are used sparingly by the respondents, they can still make distinctions that matter to practitioners” (p. 23). Ultimately, making decisions based on item information alone is a judgment call that requires the scale developer to strike a delicate balance between test length, content validity, and the relative precision of a particular item across different levels of the trait. Items that are judged to be necessary from a content validity point of view should be maintained. Uninformative items that are redundant or otherwise unnecessary in order to maintain content validity should either be modified or discarded.

Improving and Maintaining Content Validity

Perhaps even more important than technical item functioning, however, is the issue of content validity. Psychological instruments that function perfectly on a technical level are of little value if they do not capture the essence of some important facet of the targeted construct. While the purpose of this paper is primarily to analyze the psychometric properties of the TIPSA, content validity is also an important issue that deserves attention.

From the outset, it is important to recognize that the TIPSA is primarily a screening instrument designed to measure trauma experienced by the partners of sex addicts. It is not intended to diagnose trauma, but rather to measure the degree to which individuals may be experiencing it as a result of their partner's sex addiction(s). As a measurement and screening instrument, the TIPSA represents an important and valuable step forward. Evaluating and perhaps modifying the stems, or incongruent option anchors, of the problematic items identified in this study has the potential to improve the TIPSA's technical psychometric performance.

In contrast, the primary purpose of the DSM criteria is diagnosis rather than psychological measurement (American Psychiatric Association, 2015). While utilizing these criteria as a starting point for item development was a prudent decision, it may have been insufficient from a measurement or construct validity perspective. To fully capture the construct under consideration, a scale developer ought to consider any relevant aspects of the trait that are believed to be impactful. After consulting the literature, the author believes that there are at least two content areas that could be better represented in the TIPSA in order to avoid problems associated with "construct underrepresentation (or construct deficiency)" (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014, p. 12). As research on trauma and sex addiction move forward, additional content areas and nuances may need to be addressed.

First, the issue of cumulative trauma could be made more explicit in the TIPSA (Follette et al., 1996; Schumm et al., 2006). If trauma is indeed cumulative over time as the literature suggests, then that aspect of the trait ought to be made more fully captured in TIPSA items. The troubling finding that women who were sexually abused or traumatized as children are more likely to be re-traumatized later in life (Follette et al., 1996; Schumm et al., 2006) highlights the need to integrate cumulative trauma into the TIPSA. In its' current form, the TIPSA deals with cumulative nature of trauma only briefly and indirectly (e.g., Item 11).

Second, the literature also illustrates the need to consider the degree to which an individual feels personally betrayed because this has been shown to impact the intensity of trauma experienced (Freyd, 1998; Martin et al., 2013). Indeed, the relationship between the trauma survivor and the perpetrator is an "established predictor of trauma-related psychopathology, with interfamilial or interpersonal traumas being associated with more negative psychological outcomes than extrafamilial or noninterpersonal traumas" (Martin et al., 2013). That is, traumatizing experiences between close partners like spouses with children who have been married for an extended period of time are likely to be more traumatic than similar experiences between dating partners who have been together only a short while. This aspect of trauma should not be overlooked in the TIPSA.

If the purpose of the TIPSA is to measure trauma, then the scale developers should not tie themselves too tightly to a set of criteria that is designed primarily to diagnose, rather than measure. Diagnosis is fundamentally a qualitative issue that involves classifying individuals into categories. As such, criteria designed primarily to diagnose a trait may be insufficient to accurately measure the level of that trait.

Limitations

It should be noted that Skinner (2015) developed items for the TIPSA factors based on the five diagnostic criteria for PTSD found in the DSM (American Psychiatric Association, 2015). Because the purpose of this study was to examine the psychometric properties of the TIPSA and not to confirm its' factor structure, I assumed that each factor was indeed unidimensional, and that collectively they captured the essence of trauma in the partners of sex addicts. Current research is underway to confirm the TIPSA factor structure, and future iterations of the TIPSA ought to be similarly examined. I did not, however, test or confirm the factor structure of the TIPSA prior to running IRT analysis. While this study utilized CFA factor loadings to calculate reliability statistics, the CFA was not the primary objective. The factor analysis results used in this study revealed evidence of some correlated error terms and relatively low factor loadings.

In general, items that were initially highly dysfunctional and exhibited low discriminating power, disordered response categories, or low information functions (e.g., Items 6, 11, 22, 23, 27, 39, and 40) were not improved by post hoc revisions of the scoring structure. These item stems are likely to be problematic in terms of content or wording and are not likely to be improved by simply recoding the response options. Alternatively, option anchors that are incongruent with the issue described by a particular stem may also be problematic, and such anchors should be modified accordingly. For example, some important elements of trauma may not be describable on a five-point frequency continuum. In such cases, the items (and perhaps their options) ought to be carefully considered for content validity and either rewritten, or eliminated from the scale. Suggesting specific modifications to the content or wording of particular items is beyond the scope of the current project. Specialists in trauma, sex addiction, and related areas should be consulted to improve the content or wording of problematic items, and to ensure that content validity is maintained.

The proposed rescoring methods did improve technical item functioning in most cases, but the associated findings illustrate the limitations of improving the psychometric properties of a scale through post hoc scoring modifications alone. While many technical problems were mitigated through rescoring, this study illustrates that such methods cannot resolve issues due to poorly worded or ill-conceived item stems or response categories. Time spent creating and developing good item stems is just as important, if not more so, than post hoc psychometric analysis or modification of the resulting scores. An acceptance of this fact reveals an important limitation of the analyses conducted in this study, namely, that some items can only be improved through careful content or wording revisions.

In addition, there are important technical issues to consider going forward. First, the standard estimation procedures used in this study (MML and EM) ignored the skewed nature of the data, as well as the possibility that trauma is likely to have a skewed distribution in the population. That being the case, the analyses and conclusions outlined in this study are likely to be biased (Preston & Reise, 2014b), and further research is needed to investigate this issue. The sample size in this study was relatively large, which has been shown to reduce the bias introduced by skewed data (Preston & Reise, 2014b). However, the threat of bias because of skewness remains.

To deal with the issue of skewness, Preston and Reise (2014b) suggest the use of Ramsay Curve IRT (RC-IRT) (Woods, 2006; Woods & Thissen, 2006). This is a relatively new technique that combines MML and EM algorithms with a spline-based density-approximation procedure outlined mathematically by Ramsay (2000). Using splines and a variable number of knots, RC-IRT estimates the probability distribution function of the trait within population of examinees, rather than simply assuming that the distribution of theta is normal (Woods & Thissen, 2006). In their simulation study, Preston and Reise (2014b) found that "the benefits to implementing RC-IRT estimation in the accuracy of item parameters are great [when data are skewed], and there is very little consequence to implementing RC-IRT estimation when the latent trait is actually normally distributed" (Preston & Reise, 2014b). In a sense, they suggest that RC-IRT is a more flexible method, because it can more accurately handle skewed, as well as normally distributed data. Unfortunately, supporting literature on the use of RC-IRT with the NRM is also sparse, particularly as it relates to ordinal polytomous data (Woods, 2006). In addition, software applications that support RC-IRT remain rather limited. As of this writing, EQSIRT (Wu & Bentler, 2013) is one of the only applications capable of RC-IRT estimation. As IRT models such as the NRM are increasingly being used outside the field of education to assess psychological constructs that are more likely to be skewed in the population, the impact of RC-IRT on nuisance variation and parameter estimation deserves special attention in future research.

Future Research

First, the analysis in this study was confined to female respondents. This is a rather common approach since it is often assumed that sex addiction is primarily a male problem, and that the spouses of sex addicts are female (McCarthy, 2002; Milrad, 1999; Steffens & Rennie, 2006). More research is needed on the prevalence of female sex addicts, as well as how males respond to sex addiction in their partners. The Trauma Inventory for Partners of Sex Addicts cannot truly be said to measure the trauma in the partners of sex addicts, until we know how male partners respond to TIPSA items. Differential item functioning (DIF) may be a useful tool to determine the extent to which males differ from females in their conceptualization and response to trauma as a result of sexual addiction.

In addition, as discussed previously, future iterations of the TIPSA may benefit from explicitly incorporating other elements that are believed to impact trauma, such as the issues of cumulative trauma and level of betrayal.

Given what we know about the response data, and what we have reason to believe about the distribution of trauma in the population, a replication of the current study using RC-IRT would be instructive, and would contribute greatly to a literature that seems to be lacking. The use of RC-IRT on empirical data deserves more attention, and the TIPSA presents future researchers with an important and valuable opportunity to contribute.

Finally, future research on the implications of disordered response categories should be conducted with empirical data. The consequences of collapsing categories in response to disordering deserves more attention, and ought to be considered in future research. The potential tradeoff between information and category ordering also deserves theoretical and empirical attention.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

American Psychiatric Association. (2015). *Diagnostic and Statistical Manual of Mental Disorders.* Retrieved from http://dx.doi.org/10.1176/appi.books.9780890425596.dsm16

Ashley, L., Smith, A. B., Keding, A., Jones, H., Velikova, G., & Wright, P. (2013). Psychometric evaluation of the revised Illness Perception Questionnaire (IPQ-R) in cancer patients: confirmatory factor analysis and Rasch analysis. *Journal of Psychosomatic Research, 75*(6), 556-562.

Bancroft, J. (2008). Sexual behavior that is "out of control": A theoretical conceptual approach. *Psychiatric Clinics of North America, 31*(4), 593-601.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. London, UK: John Wiley & Sons.

Bee, P., Gibbons, C., Callaghan, P., Fraser, C., & Lovell, K. (2016). Evaluating and quantifying user and carer involvement in mental health care planning (EQUIP): Co-development of a new patient-reported outcome measure. *11*(3), e0149973.

Bell, R. C., Low, L. H., Jackson, H. J., Dudgeon, P. L., Copolov, D. L., & Singh, B. S. (1994). Latent trait modelling of symptoms of schizophrenia. *Psychological Medicine, 24*(02), 335-345.

Black, C. (2009). *Deceived: Facing sexual betrayal lies and secrets*. Center City, MN: Simon and Schuster.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Bourke, M., & Wallace, L. (2015). Improving objective measurement in nursing research: Rasch model analysis and diagnostics of the nursing students' clinical stress scale. *Journal of Nursing Measurement, 23*(1), E1.

Brogårdh, C., Lexell, J., & Lundgren-Nilsson, Å. (2013). Construct validity of a new rating scale for self-reported impairments in persons with late effects of polio. *Physical Medicine and Rehabilitation, 5*(3), 176-181.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Publications.

Cai, L. (2013). flexMIRT® version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Carnes, P. (2013). *Don't call it love: Recovery from sexual addiction.* New York, NY: Bantam.

Carnes, P., Murray, R. E., & Charpentier, L. (2004). Addiction interaction disorder. In R. Coombs (Ed.) *Handbook of addictive disorders: A practical guide to diagnosis and treatment*, (pp. 31-59). Hoboken, NJ: Wiley.

Centers for Disease Control and Prevention (CDC). (2015). *Reported Cases of STDs on the Rise in the U.S.* Retrieved from https://www.cdc.gov/nchhstp/newsroom/2015/std-surveillance-report-press-release.html

Centers for Disease Control and Prevention (CDC). (2017a). *Chlamydia - CDC Fact Sheet*. Retrieved from https://www.cdc.gov/std/chlamydia/stdfact-chlamydia-detailed.htm

Centers for Disease Control and Prevention (CDC). (2017b). *Gonorrhea - CDC Fact Sheet*. Retrieved from https://www.cdc.gov/std/gonorrhea/stdfact-gonorrhea-detailed.htm

Cermak, T. L. (1991). Co-addiction as a disease. *Psychiatric Annals, 21*(5), 266-272.

Conner, C. (2013). Who wastes the most time at work? Retrieved from http://www.forbes.com/sites/cherylsnappconner/2013/09/07/who-wastes-the-most-time-at-work/ - 44c34a87b3a5

Cooper, A., Delmonico, D. L., & Burg, R. (2000). Cybersex users, abusers, and compulsives: New findings and implications. *Sexual Addiction & Compulsivity: The Journal of Treatment and Prevention, 7*(1-2), 5-29.

Cooper, L. D., Balsis, S., & Zimmerman, M. (2010). Challenges associated with a polythetic diagnostic system: Criteria combinations in the personality disorders. *Journal of Abnormal Psychology, 119*(4), 886.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

das Nair, R., Moreton, B. J., & Lincoln, N. B. (2011). Rasch analysis of the Nottingham Extended Activities of Daily Living Scale. *Journal of Rehabilitation Medicine, 43*(10), 944-950.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Publications.

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

Dougherty, B. E., Nichols, J. J., & Nichols, K. K. (2011). Rasch analysis of the ocular surface disease index (OSDI). *Investigative Ophthalmology & Visual Science, 52*(12), 8630-8635.

Fearing, J. (2002). *Intervention and the sexually addicted patient*. In P. Carnes & K. Adams (Eds.). *Handbook of Multiple Addictions* (pp. 31-59). New York, NY: Brunner-Routledge.

Ferree, M. C. (2001). Females and sex addiction: Myths and diagnostic implications. *Sexual Addiction &Compulsivity: The Journal of Treatment and Prevention, 8*(3-4), 287-300.

Follette, V. M., Polusny, M. A., Bechtle, A. E., & Naugle, A. E. (1996). Cumulative trauma: The impact of child sexual abuse, adult sexual assault, and spouse abuse. *Journal of Traumatic Stress, 9*(1), 25-35.

Freimuth, M., Waddell, M., Stannard, J., Kelley, S., Kipper, A., Richardson, A., & Szuromi, I. (2008). Expanding the scope of dual diagnosis and co-addictions: Behavioral addictions. *Journal of Groups in Addiction & Recovery, 3*(3-4), 137-160.

Freyd, J. J. (1998). *Betrayal trauma: The logic of forgetting childhood abuse*. Boston, MA: Harvard University Press.

García-Pérez, M. A. (2017). An Analysis of (Dis)Ordered Categories, Thresholds, and Crossings in Difference and Divide-by-Total IRT Models for Ordered Responses. *The Spanish Journal of Psychology,* 20(10), 1-27.

Goodman, A. (1990). Addiction: Definition and implications. *British Journal of Addiction, 85*(11), 1403-1408.

Hentsch-Cowles, G., & Brock, L. J. (2013). A systemic review of the literature on the role of the partner of the sex addict, treatment models, and a call for research for systems theory model in treating the partner. *Sexual Addiction & Compulsivity, 20*(4), 323-335.

Kafka, M. P. (2010). Hypersexual disorder: A proposed diagnosis for DSM-V. *Archives of Sexual Behavior, 39*(2), 377-400.

Kim, S., Kim, S. H., & Kamphaus, R. W. (2010). Is aggression the same for boys and girls? Assessing measurement invariance with confirmatory factor analysis and item response theory. *School Psychology Quarterly, 25*(1), 45.

Kraus, S. W., Voon, V., & Potenza, M. N. (2016). Should compulsive sexual behavior be considered an addiction? *Addiction, 111*(12), 2097-2106.

Laaser, M. R. (2002). *Recovery for couples*. New York, NY: Brunner-Routledge.

Lebowitz, L., Harvey, M. R., & Herman, J. L. (1993). A stage-by-dimension model of recovery from sexual trauma. *Journal of Interpersonal Violence, 8*(3), 378-391.

MacLaren, V. V., & Best, L. A. (2010). Multiple addictive behaviors in young adults: Student norms for the Shorter PROMIS Questionnaire. *Addictive Behaviors, 35*(3), 252-255.

Martin, C. G., Cromer, L. D., DePrince, A. P., & Freyd, J. J. (2013). The role of cumulative trauma, betrayal, and appraisals in understanding trauma symptomatology. *Psychological Trauma: Theory, Research, Practice, and Policy, 5*(2), 110.

McCarthy, B. W. (2002). The Wife's Role in Facilitating Recovery from Male Compulsive Sexual Behavior. *Sexual Addiction & Compulsivity: The Journal of Treatment and Prevention, 9*(4), 275-284.

McDonald, R. P. (2013). *Test theory: A unified treatment.* New York, NY: Psychology Press.

Milrad, R. (1999). Coaddictive recovery: Early recovery issues for spouses of sex addicts. *Sexual Addiction & Compulsivity: The Journal of Treatment and Prevention, 6*(2), 125-136.

Minwalla, O. (2011). What about me and my sexuality? In S. Carnes (Ed.), *Mending a shattered heart: A guide for partners of sex addicts*. Carefree, Arizona: Gentle Path Press (pp. 93-112).

Morgan, J. P. (1991). What is codependency? *Journal of clinical psychology, 47*(5), 720-729.

Muraki, E. (1997). A generalized partial credit model. In W. J. Van Der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 153-164). New York, NY: Springer.

Murray, A., Booth, T., & Molenaar, D. (2016). When middle really means “Top” or “Bottom”: An analysis of the 16PF5 using Bock's nominal response model. *Journal of Personality Assessment, 98*(3), 319-331.

Muthen, L., & Muthen, B. (2015). MPlus (Version 7.4)[Software].

Nilsson, Å. L., Sunnerhagen, K. S., & Grimby, G. (2005). Scoring alternatives for FIM in neurological disorders applying Rasch analysis. *Acta neurologica scandinavica, 111*(4), 264-273.

Oluboyede, Y., & Smith, A. B. (2013). Evidence for a unidimensional 15-item version of the CASP-19 using a Rasch model approach. *Quality of Life Research, 22*(9), 2429-2433.

Preston, K. (2014a). *Advanced topics in IRT: Evaluating the effectiveness of each response option with the nominal response model*. Paper presented at the 94th Annual Convention of the Western Psychological Association, Portland, OR.

Preston, K. (2014b). R Code for conducting Wald tests. Retrieved from http://hssfaculty.fullerton.edu/psychology/kpreston/Wald.txt

Preston, K. (2014c). R Code for plotting Category Response Curves. Retrieved from http://hssfaculty.fullerton.edu/psychology/kpreston/Plotting.txt

Preston, K., Parral, S., Gottfried, A., Oliver, P., Gottfried, A., Ibrahim, S., & Delany, D. (2015). Applying the Nominal Response Model within a longitudinal framework to construct the Positive Family Relationships Scale. *Educational and Psychological Measurement*, 75(6), 901-930.

Preston, K., & Reise, S. (2014a). Detecting faulty within-item category functioning with the Nominal Response Model *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 386-405). New York, NY: Routledge.

Preston, K., & Reise, S. (2014b). Estimating the Nominal Response Model under nonnormal conditions. *Educational and Psychological Measurement, 74*(3), 377-399.

Preston, K., Reise, S., Cai, L., & Hays, R. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement, 71*(3), 523-550.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/.

Ramsay, J. O. (2000). Differential equation models for statistical functions. *Canadian Journal of Statistics, 28*(2), 225-240.

Raykov, T. (2009). Evaluation of scale reliability for unidimensional measures using latent variable modeling. *Measurement and Evaluation in Counseling and Development, 42*(3), 223-232.

Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.

RStudio Team. (2015). *RStudio: Integrated development for R*. Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/.

Schneider, J. P., & Irons, R. R. (2001). Assessment and treatment of addictive sexual disorders: Relevance for chemical dependency relapse. *Substance Use & Misuse, 36*(13), 1795-1820.

Schumm, J. A., Briggs-Phillips, M., & Hobfoll, S. E. (2006). Cumulative interpersonal traumas and social support as risk and resiliency factors in predicting PTSD and depression among inner-city women. *Journal of Traumatic Stress, 19*(6), 825-836.

Skinner, K. B. (2015). *The Lasting Effects of Sexual Betrayal*. Retrieved from https://www.psychologytoday.com/blog/inside-porn-addiction/201508/the-lasting-effects-sexual-betrayal

Stafford, L. L. (2001). Is codependency a meaningful concept? *Issues in mental health nursing, 22*(3), 273-286.

Steffens, B., & Rennie, R. (2006). The traumatic nature of disclosure for wives of sexual addicts. *Sexual Addiction & Compulsivity, 13*(2-3), 247-267.

Sussman, S., Lisha, N., & Griffiths, M. (2011). Prevalence of the addictions: A problem of the majority or the minority? *Evaluation & The Health Professions, 34*(1), 3-56.

The Economist. (2015). *Naked Capitalism*. Retrieved from http://www.economist.com/news/international/21666114-internet-blew-porn-industrys-business-model-apart-its-response-holds-lessons

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26*(2), 161-176.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8-14.

Ura, M., Preston, K., & Mearns, J. (2015). A measure of prejudice against accented English (MPAAE): scale development and validation. *Journal of Language and Social Psychology*, 34(5), 539-563. doi: 0261927X15571537.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics, 16*(2), 117-186.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Chichester, UK: John Wiley & Sons.

Washton, A. (1989). Cocaine may trigger sexual compulsivity. *US Journal of Drug and Alcohol Dependency, 13*(6), 8-14.

Woods, C. M. (2006). Ramsay-Curve Item Response Theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods, 11*(3), 253-270.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*(2), 281-301.

Wu, E. J. C., & Bentler, P. M. (2013). *EQSIRT: A comprehensive item response theory program* [Computer software]. Encino, CA: Multivariate Software.

APPENDIX A: TIPSA Items

Table 15

TIPSA Items

Item and Factor #	Item Anchor
Factor 1: Exposed to threat	
1	I experience intense feelings of indescribable fear since discovering my partner’s sexual misbehaviors.
2	I have feelings of helplessness since discovering my partner’s behaviors.
3	My partner threatens to hurt me in some way if I do not comply with his/her sexual fantasies.
4	Due to my partner’s sexual behaviors I have become concerned that I might contract a sexually transmitted disease.
5	Since discovering my partner’s behaviors, I have a hard time determining who is safe to be around and who is not safe to be around.
6	I felt safe with my partner until I discovered his/her sexual behaviors. (Reverse coded)
7	I feel violated due to my partner’s sexual behaviors.
8	I feel like my partner will never stop sexually acting out.
9	Since learning of my partner's behaviors, he/she has hurt, hit, or threatened me.
10	Since learning of my partner's behaviors, I am afraid of my partner.
11	My partner’s behaviors remind me of experiences I had earlier in my life.
Factor 2: Reliving the event	
12	Since discovering my partner’s behaviors, I can’t look at him without thinking about them.
13	I have strong memories that remind me of my partner’s participation in sexually inappropriate behaviors.
14	I have disturbing dreams that remind me of my partner’s sexual problems.
15	When my partner tries to get close to me or we are sexually intimate I cannot help but question whether my partner is thinking about me or things he/she has done.
16	I have episodes where I feel like I am reliving the event over and over again.
17	I have a hard time with media because so many things remind me of what my partner has done.
18	I have a hard time being in public places with my partner because I have become highly sensitive to what my partner is looking at.
19	Since discovering my partner’s behavior, when I see sexually suggestive images I feel anxious.
20	If I am exposed to things that remind me of what my partner has done, I suddenly become physically ill (i.e., nauseous, head- aches, anxiety, vomit).
21	I struggle to think of other things besides what my partner has done.

(table continued)

Table 15 (continued)

Item and Factor #	Item Anchor
Factor 3: Avoidance	
22	I spend a lot of energy trying to avoid thinking about my partner's behaviors.
23	I engage in behaviors that distract me (i.e., excessive reading, sleeping, eating, drinking) from thinking about my partner's behavior.
24	I avoid sexual contact with my partner since discovering his/her behavior.
27	I avoid going to places or locations where people could be dressed scantily (e.g., mall, swimming pool, parks)
28	I intentionally plan activities to avoid being around my partner.
29	It is hard for me to be around others when they say something positive about their partner.
Factor 4: Cognition and negative mood	
25	Since discovering my partner's behavior I get distracted easily.
26	Since learning of my partners' behaviors, I have a hard time participating in things that I previously enjoyed.
30	I feel like my partner acts out because I am not good enough.
31	I feel like it is my fault that my partner sexually acts out.
32	After what my partner has done I feel like it is hard to trust anyone.
33	I feel ashamed because of what my partner has done.
34	Since I discovered my partner's behavior I hold back from people who used to be close to me.
35	I am still able enjoy things since learning of my partner's behavior. (Reverse coded)
36	I feel like I am a bad person because of what my partner has done.
37	When I am in social settings I don't feel like I belong anymore.
38	Since discovering my partner's behaviors, I feel like I am different than everyone else.
39	I feel people, in general, are safe. (Reverse coded)
40	I feel like my spouse would not be this way if society was not so bad.
Factor 5: Emotional arousal	
41	Since learning of my partner's behavior I have difficulty falling asleep.
42	After discovering my partner's sexual behaviors, I find that I am increasingly angry in response to my partner.
43	I find that I am more critical in conversations with my partner since discovering his/her behavior.
44	I feel like I am emotionally on edge more now than I used to be before all this happened.
45	I have been surprisingly calm since discovering my partners' behaviors. (Reverse coded)
46	I find it harder to focus on what is going on around me since I discovered my partner's behavior.
47	I find that I lose things since I learned of my partner's behaviors.
48	I closely monitor my partner's behaviors.
49	When I am around my partner, I am constantly trying to read his/her emotions.
50	I feel like I need to check up on my partner.
51	I feel more anxious since I learned of my partner's behavior.
52	I feel suicidal due to this experience with my partner.
53	I am worried that I may follow through on an impulse to hurt myself.

APPENDIX B: flexMIRT Syntax

Sample flexMIRT syntax for modeling Factor 1: Exposed to Threat using the GPCM. Subsequent factors were analyzed using the same syntax by selecting the appropriate variables and changing the output file names.

```
<Project>
  Title = "TIPSA";
  Description = "IRT GPCM Analysis TIPSA Exposed to Threat";

<Options>
  Mode = Calibration;
        SE = SEM;
        smartSEM = Yes;
        SaveSCO = Yes;
        SavePRM = Yes;
        SaveDBG = Yes;
        SaveINF = Yes;
        SaveCOV = Yes;
        FisherInf = 81,4.0;
        Score = EAP;
        GOF = Extended;
        M2 = Full;
        FitNullModel = Yes;

<Groups>
  %Group1%
  File = "F:\5.2 reversed TIPSA.dat";
  Varnames = v1-v53;
  Select = v1-v11;
  Missing = -99;
  N = 2339;
  Ncats(v1-v11) = 5;
  Model(v1-v11) = gpc(5);

<Constraints>
```

Sample flexMIRT syntax for modeling Factor 1: Exposed to Threat using the NRM. Subsequent factors were analyzed using the same general syntax by selecting the appropriate variables and changing the output file names.

```
<Project>
  Title = "TIPSA";
  Description = "IRT NRM Analysis TIPSA Exposed to Threat";

<Options>
  Mode = Calibration;
        SE = SEM;
        smartSEM = Yes;
        SaveSCO = Yes;
        SavePRM = Yes;
        SaveDBG = Yes;
        SaveINF = Yes;
        SaveCOV = Yes;
        FisherInf = 81,4.0;
        Score = EAP;
        GOF = Extended;
        M2 = Full;
        FitNullModel = Yes;

<Groups>
  %Group1%
  File = "F:\5.2 reversed TIPSA.dat";
  Varnames = v1-v53;
  Select = v1-v11;
  Missing = -99;
  N = 2339;
  Ncats(v1-v11) = 5;
  Model(v1-v11) = Nominal(5);
  Ta(v1-v11) =
        (0 0 0 0,
         1 0 0 0,
         1 1 0 0,
         1 1 1 0,
         1 1 1 1);
  Tc(v1-v11) = Trend;
<Constraints>
```

APPENDIX C: R Syntax

Sample R code used to calculate CBD parameters, category intersection parameters, and plot category response and information functions using flexMIRT output files (Preston, 2014c). Subsequent factors were analyzed by modifying the “flexname” variable to coincide with the appropriate flexMIRT output files. Some modifications were also made to this general syntax to plot the superimposed Test Information Functions (Figure 1).

```
wd <- “f:/”
flexname <- “Factor 1 Exposure NRM_2.0 R1”

itemp <- read.delim(file=paste(wd,flexname,"-inf.txt",sep=""), header=FALSE, sep = "\t")
nitems <- nrow(itemp)-1

iteminfo <- as.matrix(itemp[1:nitems,3:83],dimnames=NULL)
testinfo <- itemp[(nitems+1),3:83]

xtemp <- scan(file=paste(wd,flexname,"-irt.txt",sep=""), what="real")

for (i in 1:length(xtemp)) {
  if (xtemp[i] == "Categories") newx <- xtemp[(i-1):((i+(nitems+1)*5)-2)]
}
newx <- matrix(newx,(nitems+1),5,byrow=TRUE)
newx <- matrix(as.numeric(newx[2:(nitems+1),1:2]),nitems,2)
rnames <- as.list(newx[,1])
mincat <- min(newx[,2])
maxcat <- max(newx[,2])
for(k in mincat:maxcat){
 cat <- matrix(0,nitems,1)
for(j in 1:nitems){
  if(newx[j,2]==k) cat[j]<-newx[j,1]
  else cat[j] <- NA
}
cat<-as.list(na.omit(cat))
assign(paste("cat",k,sep=""),cat)
}
for (i in 1:length(xtemp)) {
  if (xtemp[i] == "(Bock,") newz <- xtemp[(i-1):(i+17+(nitems*2)*maxcat*2)]
}

theta <- seq(-4.0,4.0,0.1)
cpar <- matrix(0,nitems,maxcat)
apar <- matrix(0,nitems,maxcat)
CBD <- matrix(0,nitems,(maxcat-1))
int <- matrix(0,nitems,(maxcat-1))
P <- NULL
```

```
info <- NULL
relinfo <- NULL

for(q in mincat:maxcat){
for (k in get(paste("cat",q,sep=""))) {
  for (i in 1:length(newz)) {
    if ((newz[i] == k) && (newz[i+2]=="a"))  {
      apar[k,1:q] <- newz[(i+3):(i+2+q)]
      cpar[k,1:q] <- newz[(i+4+q):(i+3+q*2)]
      rnames[[k]] <- as.numeric(gsub("[^[:digit:]]", "",newz[i+1]))
      break
}}}}

apar <- matrix(as.numeric(apar),nitems,maxcat)
cpar <- matrix(as.numeric(cpar),nitems,maxcat)

logit <- NULL
catP <- matrix(0,length(theta),maxcat)
catinfo <- matrix(0,maxcat,ncol(iteminfo))
for(q in mincat:maxcat){
for(k in get(paste("cat",q,sep=""))) {
for(j in 1:q){
  logit[[j]]<-exp(apar[k,j] * theta + cpar[k,j])}
total <- Reduce("+",logit)
for(j in 1:q){
  catP[,j] <- logit[[j]]/total}
  P[[k]] <- catP
for(j in 1:q){
    catinfo[j,] <- iteminfo[k,] * catP[,j]
  }
  info[[k]] <- rbind(iteminfo[k,],catinfo)
  relinfo[[k]] <- info[[k]]/q
}}

for(k in mincat:maxcat){
for(m in 1:nitems){
  if(apar[m,k]==0){apar[m,k]<-NA}
  if(cpar[m,k]==0){cpar[m,k]<-NA}
}}

for(q in 1:(maxcat-1)){
CBD[,q]<-(apar[,(q+1)] - apar[,q])
int[,q] <- (cpar[,q] - cpar[,(q+1)]) / (CBD[,q])
}
rownames(CBD) <- unlist(rnames)
colnames(CBD) <- c(paste("CBD", 1:(maxcat-1)))
```

```
print(CBD)
rownames(int) <- unlist(rnames)
colnames(int) <- c(paste("Int", 1:(maxcat-1)))
print(round(int,2))

setwd(wd)

for(m in 1:(nitems)){
  pdf(paste(flexname,"CRC Item",rnames[[m]],".pdf"))
  plotP <- P[[m]]
  matplot(theta,plotP[,1:newx[m,2]],ylim=c(0,1),xlim=c(-
4,4),xlab=expression(theta),ylab="Probability",
type="l",lty=1,lwd=3,col=c(2:(maxcat+1)),main=paste("Category Response Curves
Item",rnames[[m]]))
}

graphics.off()
for(m in 1:(nitems)){
  plotinfo <- t(relinfo[[m]])
  pdf(paste(flexname,"Info - Item",rnames[[m]],".pdf"))
  matplot(theta,plotinfo[,1:(newx[m,2]+1)],ylim=c(0,1),xlim=c(-
4,4),xlab=expression(theta),ylab="Information",
type="l",lty=1,lwd=3,col=c(1:(maxcat+1)),main=paste("Item and Category Information
Item", rnames[[m]]))
}

graphics.off()
pdf(paste(flexname,"Test Info.pdf",sep=""))
totalpars <- sum(newx[,2])-nitems
reltest <- testinfo/totalpars
matplot(theta,t(reltest),ylim=c(0,1),xlim=c(-4,4),xlab=expression(theta),ylab="Relative
Information", type="l",lty=c(1),lwd=c(3),col=c(1),main=paste("Relative Test
Information"))
graphics.off()
```

Sample R code used to calculate Wald tests on the CBD parameters using flexMIRT output files (Preston, 2014b). Subsequent factors were analyzed by modifying the “flexname” variable to coincide with the appropriate flexMIRT output files.

```
wd <- “f:/”
flexname <- “Factor 1 Exposure NRM_2.0 R1”

itemp <- read.delim(file=paste(wd,flexname,"-inf.txt",sep=""), header=FALSE, sep = "\t")
nitems <- nrow(itemp)-1

iteminfo <- as.matrix(itemp[1:nitems,3:83],dimnames=NULL)
testinfo <- itemp[(nitems+1),3:83]

xtemp <- scan(file=paste(wd,flexname,"-irt.txt",sep=""), what="real")

for (i in 1:length(xtemp)) {
  if (xtemp[i] == "Categories") newx <- xtemp[(i-1):((i+(nitems+1)*5)-2)]
}
newx <- matrix(newx,(nitems+1),5,byrow=TRUE)
newx <- matrix(as.numeric(newx[2:(nitems+1),1:2]),nitems,2)
rnames <- as.list(newx[,1])
mincat <- min(newx[,2])
maxcat <- max(newx[,2])
for(k in mincat:maxcat){
  cat <- matrix(0,nitems,1)
  for(j in 1:nitems){
    if(newx[j,2]==k) cat[j]<-newx[j,1]
    else cat[j] <- NA
  }
  cat<-as.list(na.omit(cat))
  assign(paste("cat",k,sep=""),cat)
}
for (i in 1:length(xtemp)) {
  if (xtemp[i] == "(Bock,") newz <- xtemp[(i-1):(i+17+(nitems*2)*maxcat*2)]
}


theta <- seq(-4.0,4.0,0.1)
cpar <- matrix(0,nitems,maxcat)
apar <- matrix(0,nitems,maxcat)
CBD <- matrix(0,nitems,(maxcat-1))
int <- matrix(0,nitems,(maxcat-1))

for(q in mincat:maxcat){
  for (k in get(paste("cat",q,sep=""))) {
    for (i in 1:length(newz)) {
```

```
      if ((newz[i] == k) && (newz[i+2]=="a"))  {
        apar[k,1:q] <- newz[(i+3):(i+2+q)]
        cpar[k,1:q] <- newz[(i+4+q):(i+3+q*2)]
        rnames[[k]] <- as.numeric(gsub("[^[:digit:]]", "",newz[i+1]))
        break
      }}}}

apar <- matrix(as.numeric(apar),nitems,maxcat)
cpar <- matrix(as.numeric(cpar),nitems,maxcat)

for(k in mincat:maxcat){
  for(m in 1:nitems){
    if(apar[m,k]==0){apar[m,k]<-NA}
    if(cpar[m,k]==0){cpar[m,k]<-NA}
  }}

for(q in 1:(maxcat-1)){
  CBD[,q]<-(apar[,(q+1)] - apar[,q])
  int[,q] <- (cpar[,q] - cpar[,(q+1)]) / (CBD[,q])
}
pnum <- matrix(0,nitems,1)
for (k in 1:nitems) {
for(n in 1:length(xtemp)){
  if (xtemp[n] == paste("v",rnames[[k]], sep="")) {
    pnum[k,] <- as.numeric(xtemp[n+1])
    break
}}}

for(q in mincat:maxcat){
  for (k in get(paste("cat",q,sep=""))) {
    pnum[k,] <- pnum[k,]-(q-2)
  }}

sigma <- read.csv(file=paste(wd,flexname,"-cov.txt",sep=""), header=FALSE,sep=",",
dec=".")

ahat <- matrix(0,(maxcat-1),1)
sigmahat <- matrix(0,(maxcat-1),(maxcat-1))
Q <- matrix(0,nitems,1)
pro <- matrix(0,nitems,1)
df <- matrix(0,nitems,1)

for(q in mincat:maxcat){
  for (k in get(paste("cat",q,sep=""))) {
    ahat <- matrix(0,(q-1),1)
    sigmahat <- matrix(0,(q-1),(q-1))
```

```
    ahat <- as.matrix(CBD[k,1:(q-1)])
    sigmahat <- as.matrix(sigma[pnum[k]:(pnum[k]+(q-2)),pnum[k]:(pnum[k]+(q-2))])

  l <- t(contr.poly(q-1))
  lahat <- l %*% ahat
  lsigmalt <- l %*% sigmahat %*% t(l)

  Q[k,] <- t(lahat) %*% solve(lsigmalt) %*% lahat
  pro[k,] <- pchisq(Q[k,], df= (q-2), lower.tail=FALSE)
  df[k,] <- (q-2)
}}
# 2 because constraint on both a and c
(  out <- round(cbind(Q,df,pro),digits=3))

rownames(out) <- unlist(rnames)
colnames(out) <- c("Q",'df',"p-val")
print(out)
```

APPENDIX D: Items with a Poorly Functioning Bottom Category

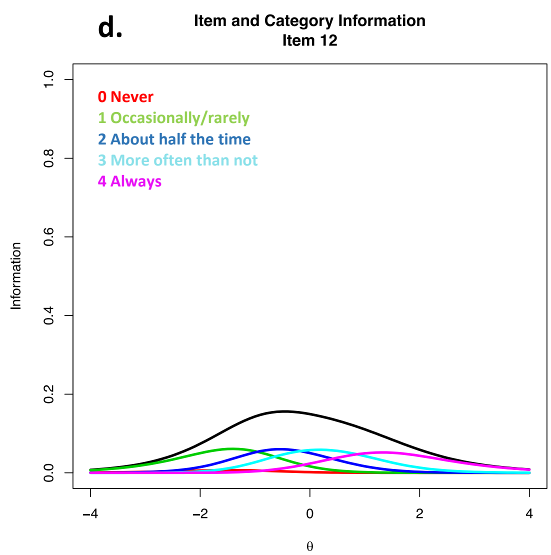
a.

0 Never
1 Occasionally/rarely
2 About half the time
3 More often than not
4 Always

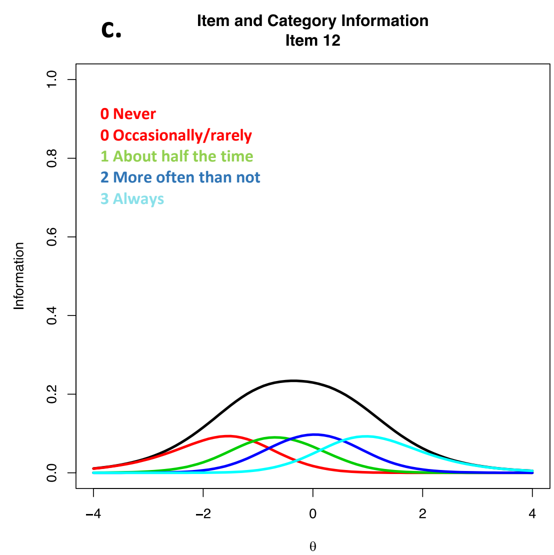
b.

0 Never
0 Occasionally/rarely
1 About half the time
2 More often than not
3 Always

d.



c.



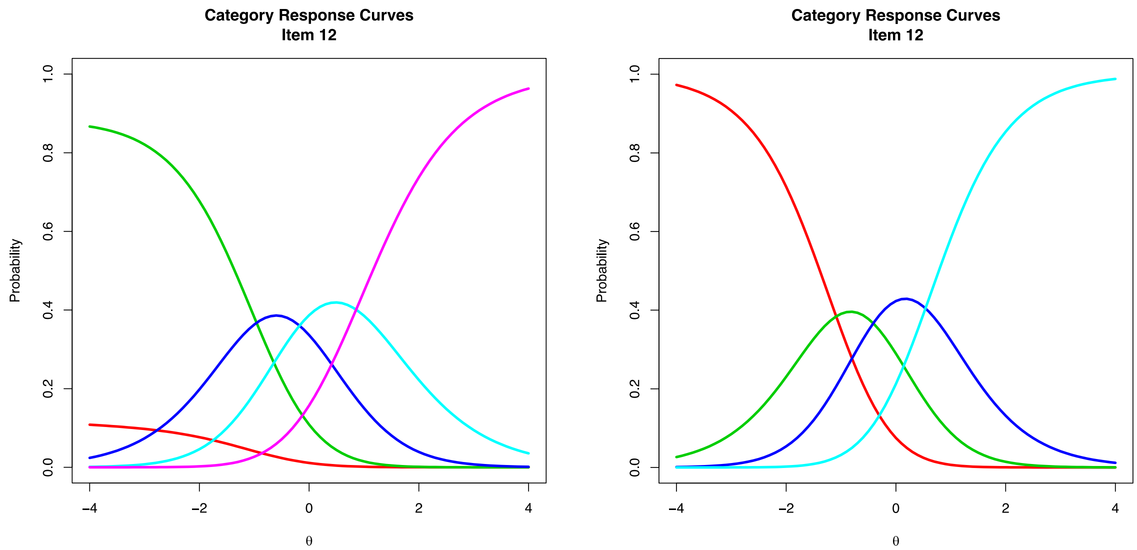


Figure 11. Factor 2 Item 12: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

0 Never
0 Occasionally/rarely
1 About half the time
2 More often than not
3 Always

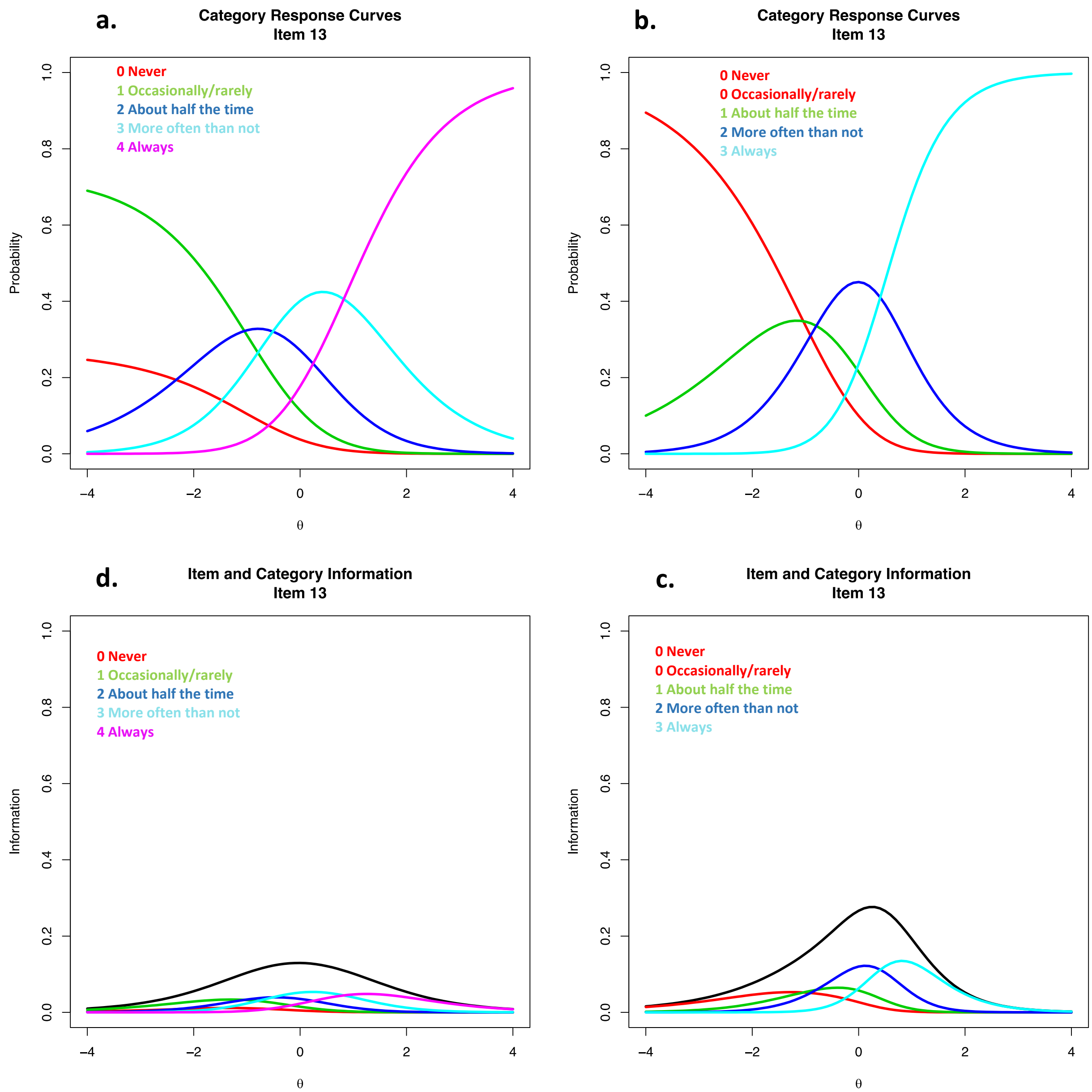


Figure 12. Factor 2 Item 13: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

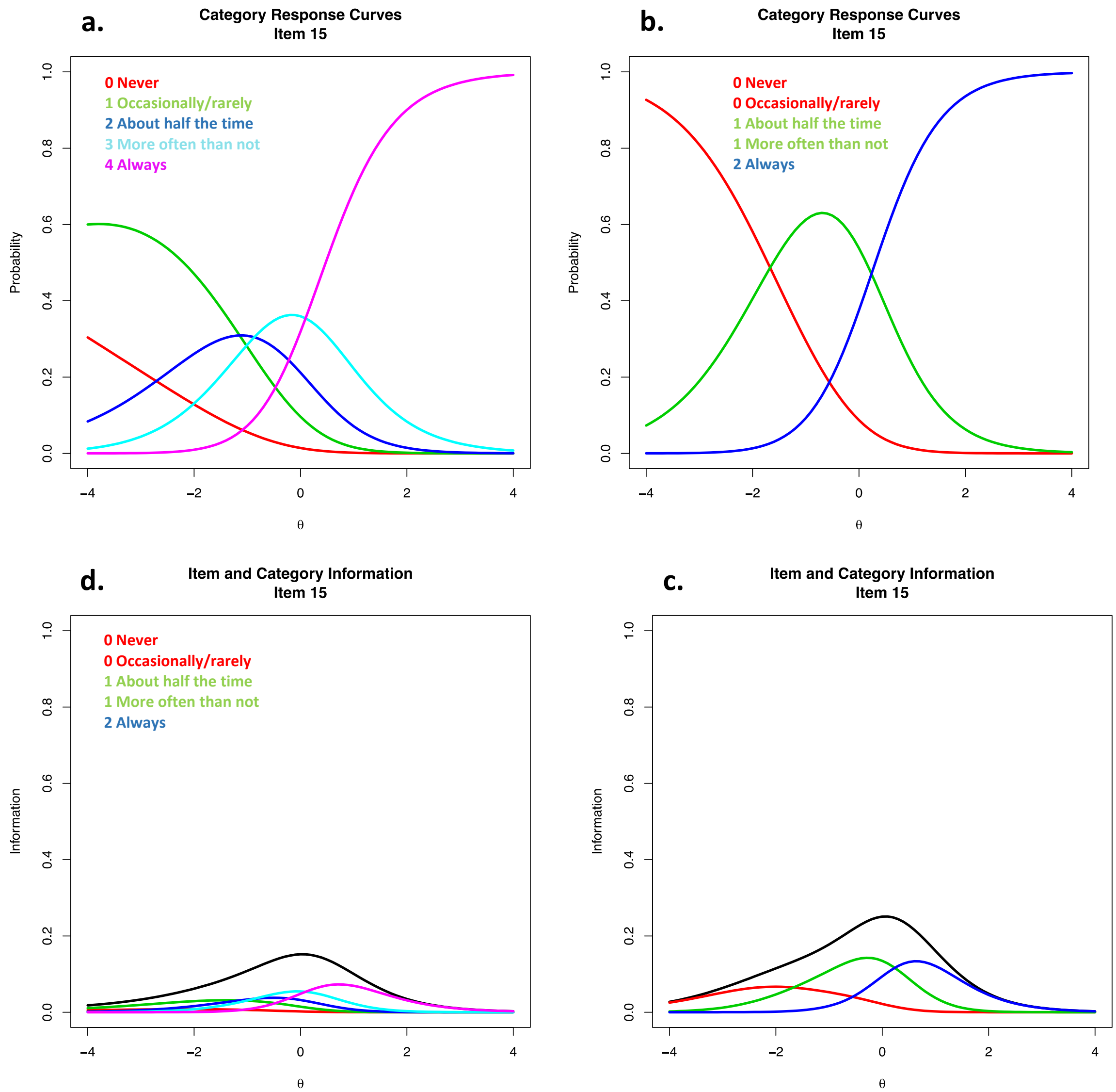


Figure 13. Factor 2 Item 15: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

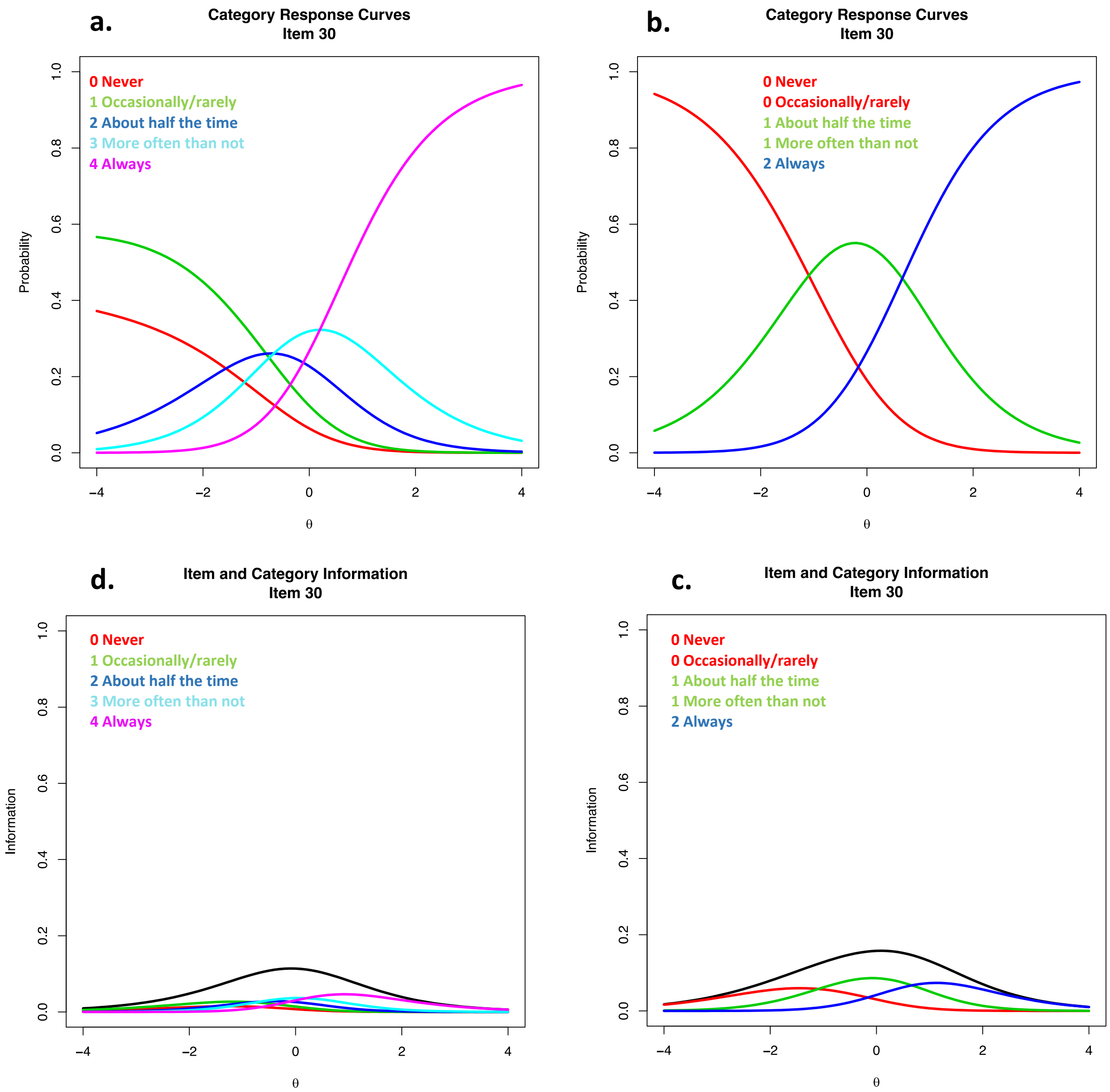


Figure 14. Factor 4 Item 30: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

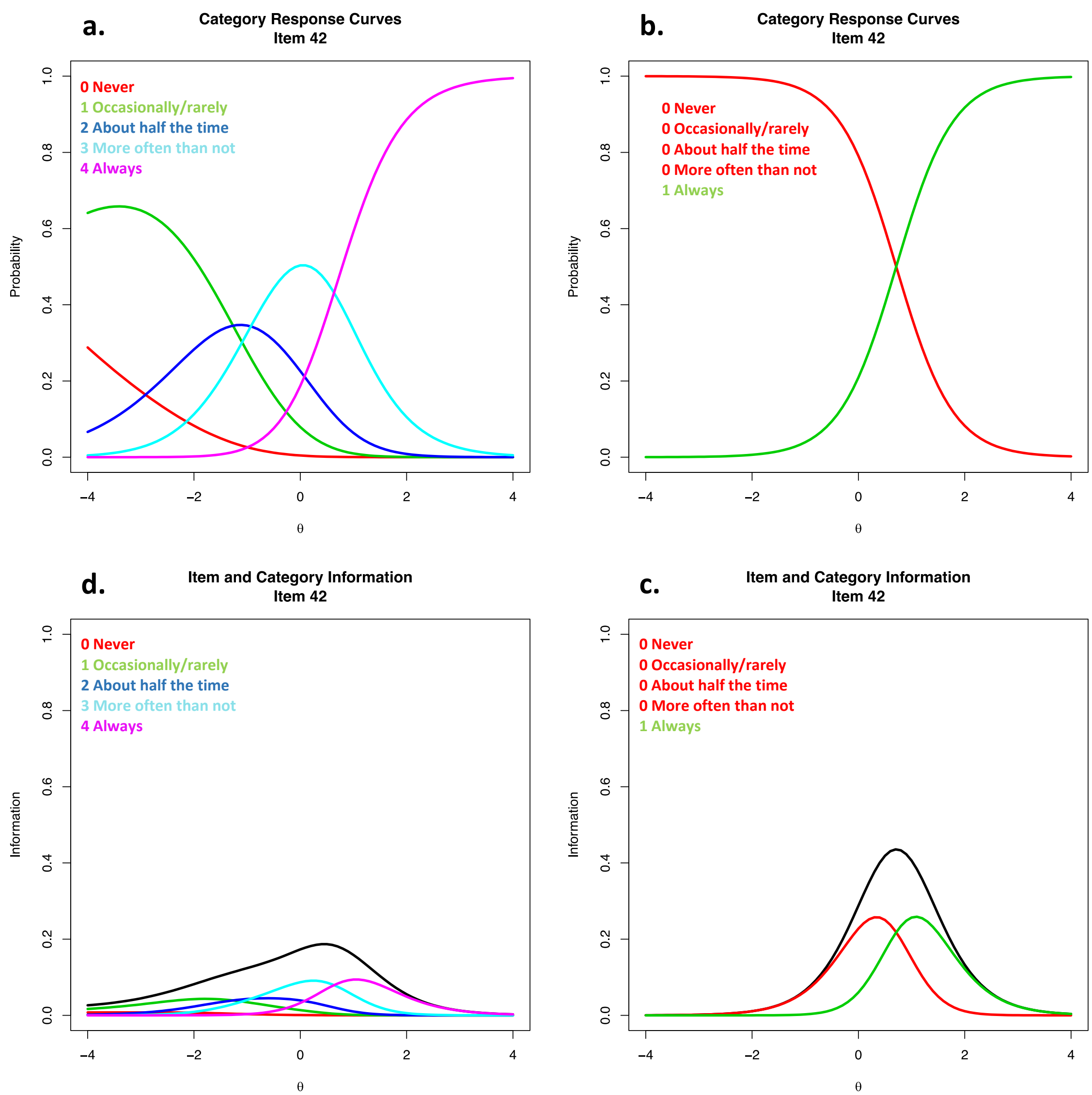


Figure 15. Factor 5 Item 42: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

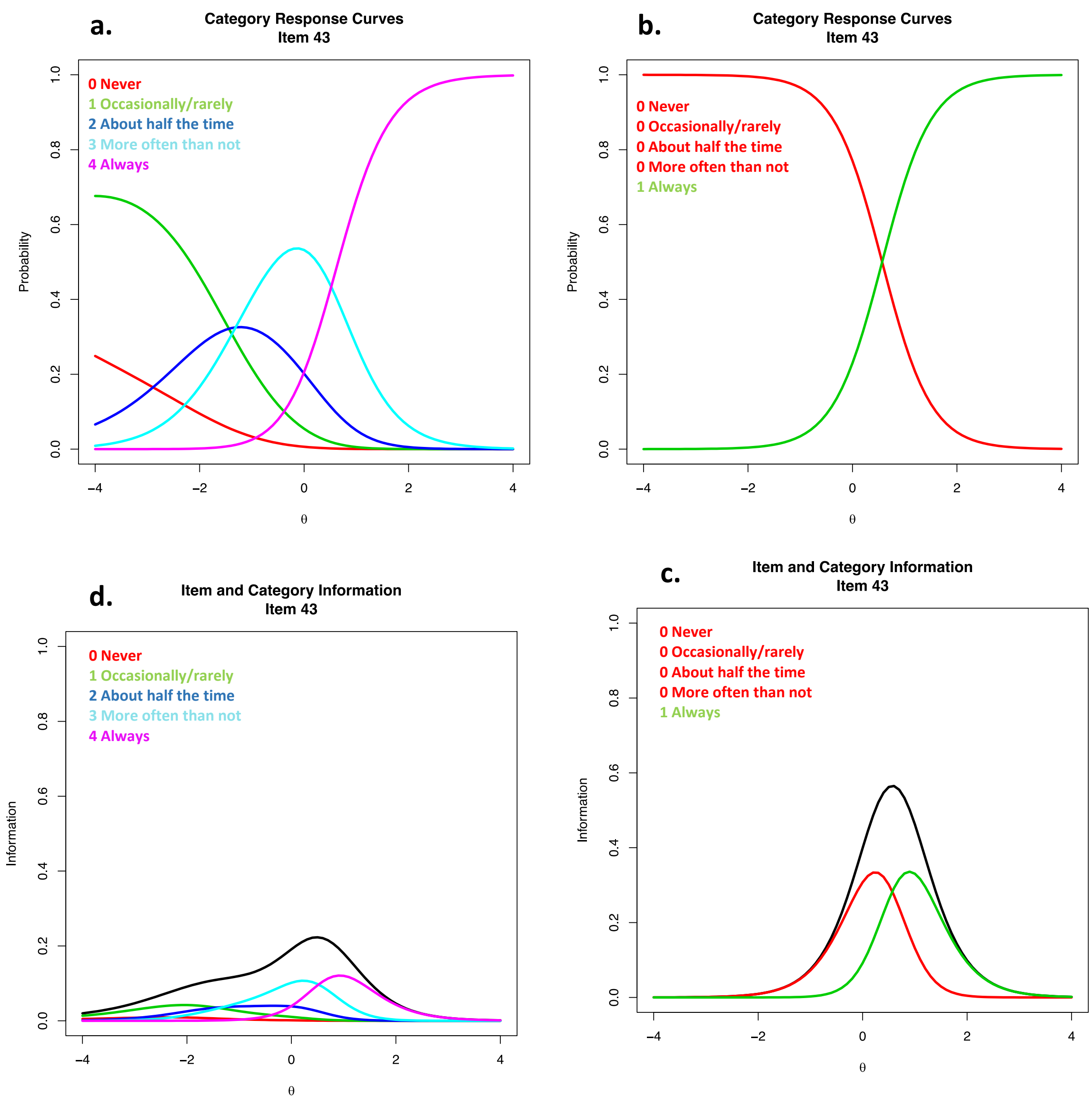


Figure 16. Factor 5 Item 43: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

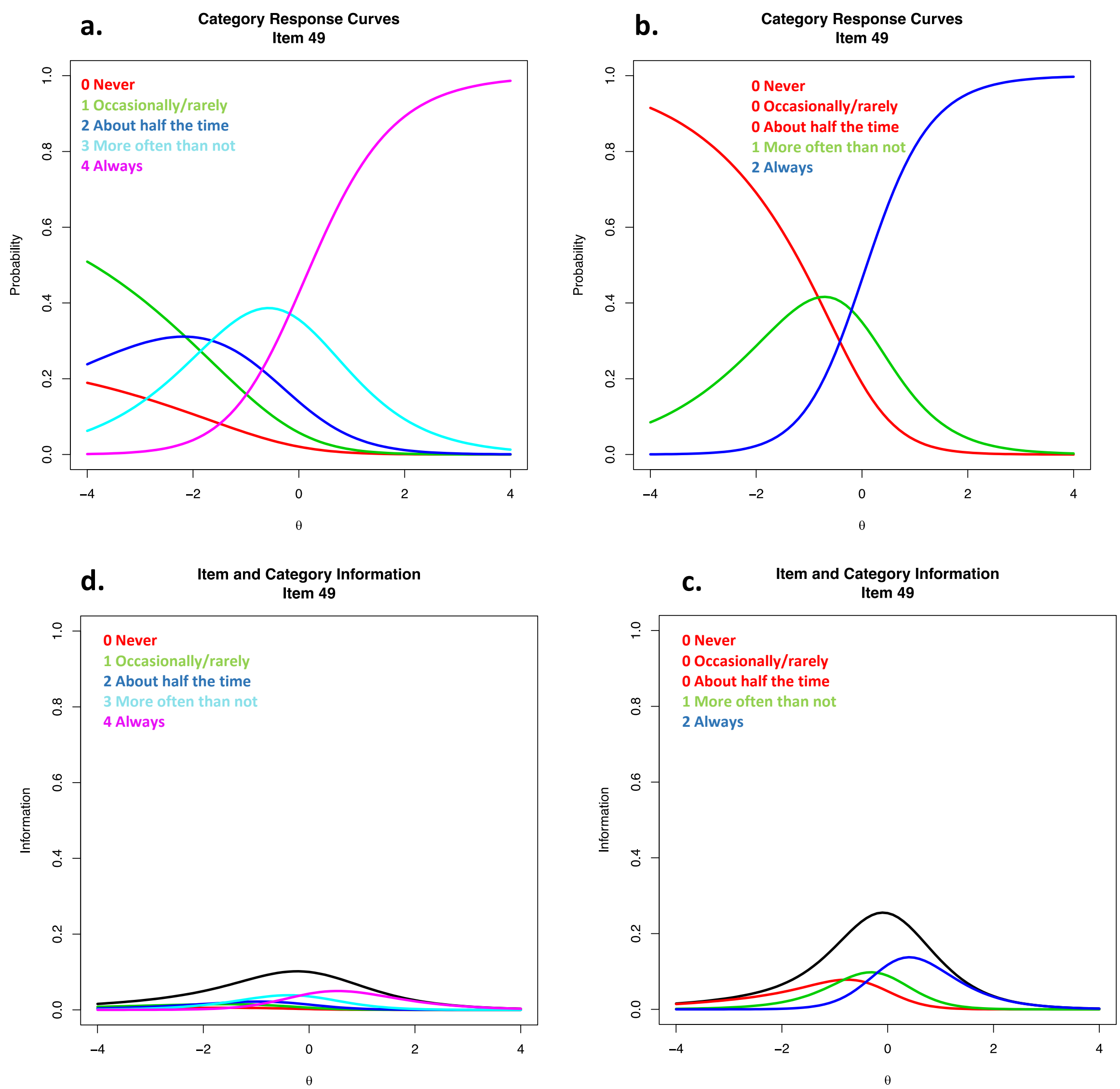


Figure 17. Factor 5 Item 49: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

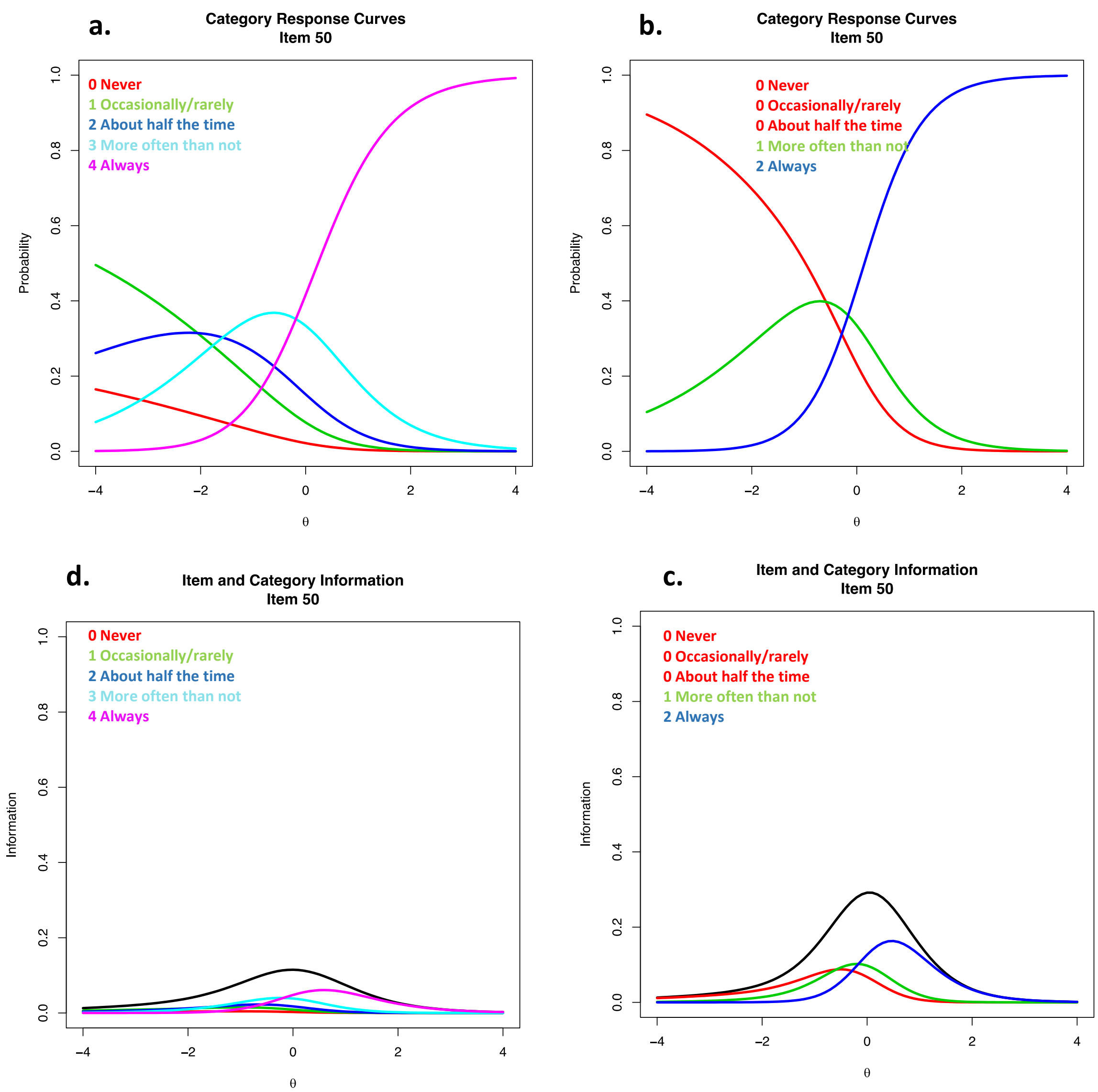


Figure 18. Factor 5 Item 50: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

APPENDIX E: Items with a Poorly Functioning Middle Option

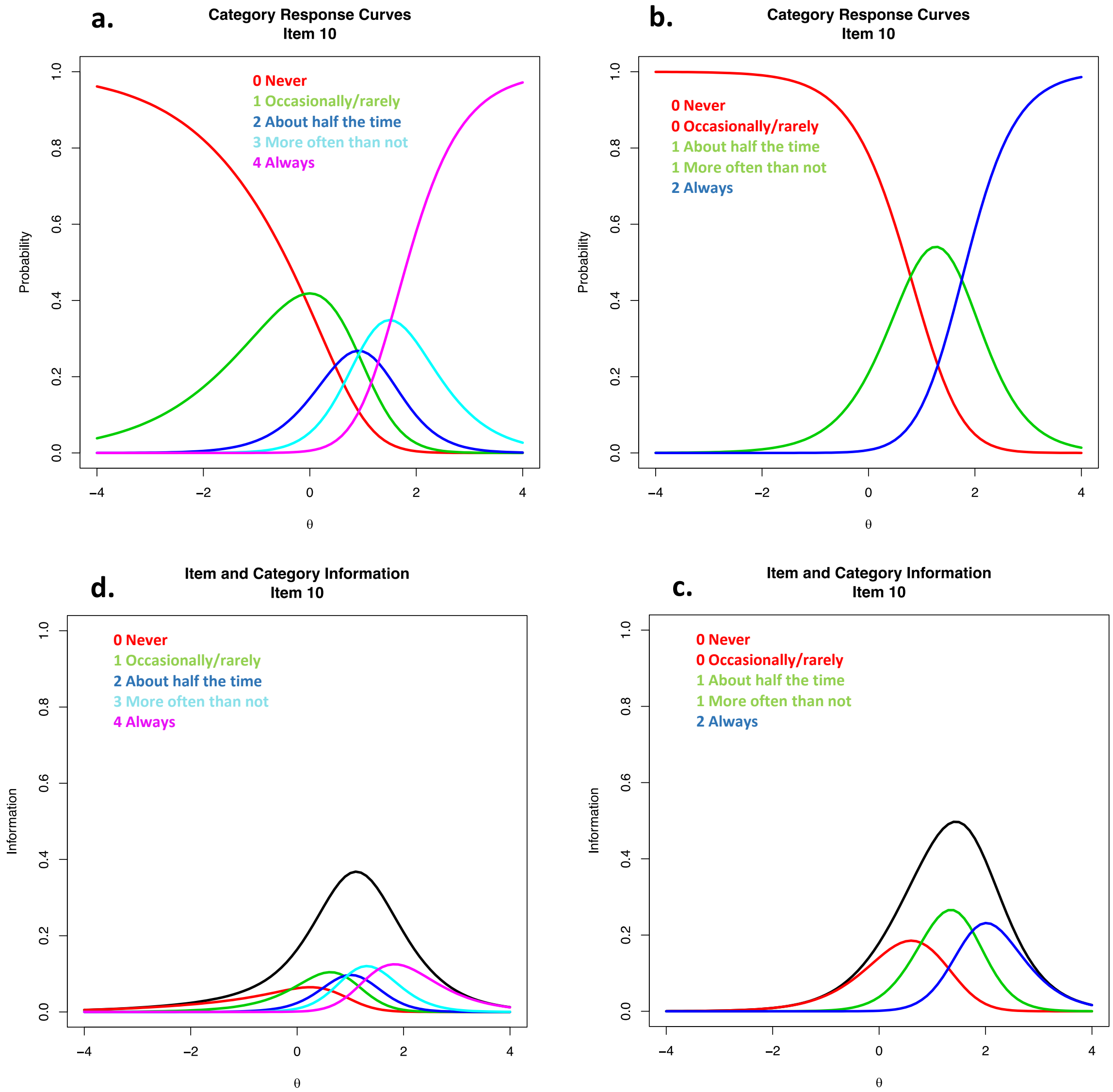


Figure 19. Factor 1 Item 10: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

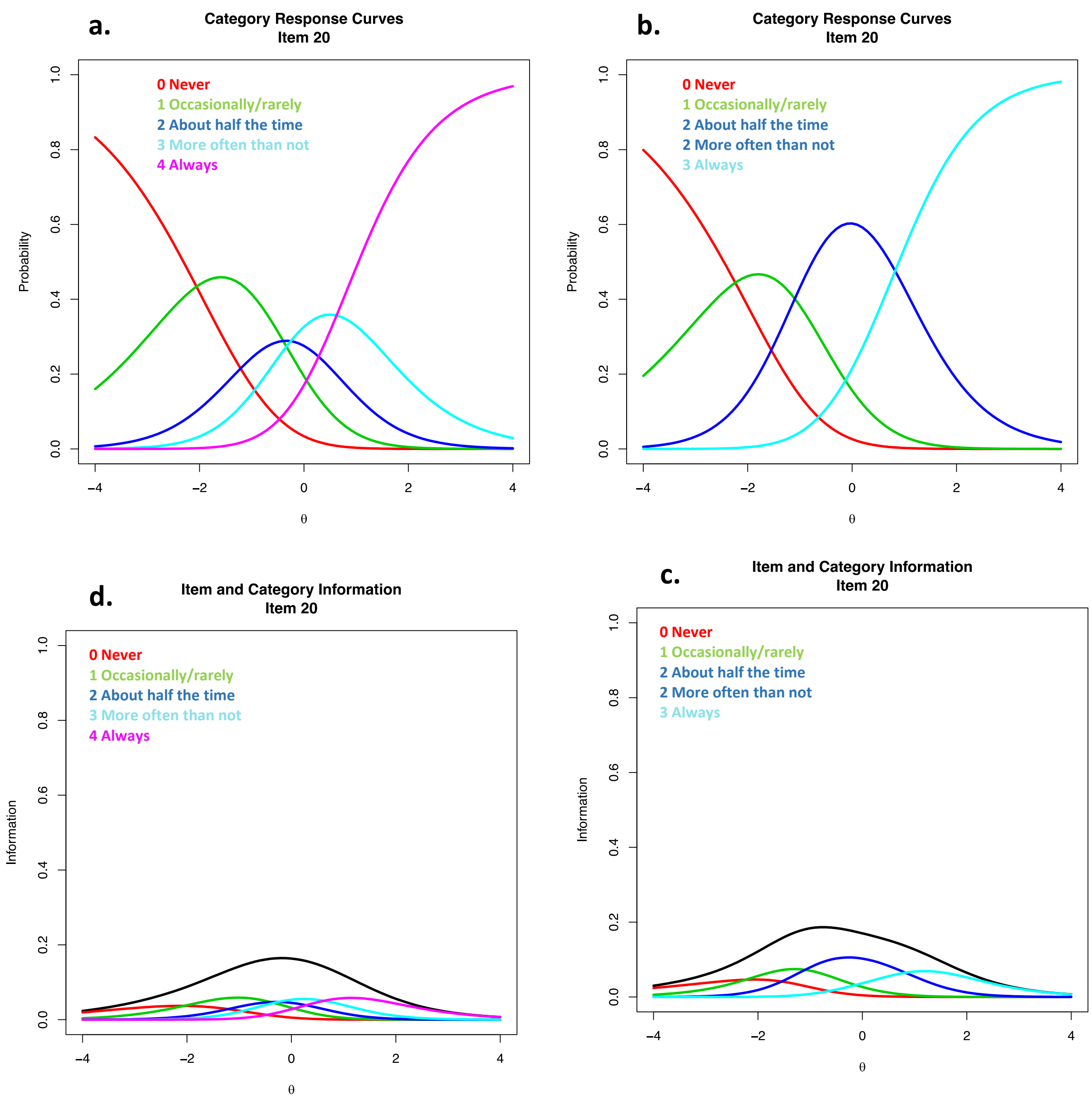


Figure 20. Factor 2 Item 20: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

APPENDIX F: Items With Problematic Parameter Estimates In The Presence Of Revised Options

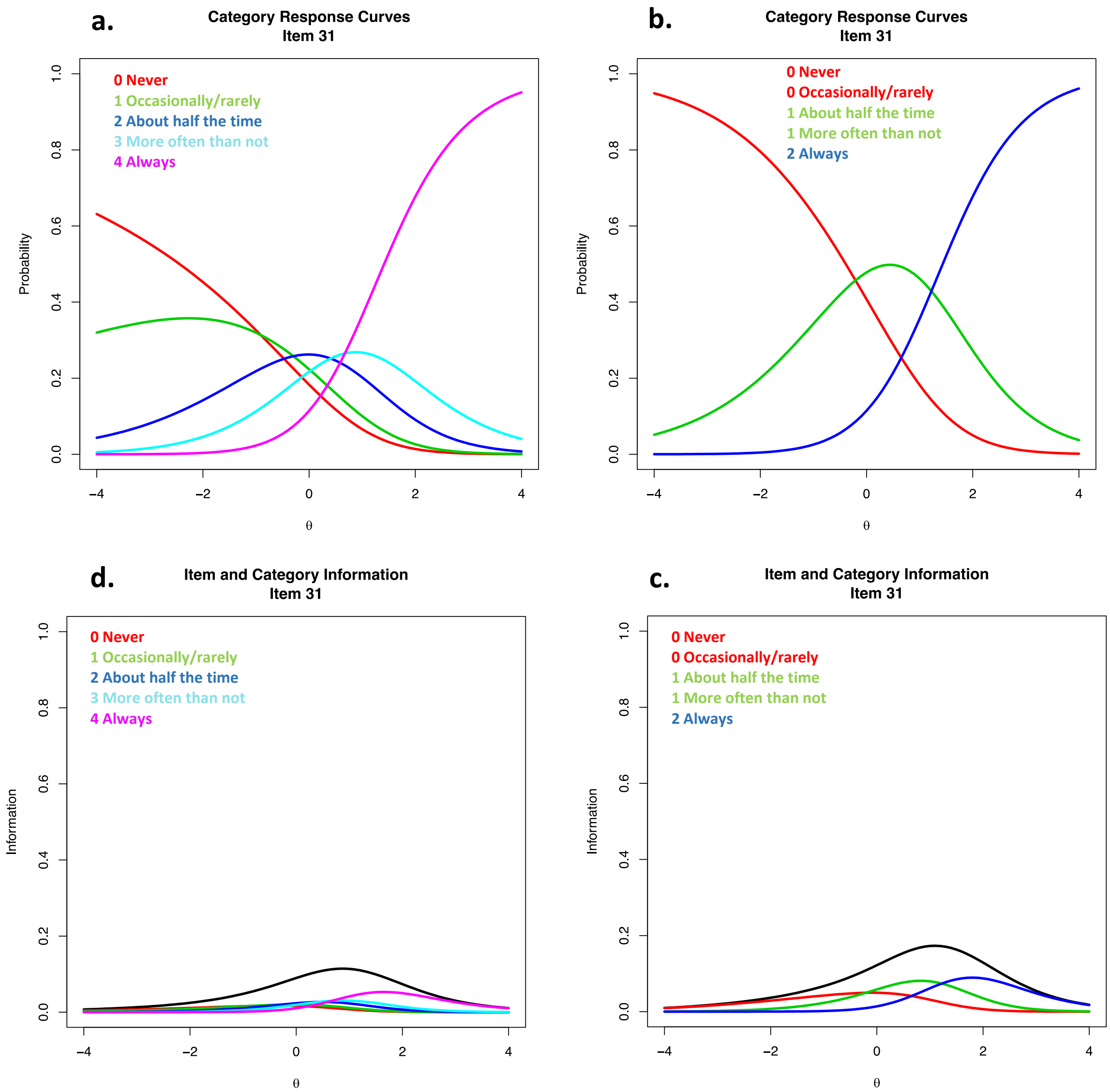


Figure 21. Factor 4 Item 31: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

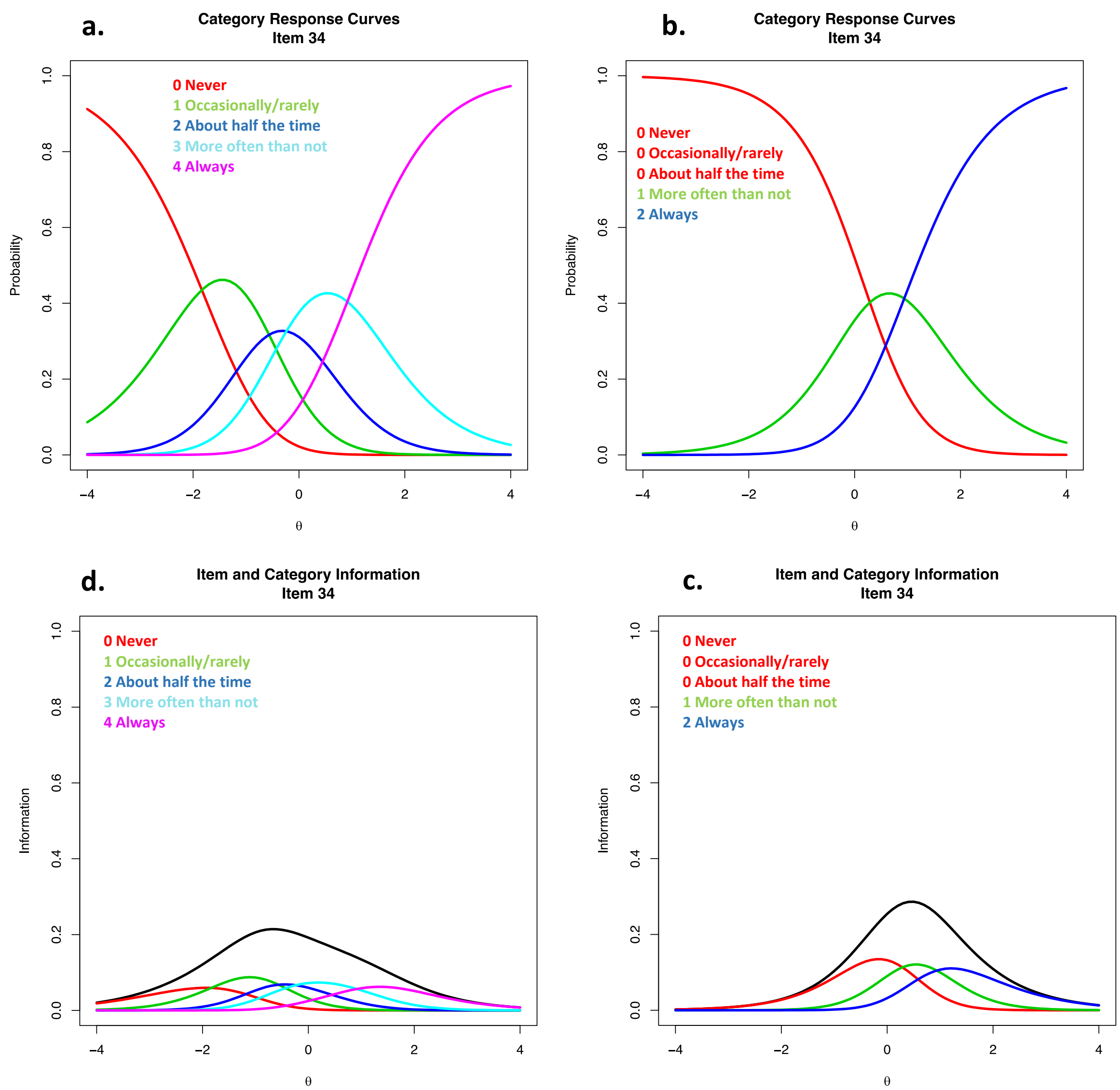


Figure 22. Factor 4 Item 34: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

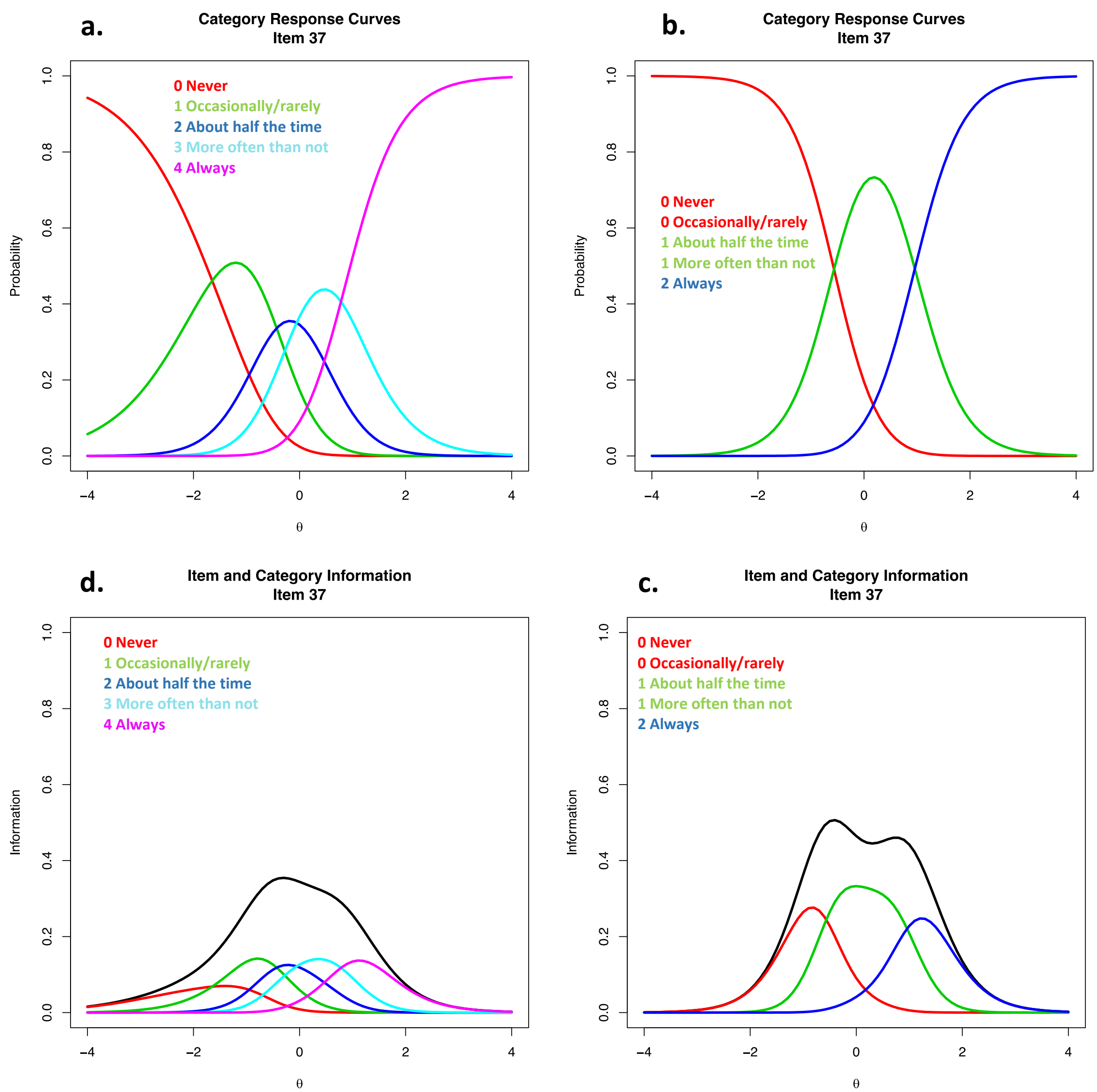


Figure 23. Factor 4 Item 37: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

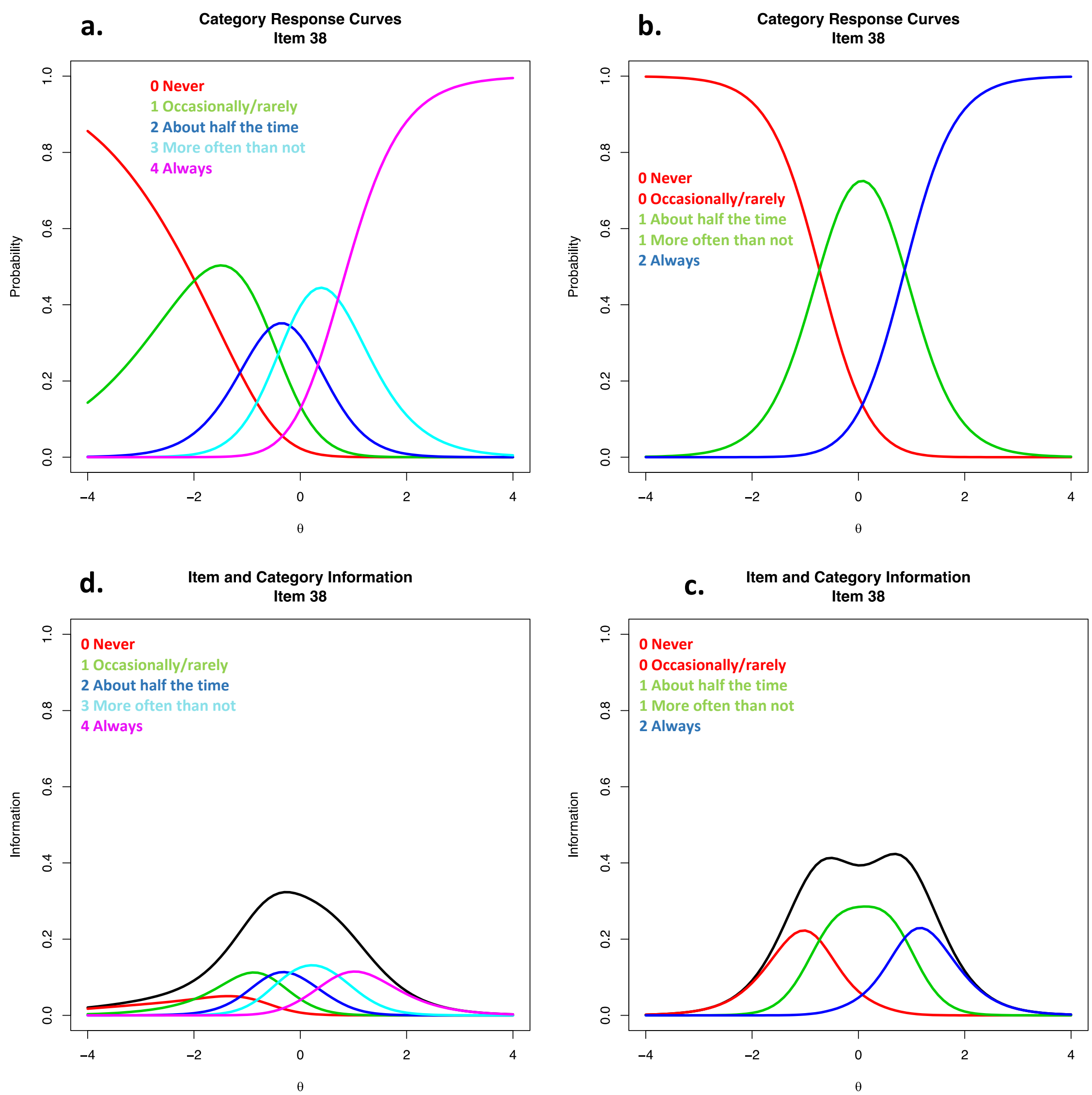


Figure 24. Factor 4 Item 38: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

APPENDIX G: Items That Did Not Improve With Rescoring, And Are In Need Of Revision

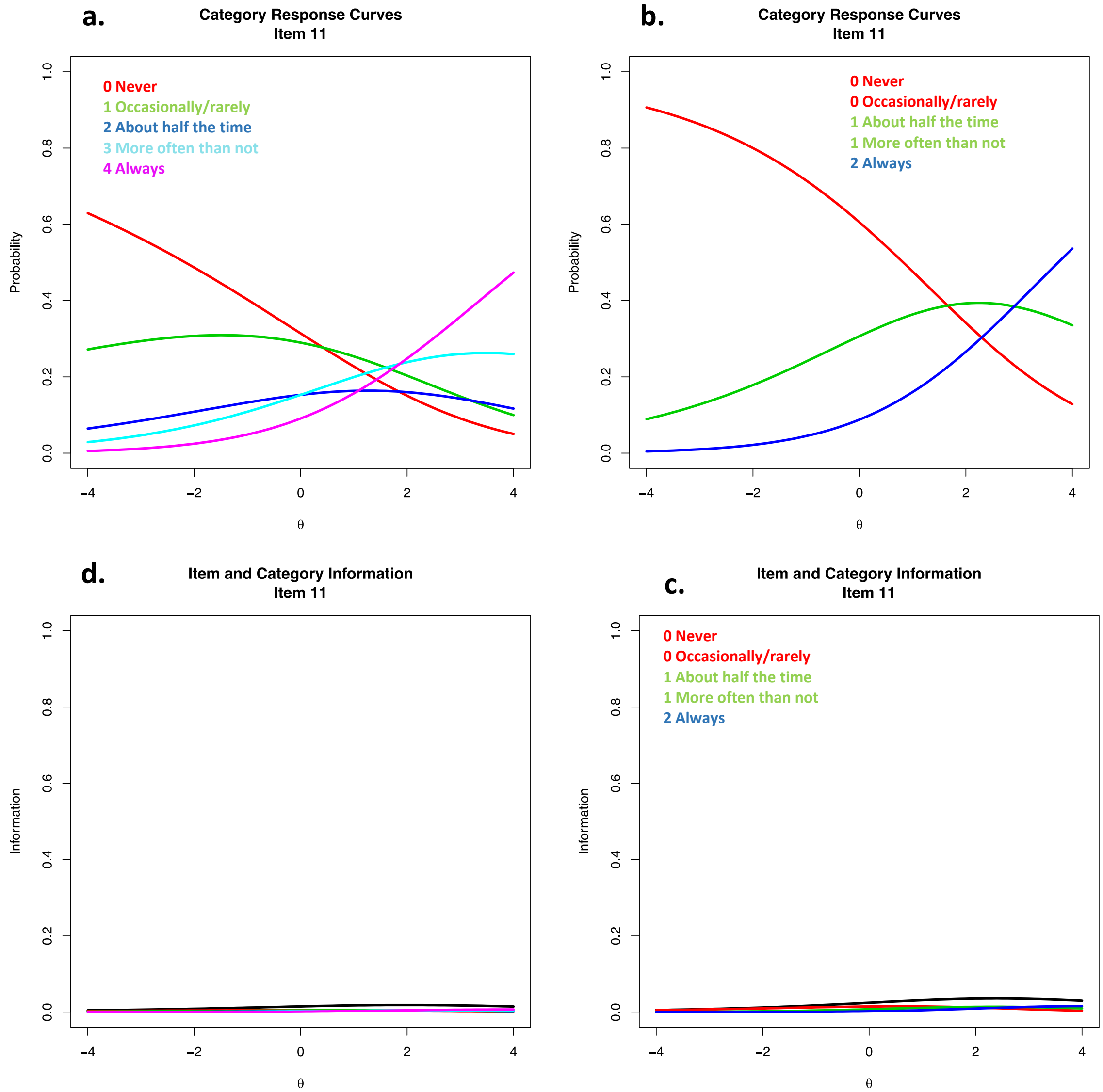


Figure 25. Factor 1 Item 11: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

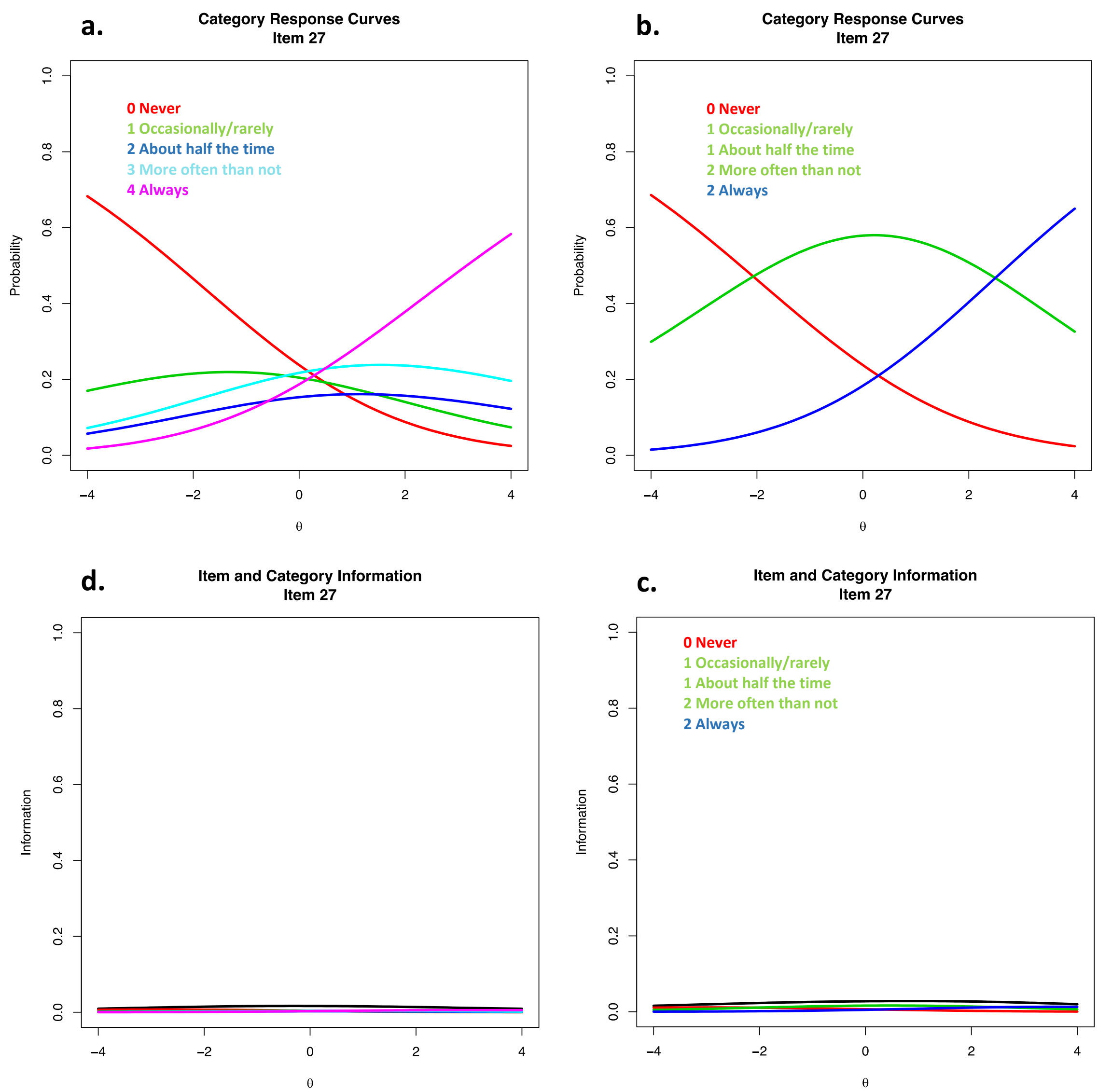


Figure 26. Factor 3 Item 27: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

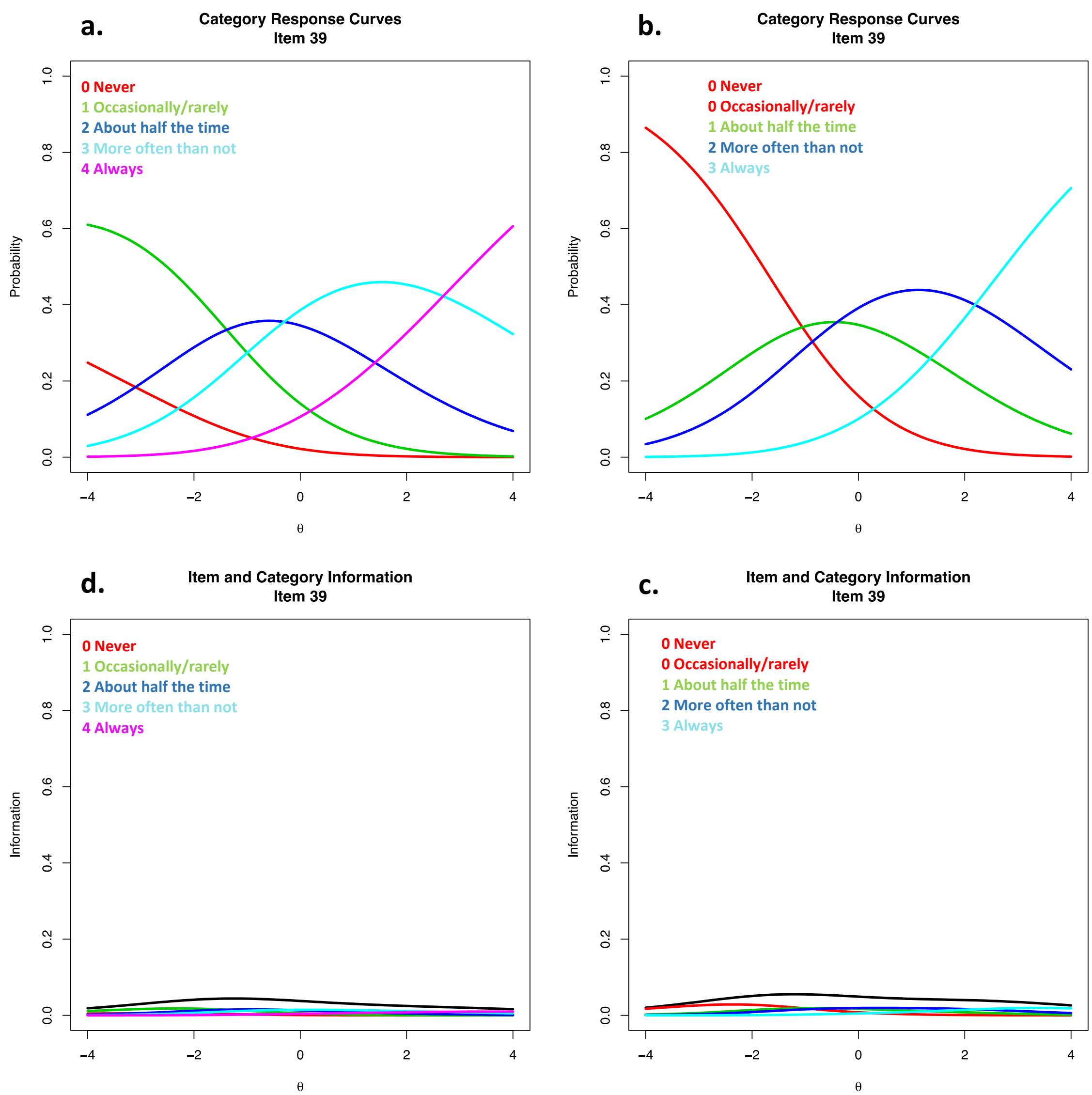


Figure 27. Factor 4 Item 39: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

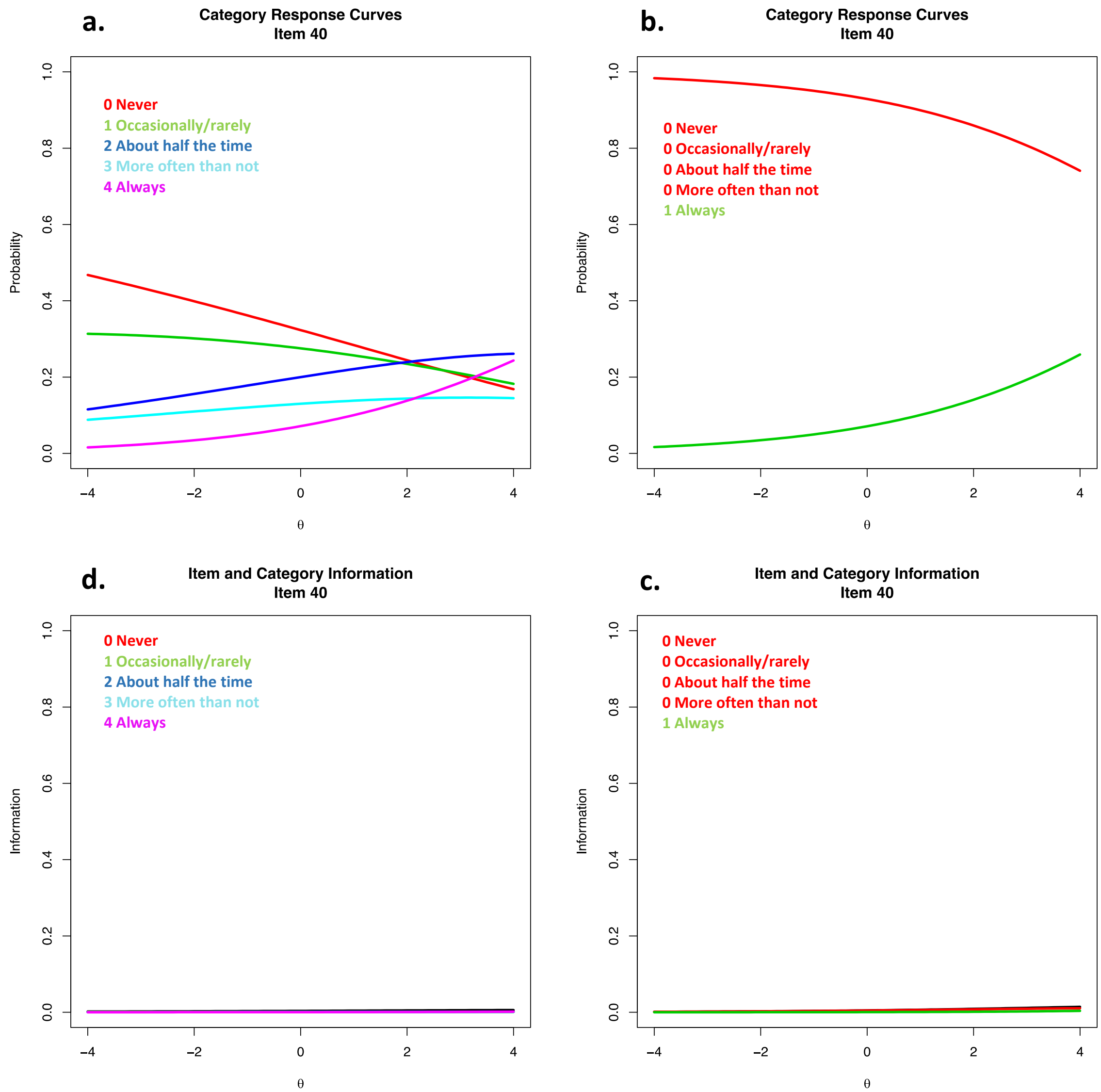


Figure 28. Factor 4 Item 40: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

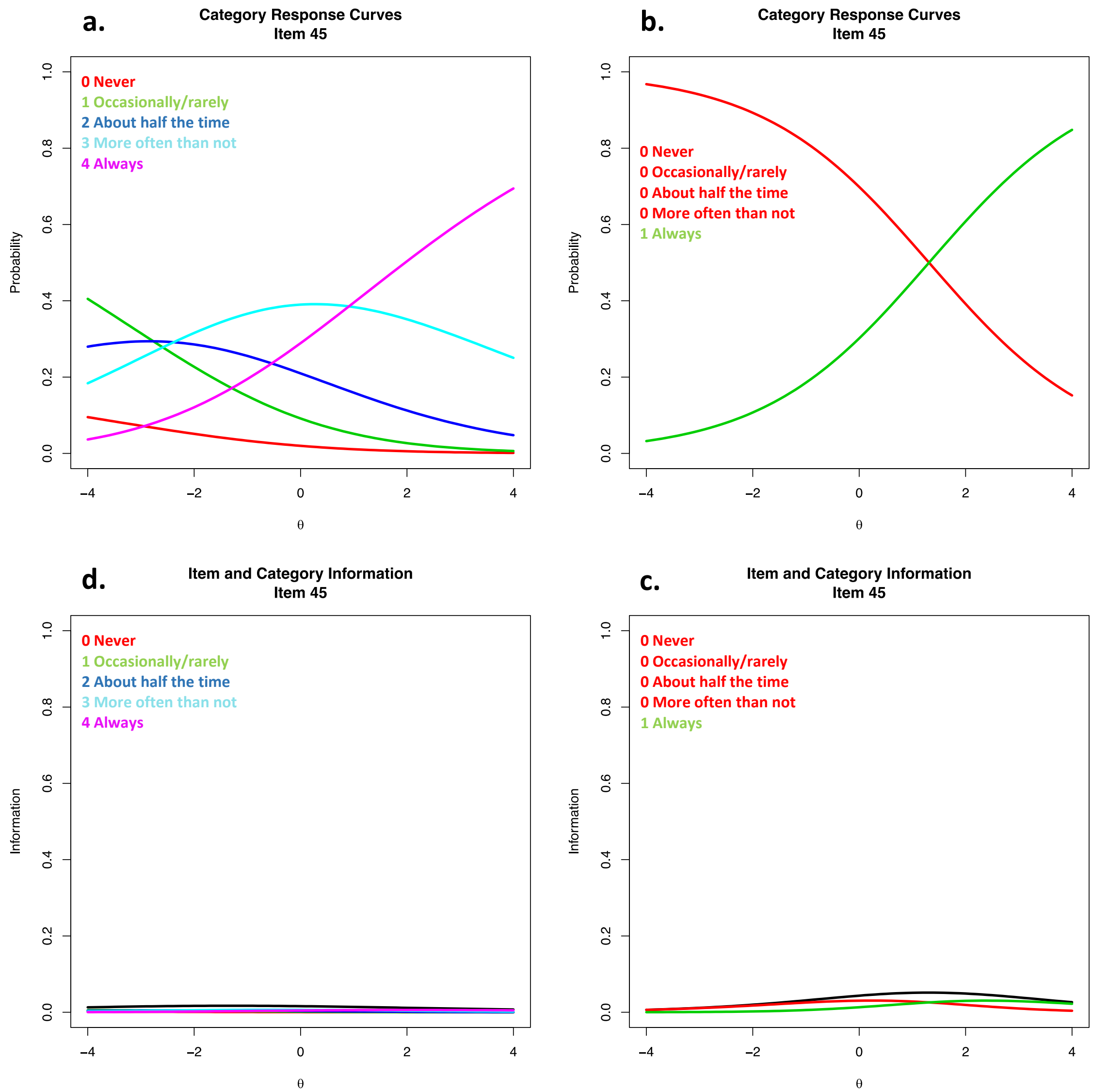


Figure 29. Factor 5 Item 45: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

APPENDIX H: Properly Functioning Items That Were Not Rescored

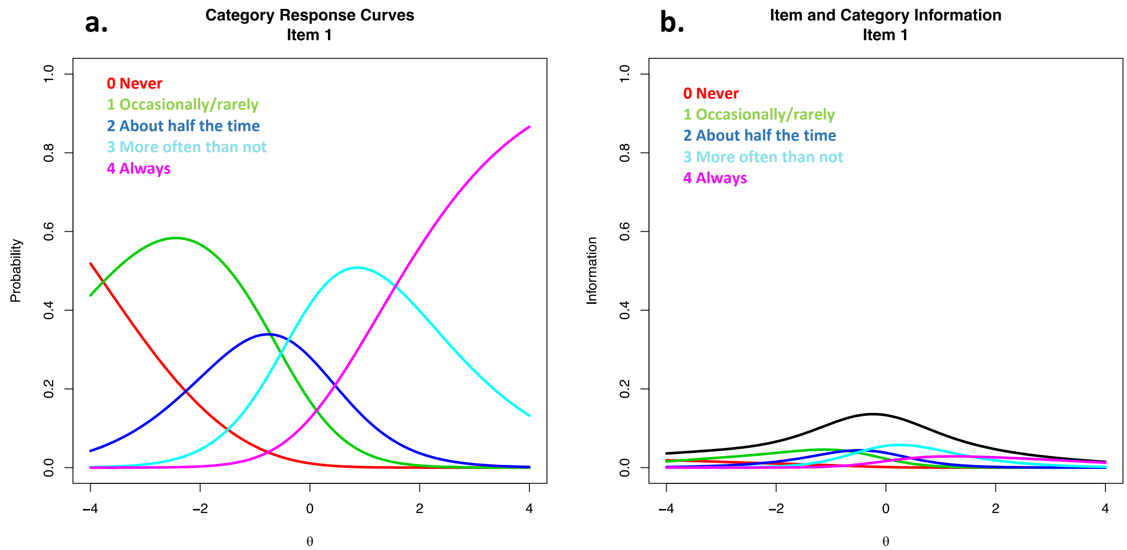


Figure 30. Factor 1 Item 1: Category Response Curves (a) and Item Category Information Functions (b).

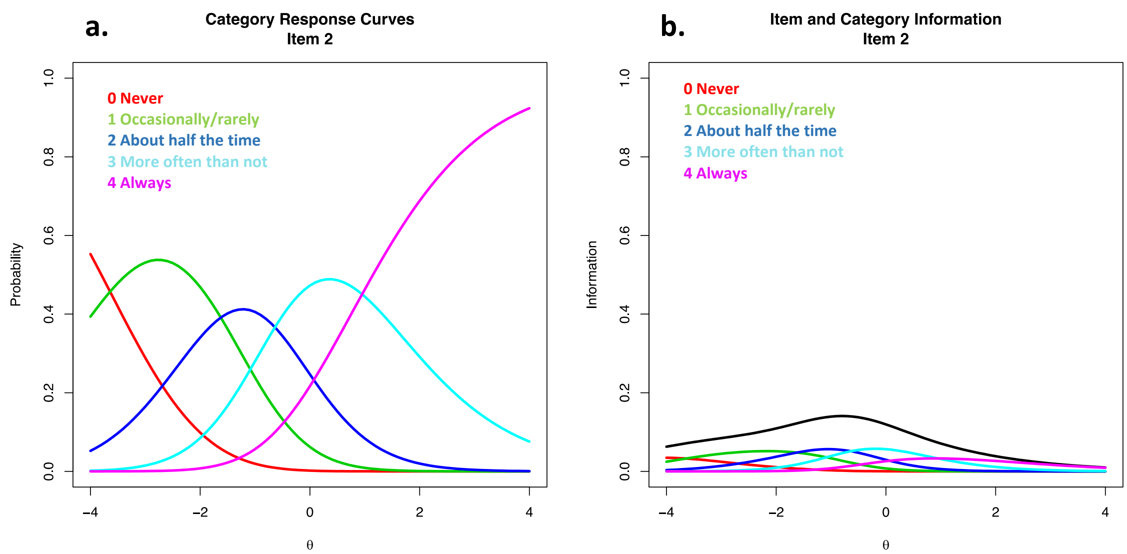


Figure 31. Factor 1 Item 2: Category Response Curves (a) and Item Category Information Functions (b).

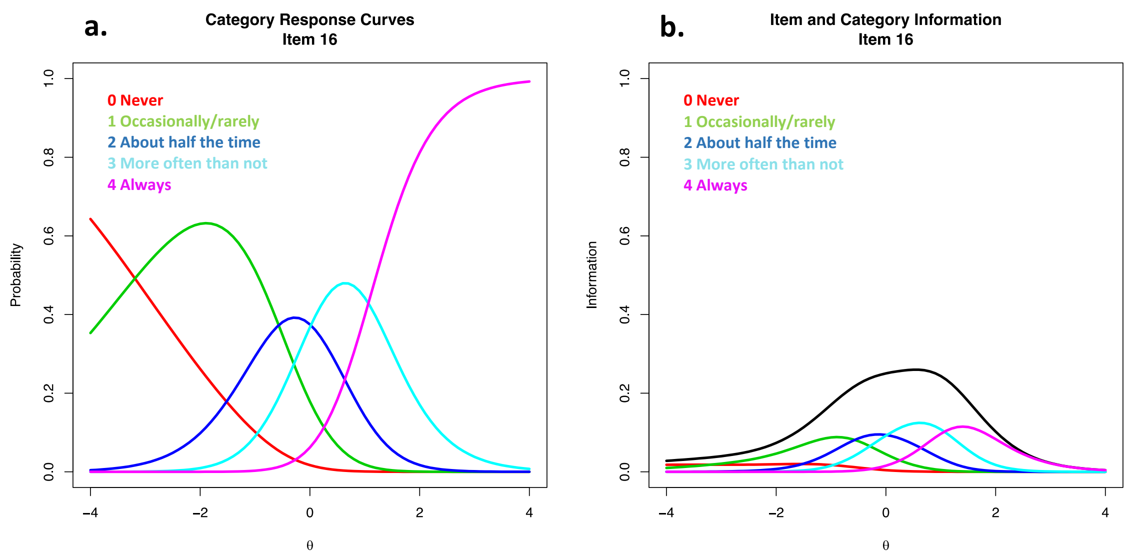


Figure 32. Factor 2 Item 16: Category Response Curves (a) and Item Category Information Functions (b).

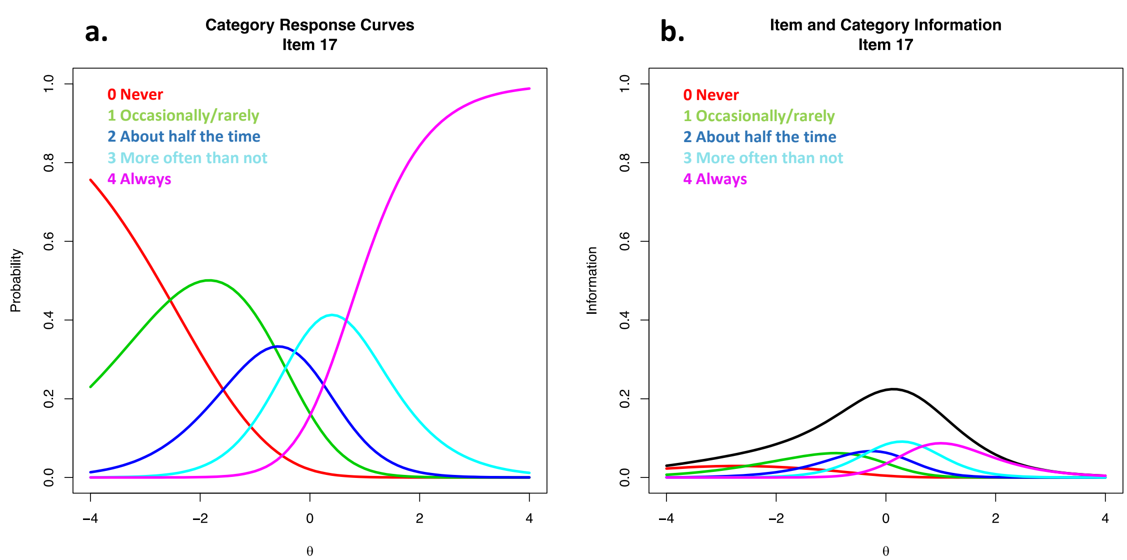


Figure 33. Factor 2 Item 17: Category Response Curves (a) and Item Category Information Functions (b).

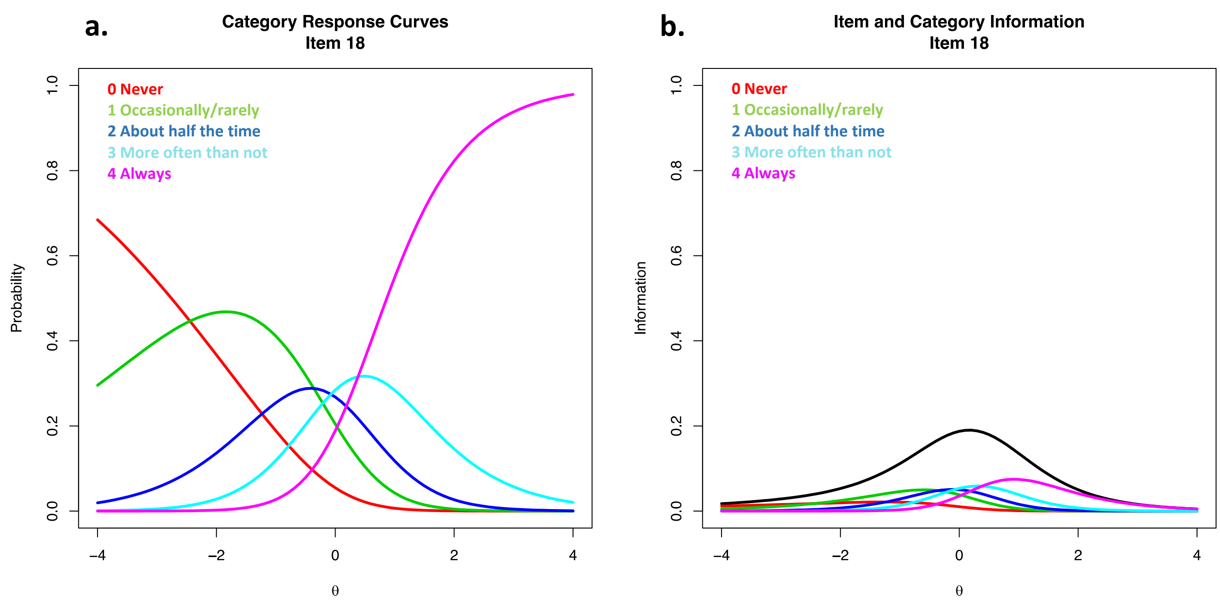


Figure 34. Factor 2 Item 18: Category Response Curves (a) and Item Category Information Functions (b).

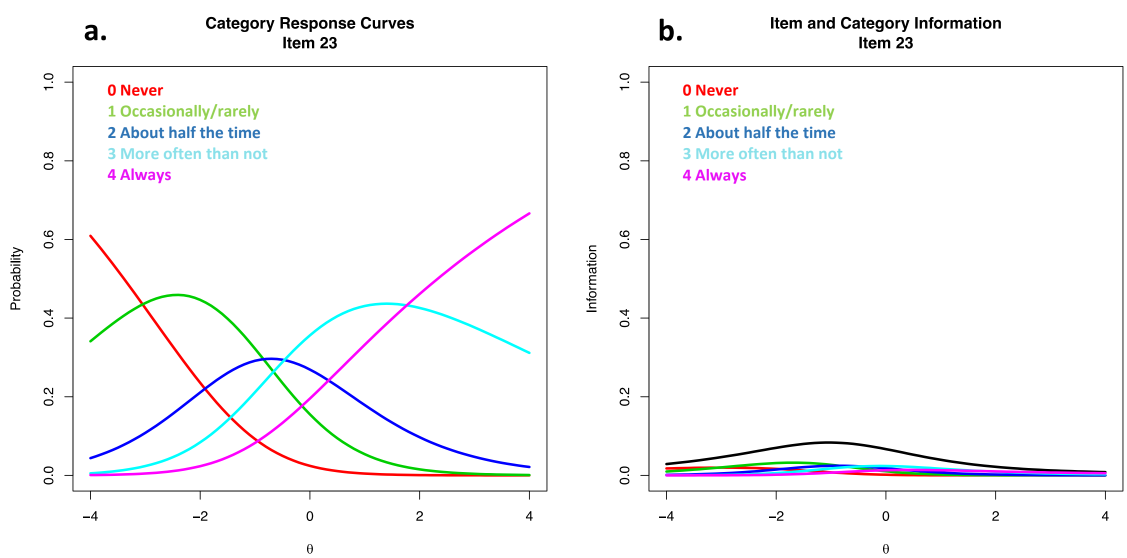


Figure 35. Factor 3 Item 23: Category Response Curves (a) and Item Category Information Functions (b).

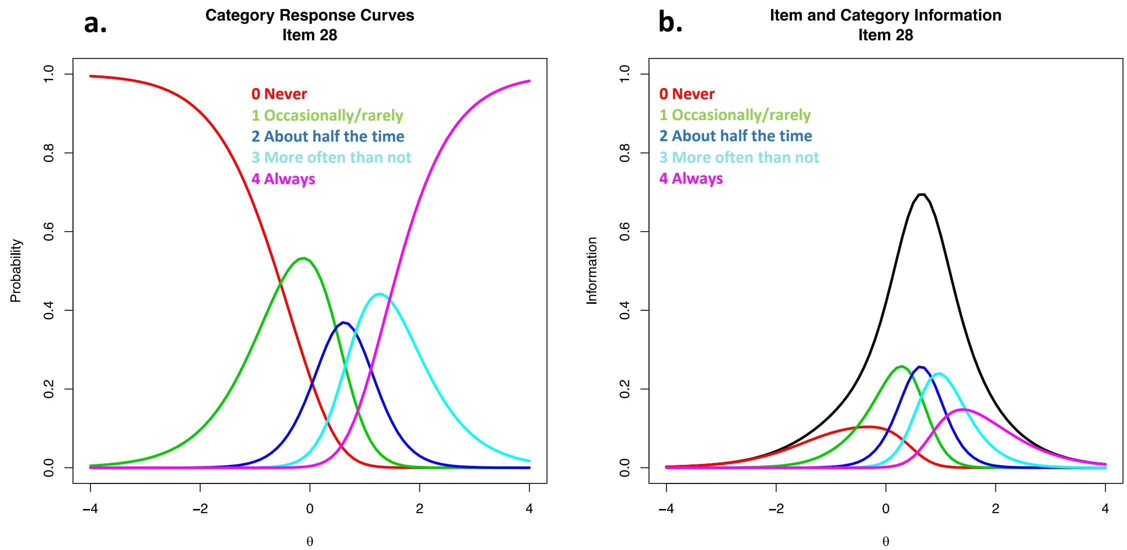


Figure 36. Factor 3 Item 28: Category Response Curves (a) and Item Category Information Functions (b).

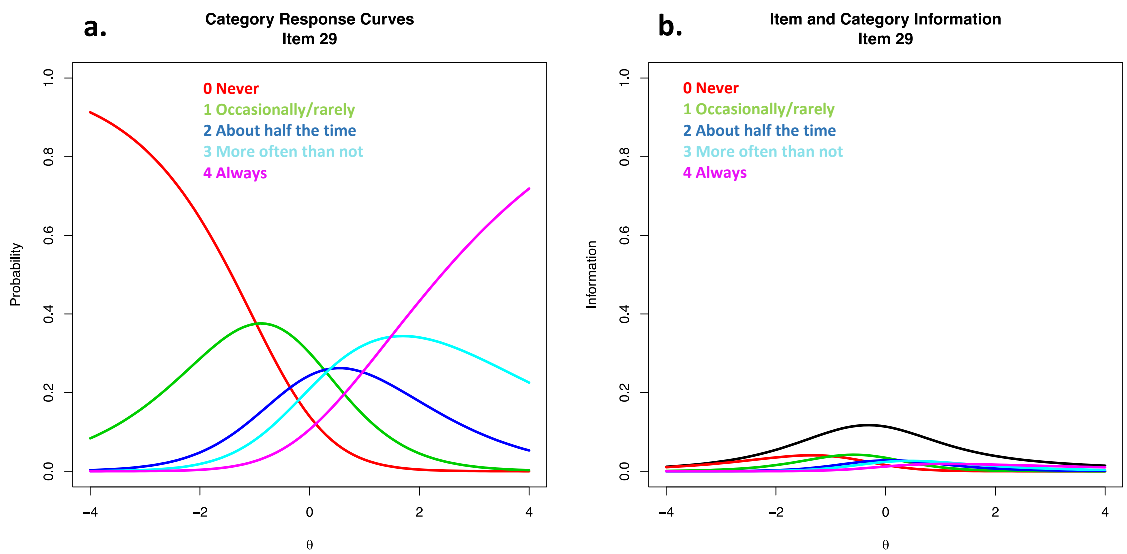


Figure 37. Factor 3 Item 29: Category Response Curves (a), Item Category Information Functions (b), and original vs. rescored Item Information Curves (c).

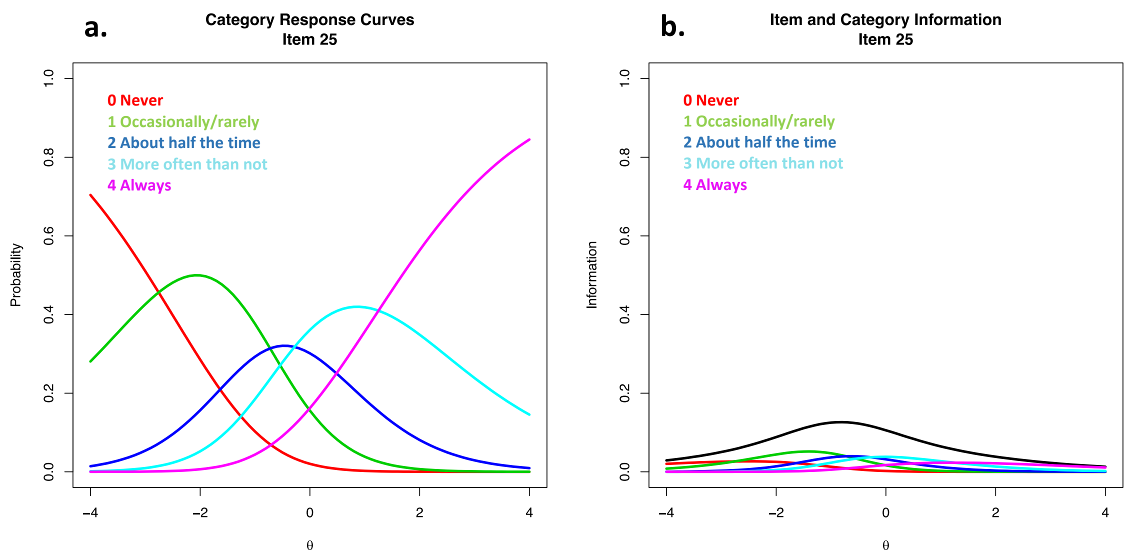


Figure 38. Factor 4 Item 25: Category Response Curves (a) and Item Category Information Functions (b).

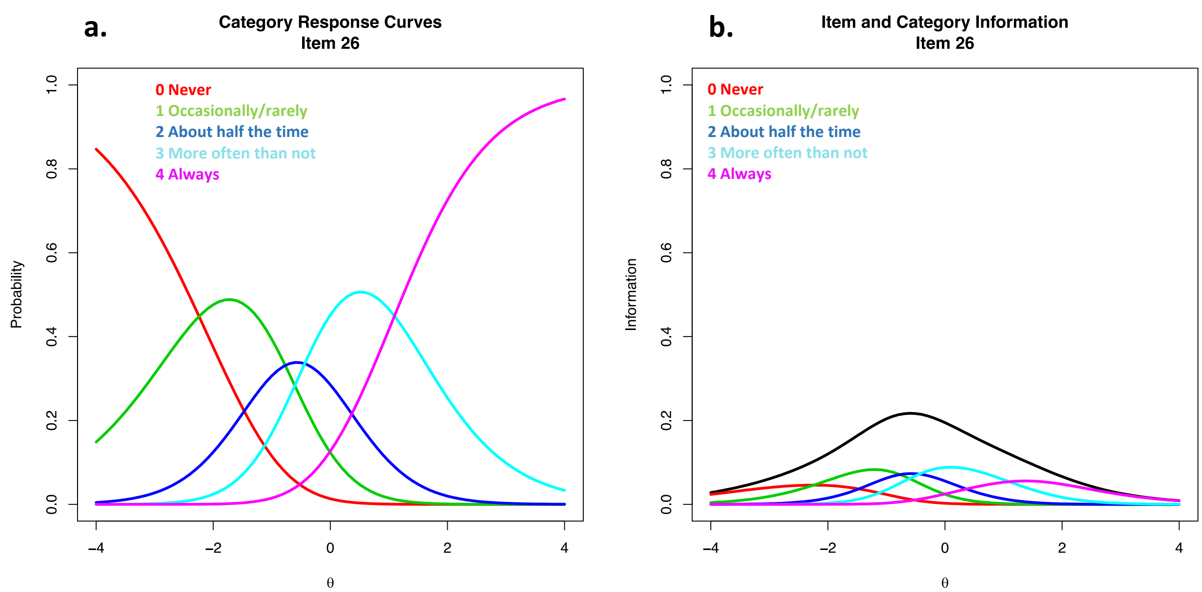


Figure 39. Factor 4 Item 26: Category Response Curves (a) and Item Category Information Functions (b).

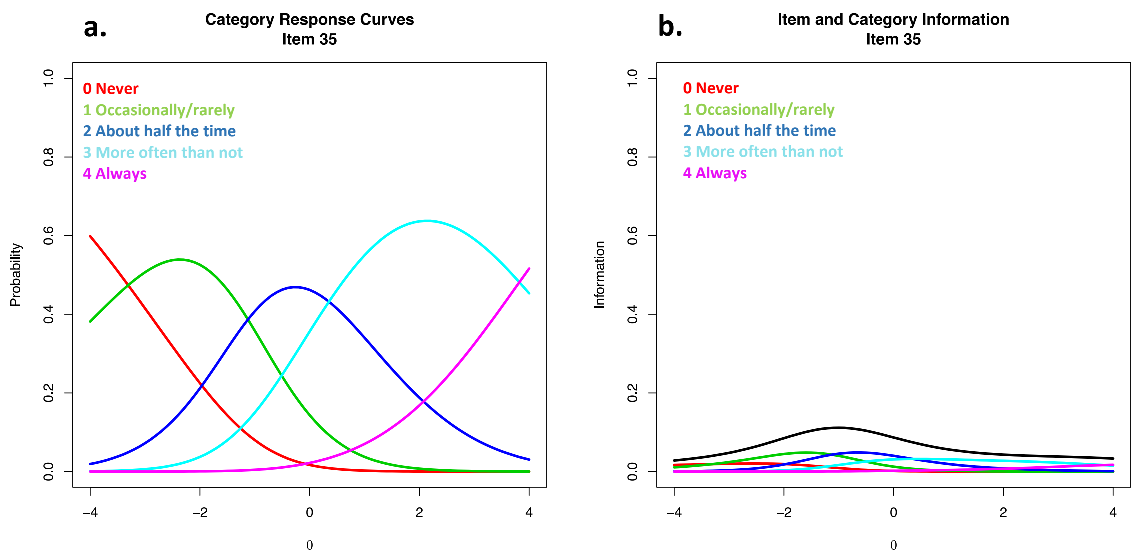


Figure 40. Factor 4 Item 35: Category Response Curves (a) and Item Category Information Functions (b).

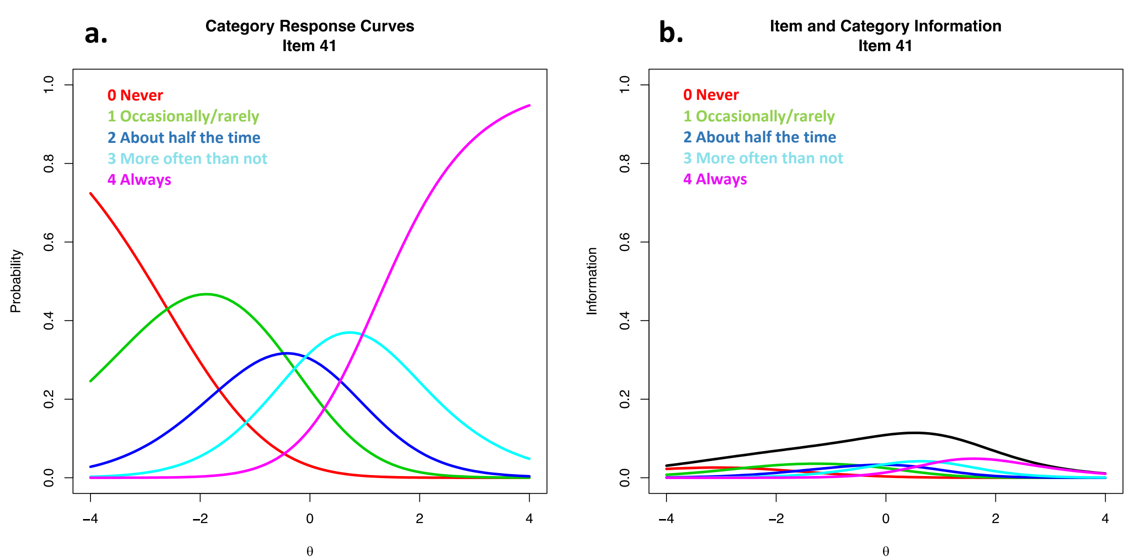


Figure 41. Factor 5 Item 41: Category Response Curves (a) and Item Category Information Functions (b).

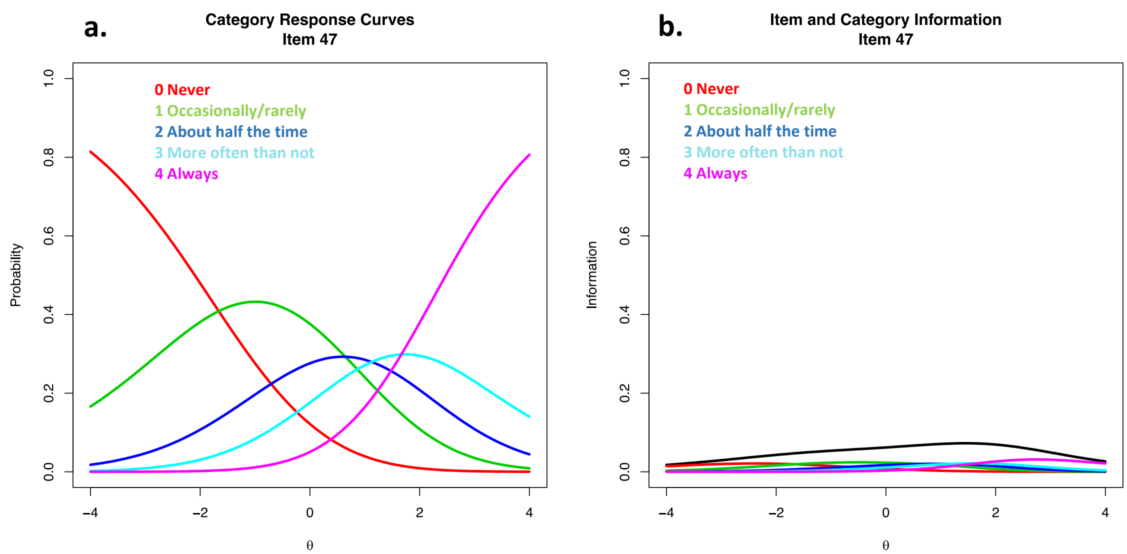


Figure 42. Factor 5 Item 47: Category Response Curves (a), Item Category Information Functions (b), and original vs. rescored Item Information Curves (c).

APPENDIX I: Additional Category Response and Item Information Curves

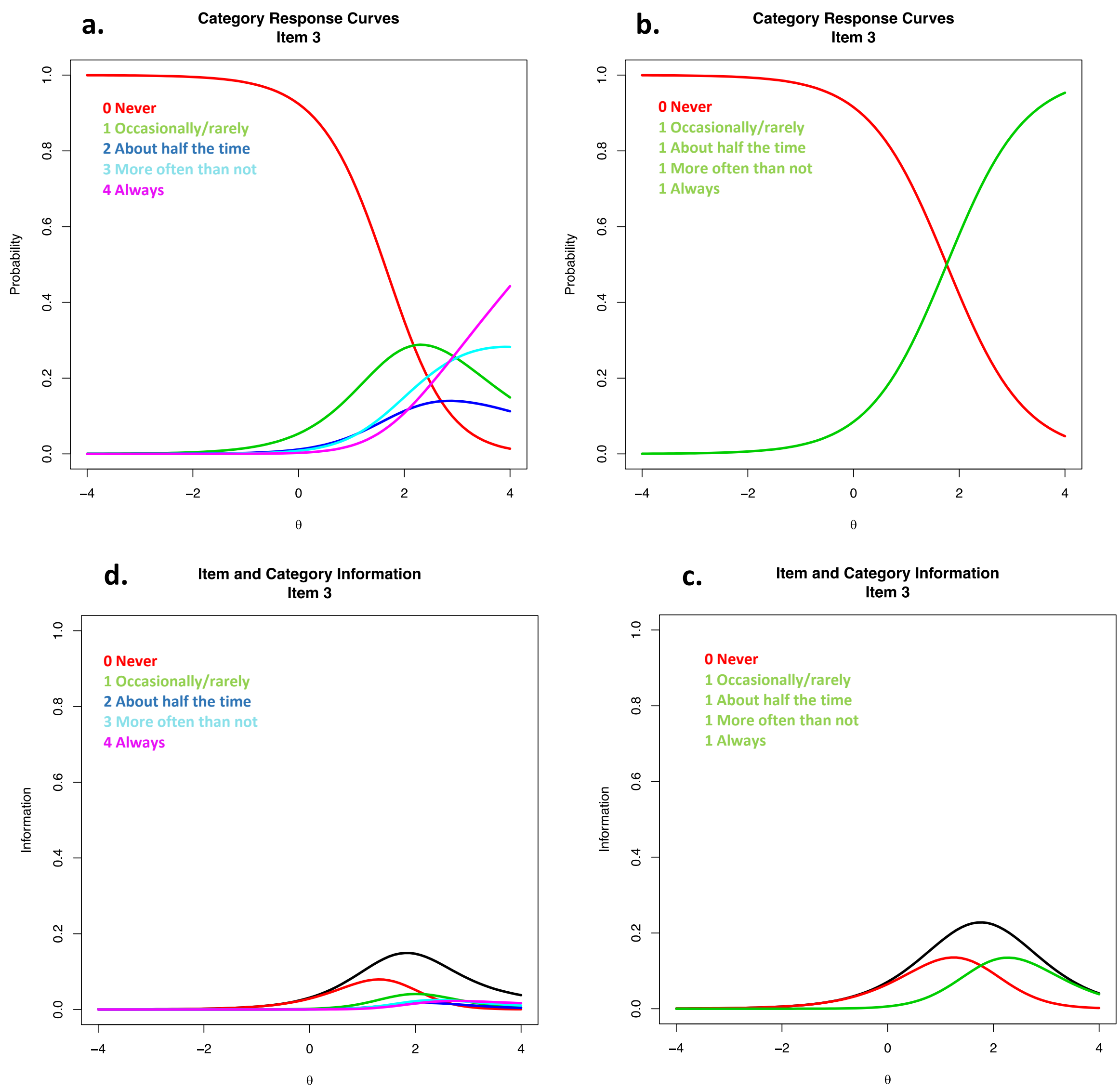


Figure 43. Factor 1 Item 3: Original Category Response Curves (a), rescored category response curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

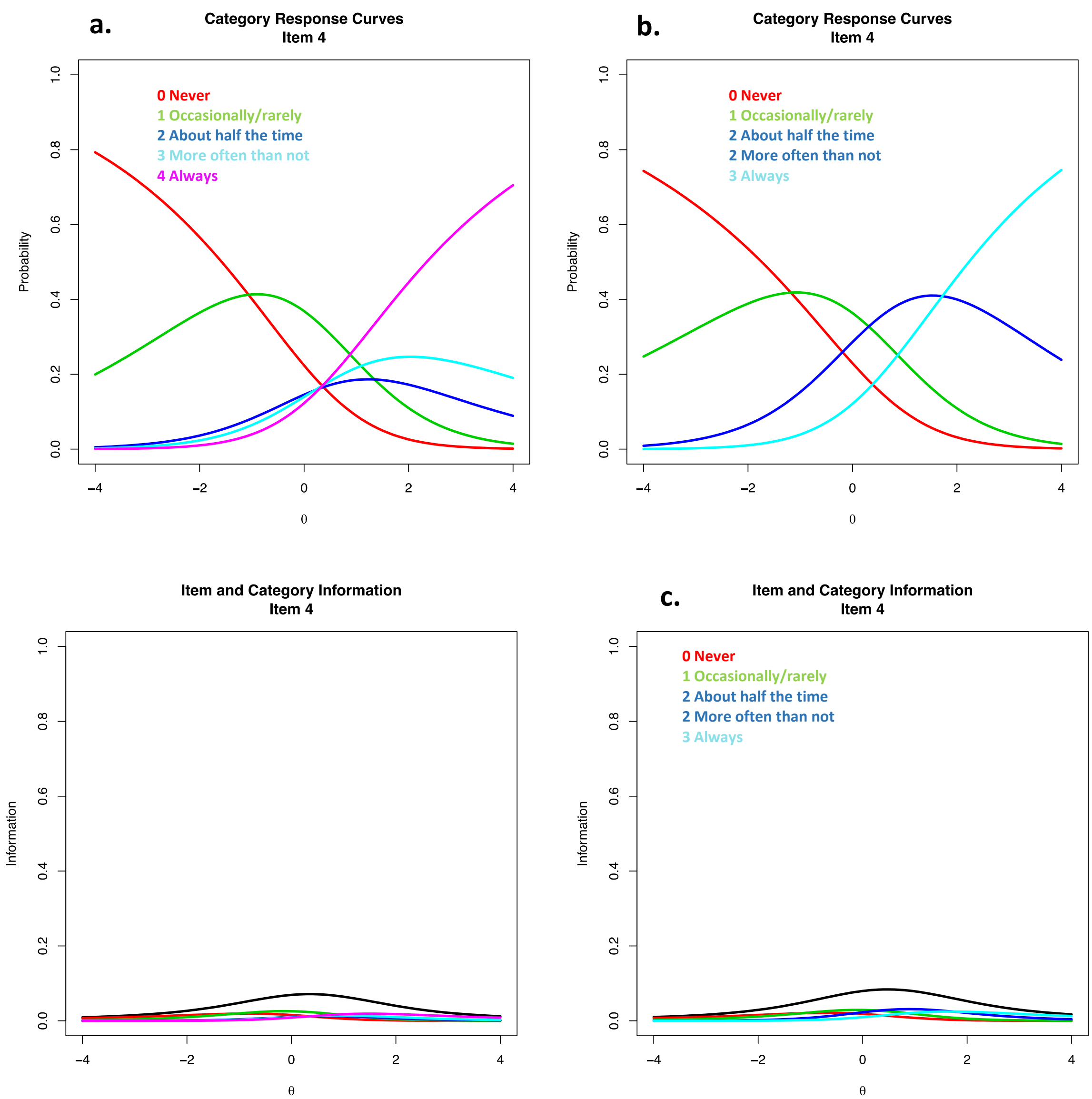


Figure 44. Factor 1 Item 4: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

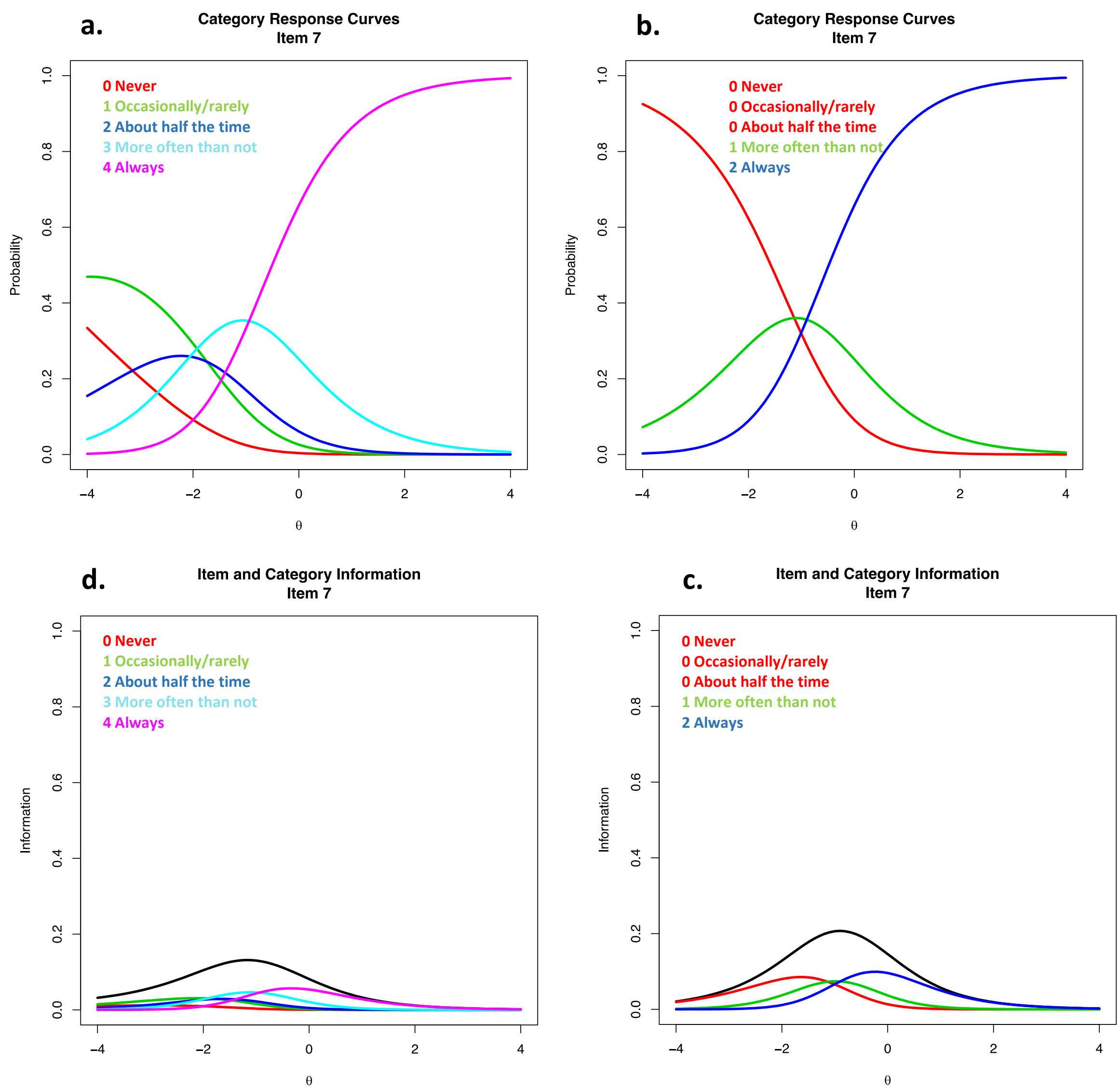


Figure 45. Factor 1 Item 7: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

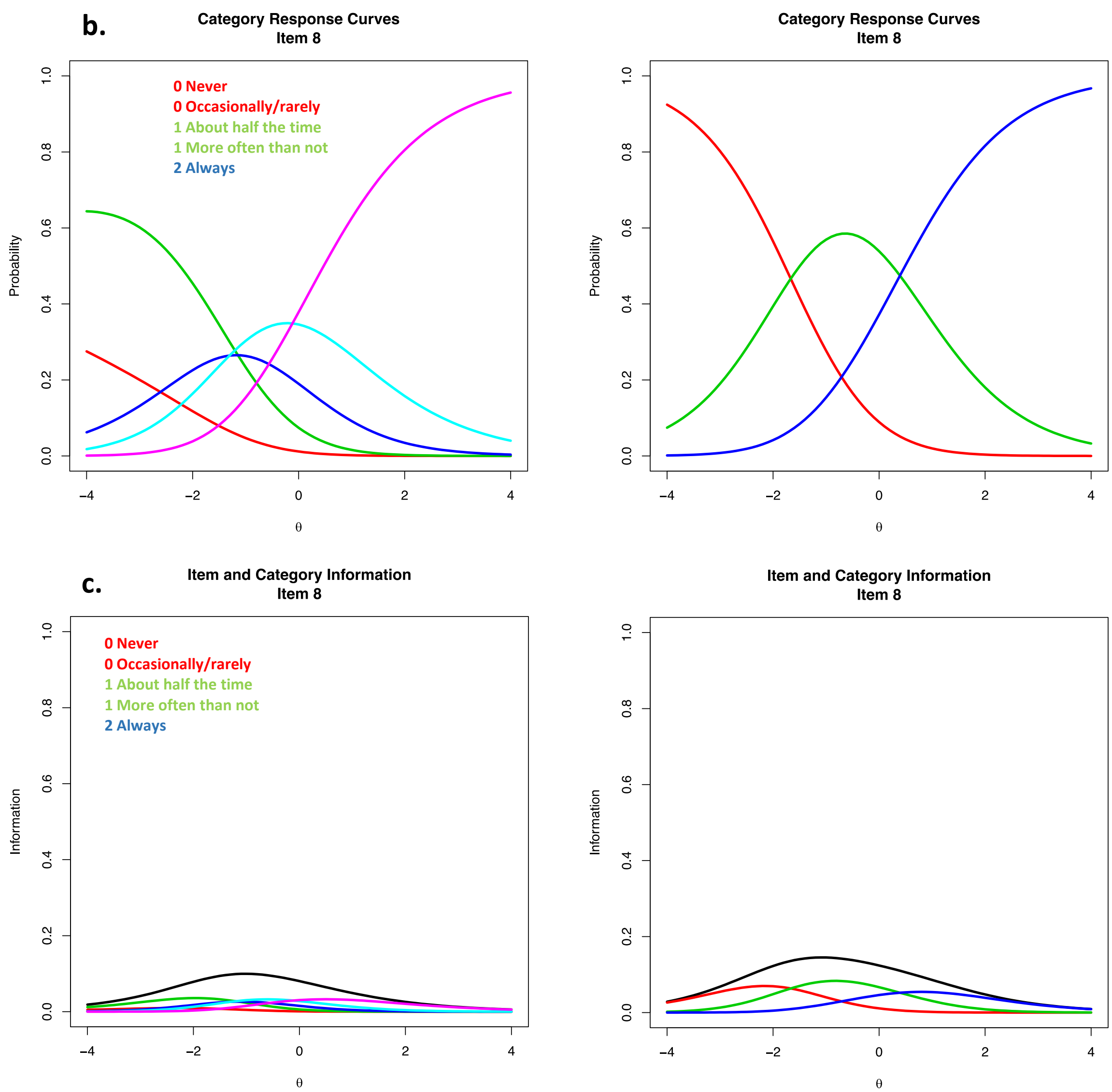


Figure 46. Factor 1 Item 8: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

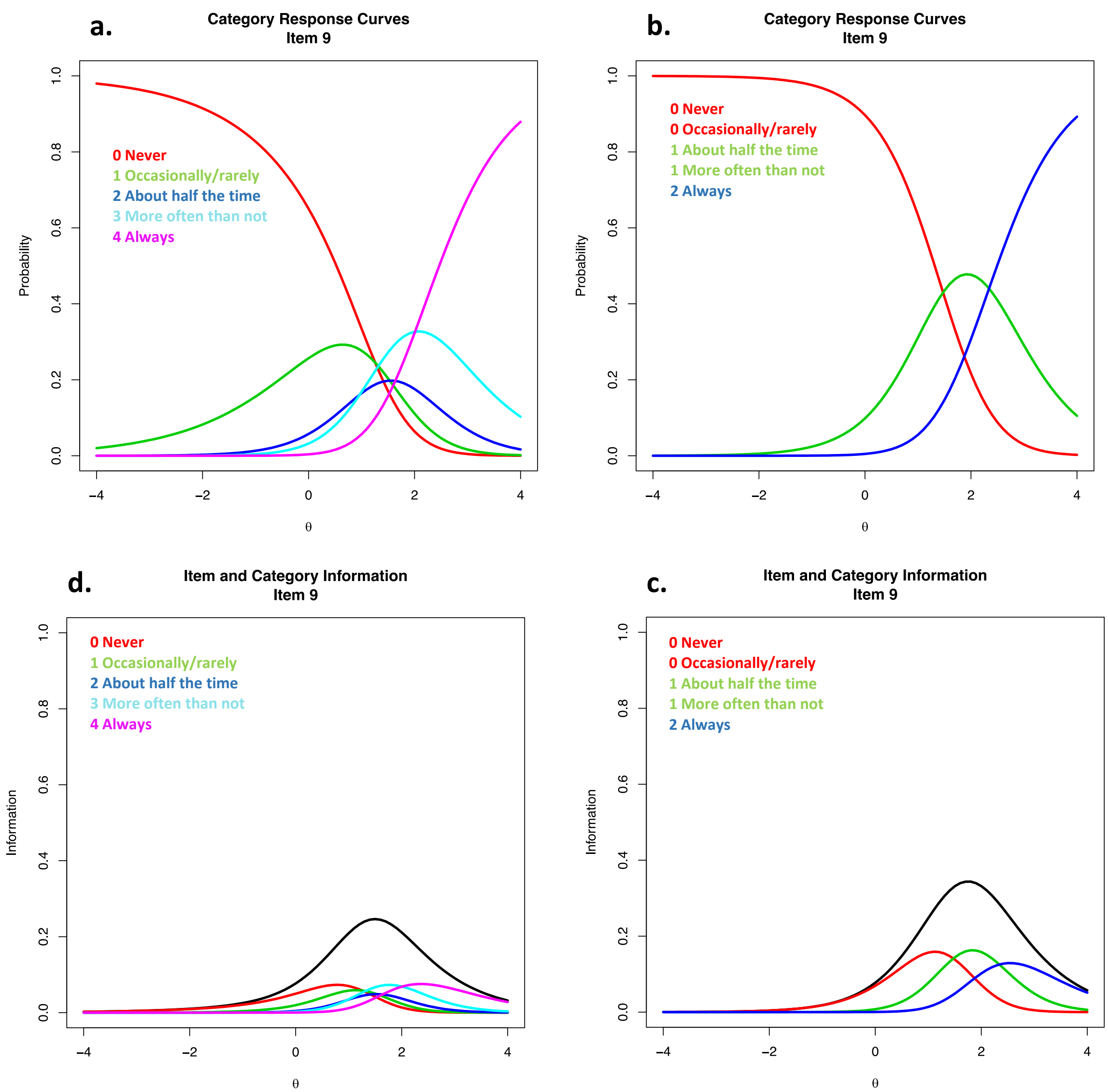


Figure 47. Factor 1 Item 9: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

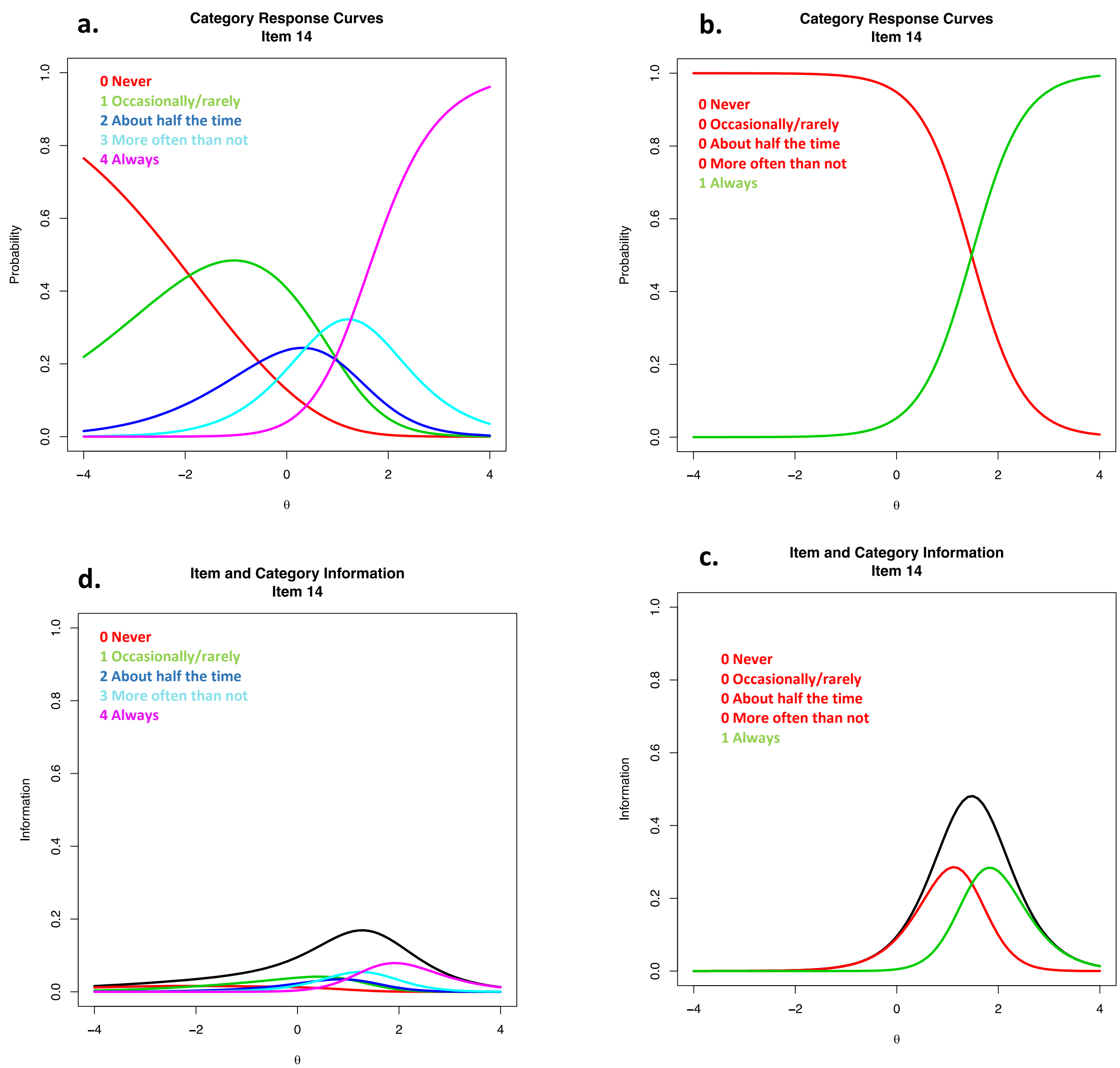


Figure 48. Factor 2 Item 14: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

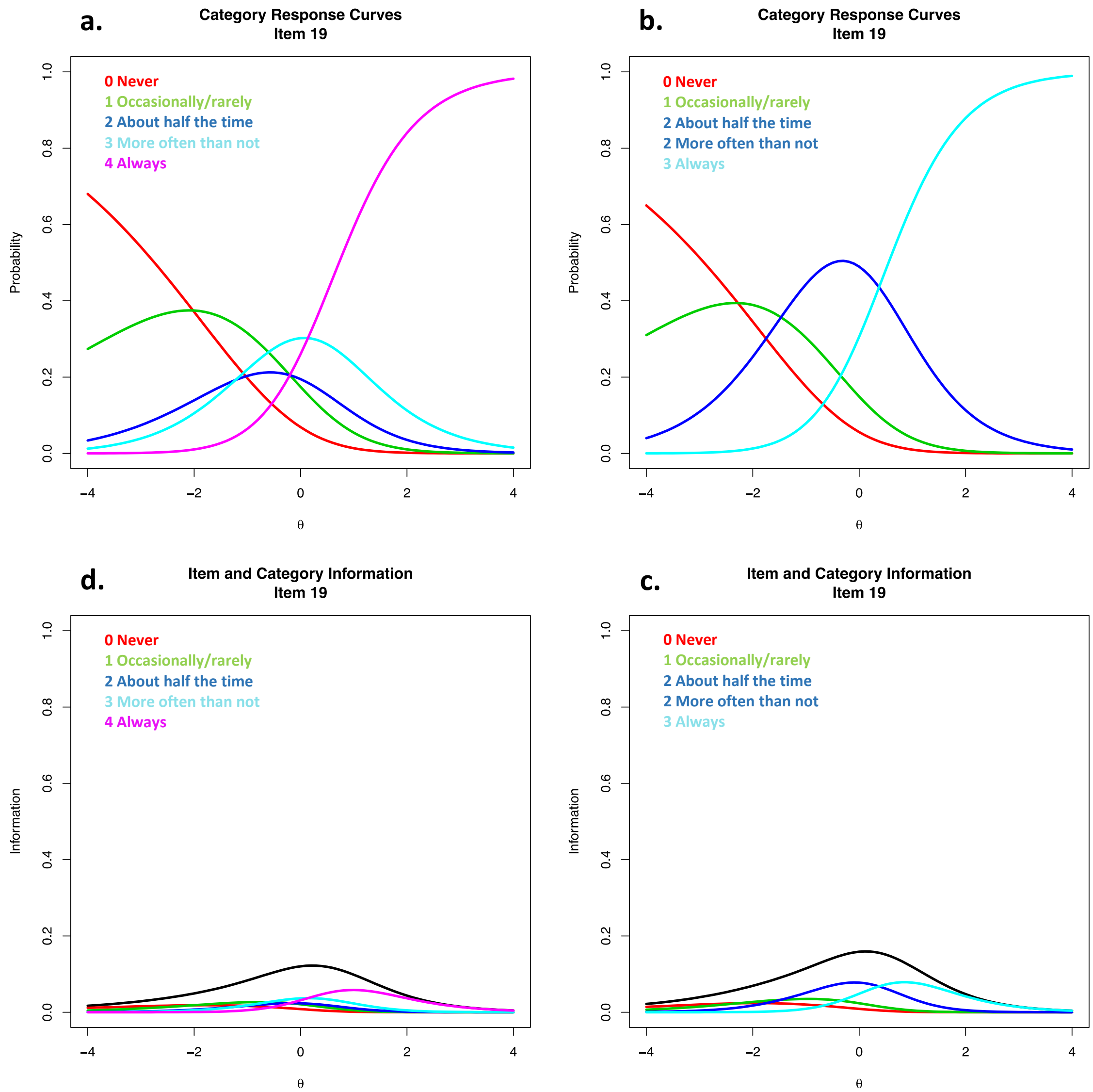


Figure 49. Factor 2 Item 19: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

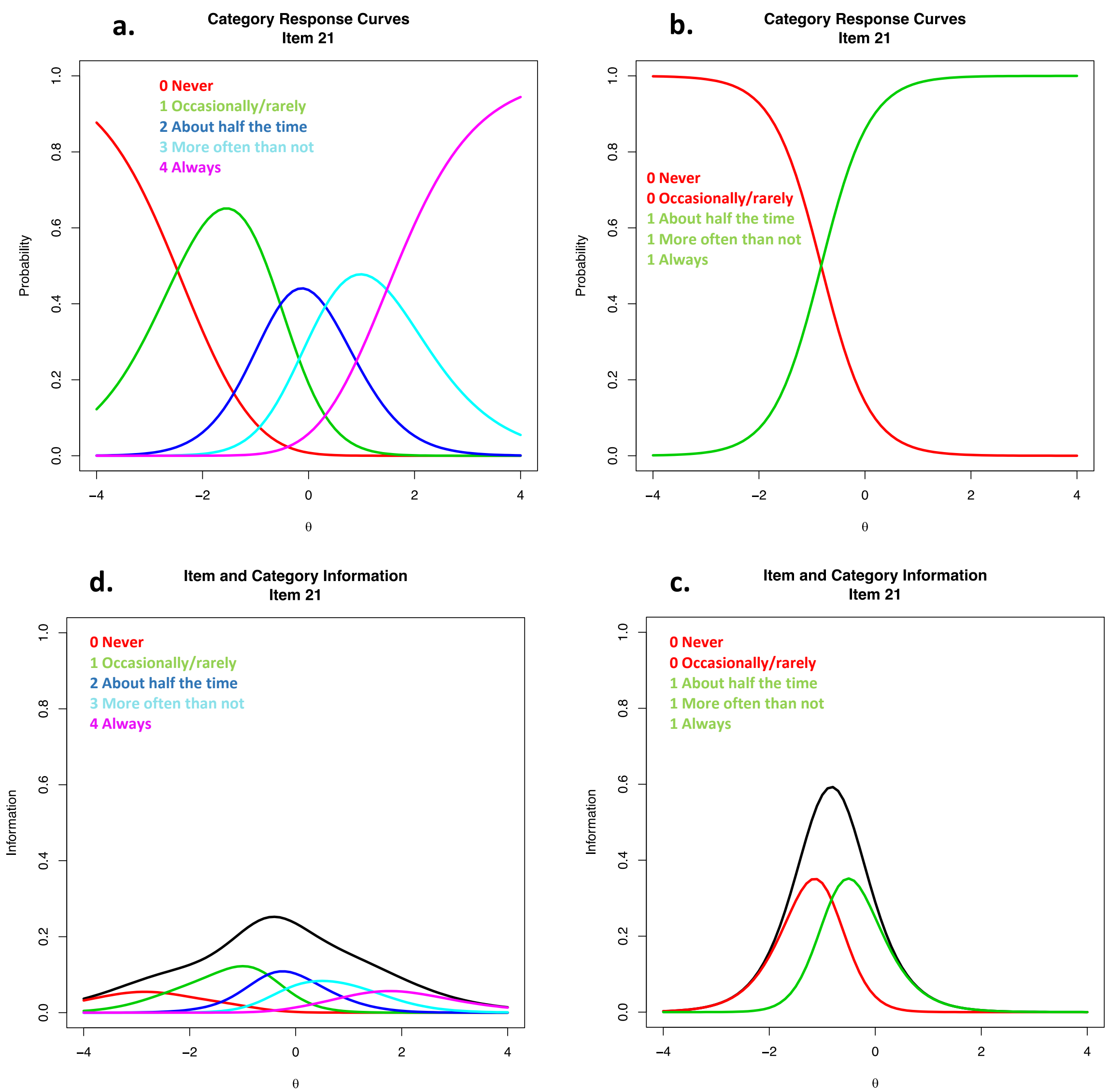


Figure 50. Factor 2 Item 21: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

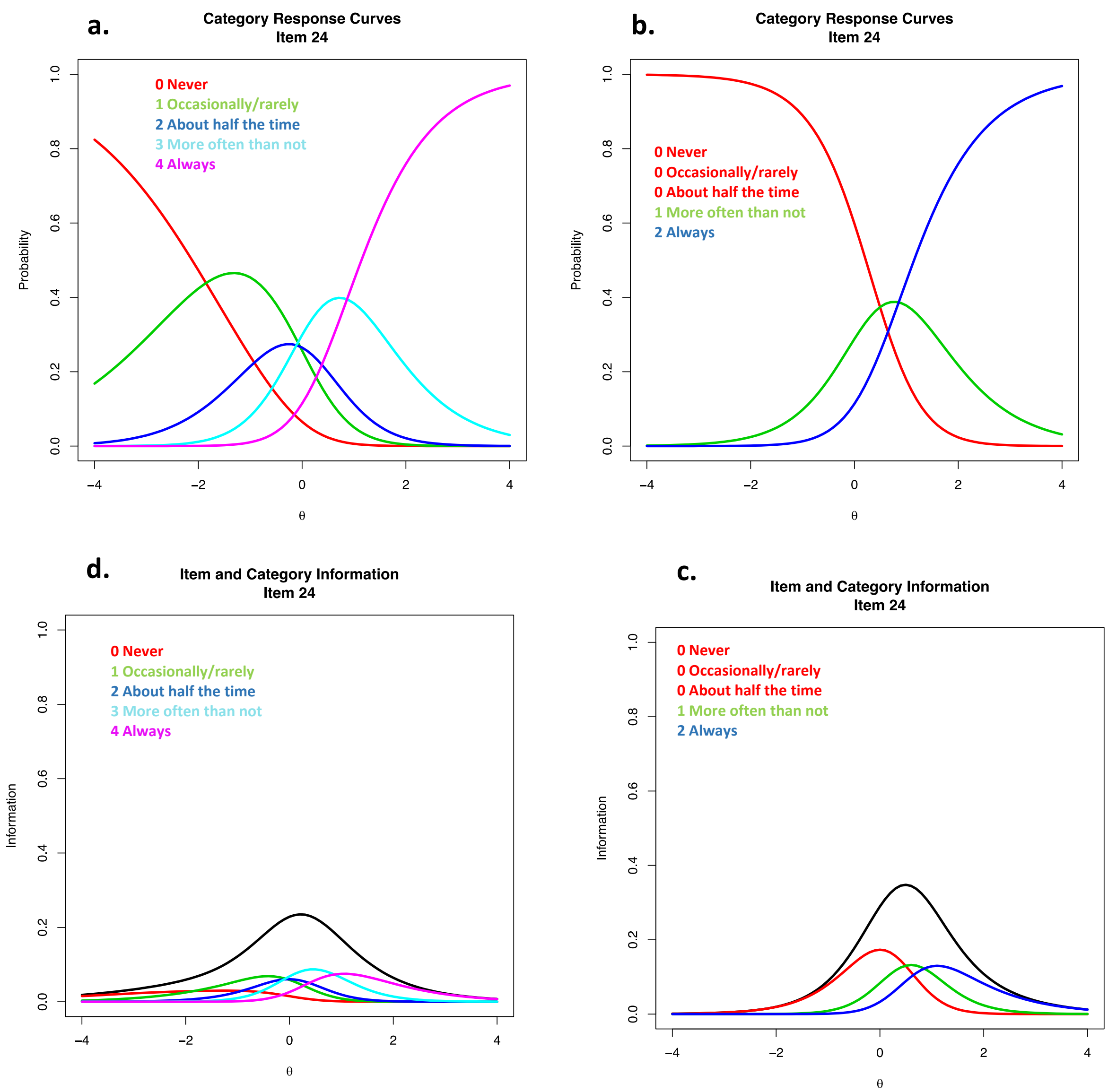


Figure 51. Factor 3 Item 24: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

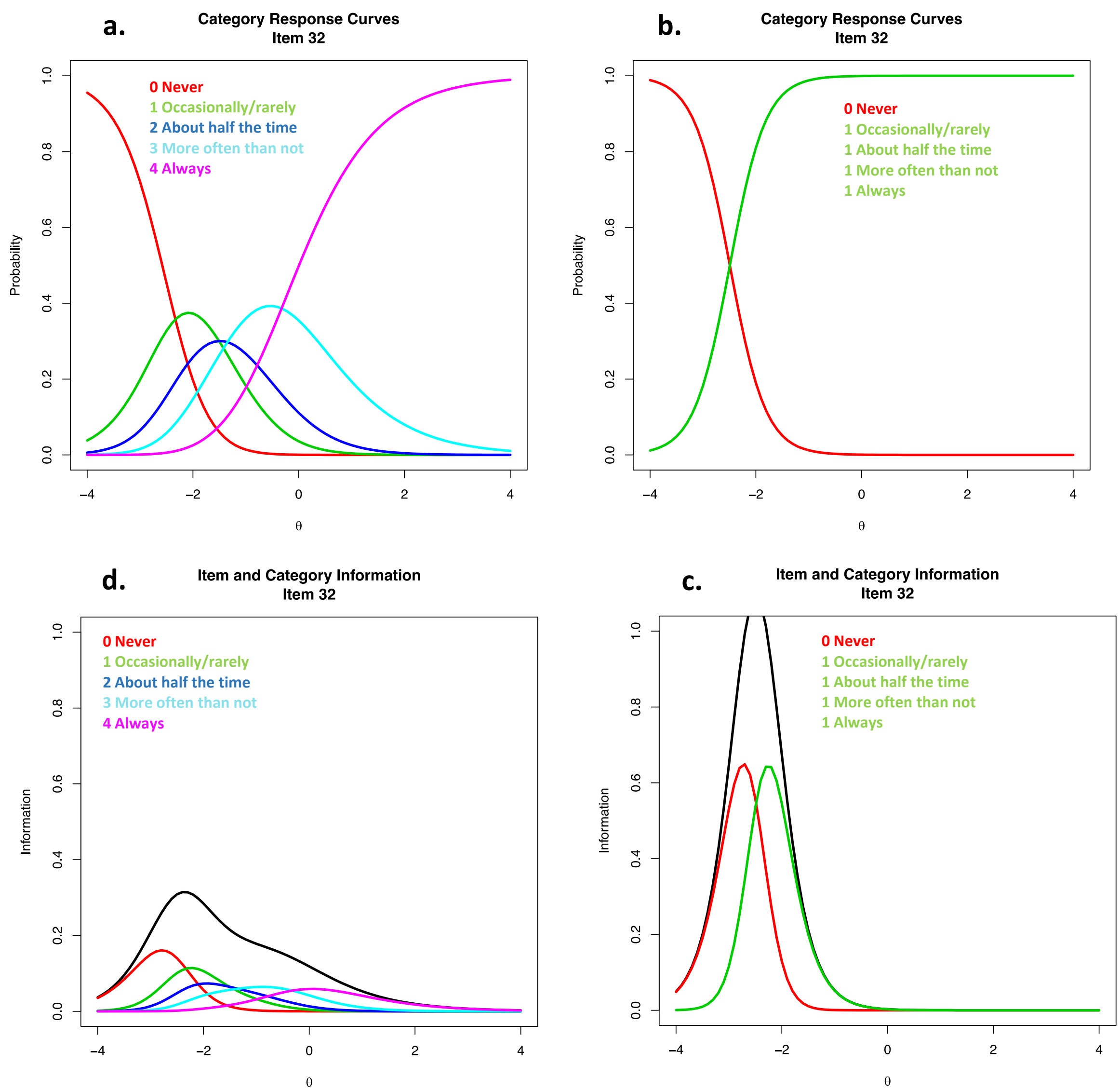


Figure 52. Factor 4 Item 32: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

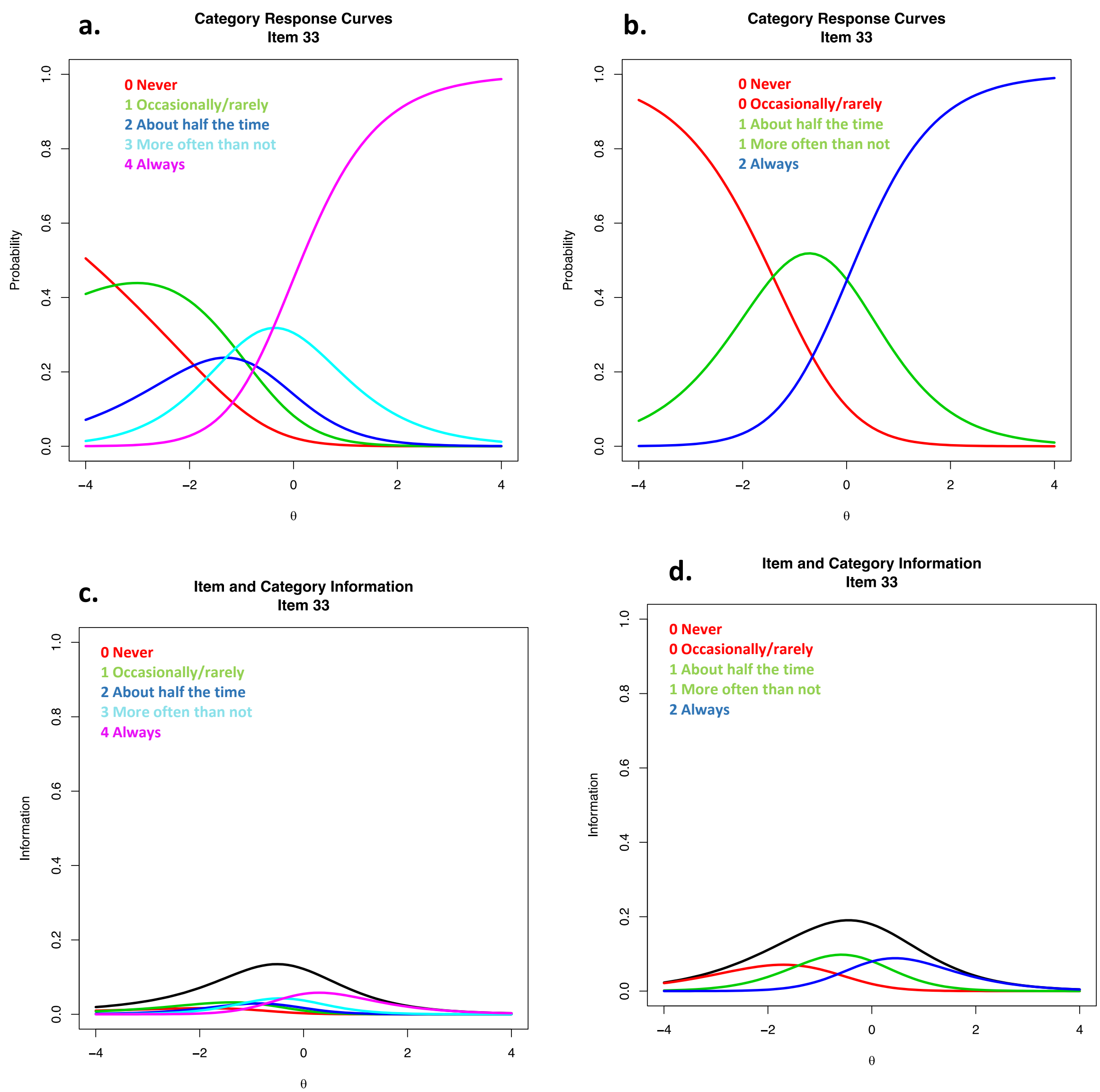


Figure 53. Factor 4 Item 33: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

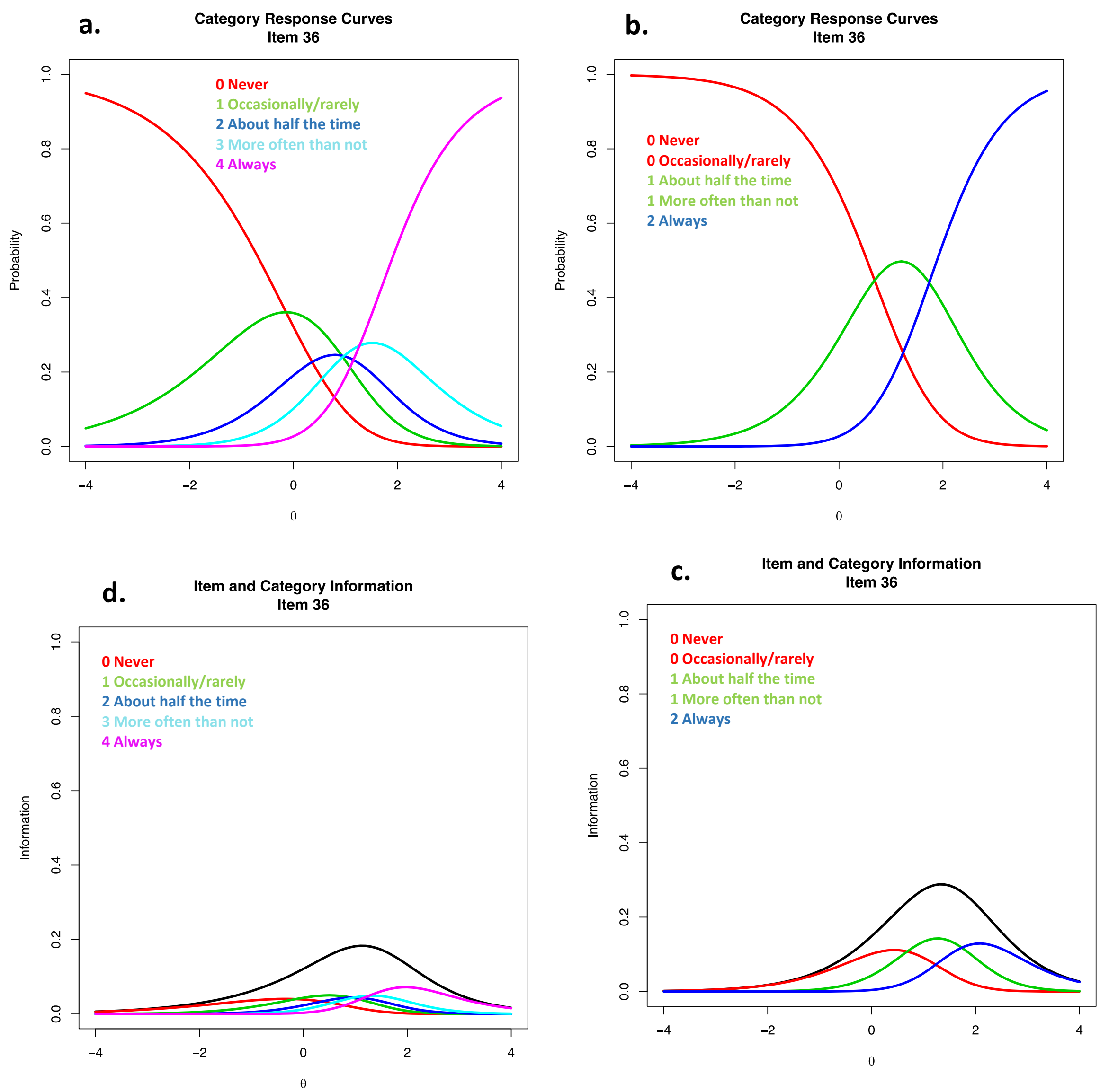


Figure 54. Factor 4 Item 36: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

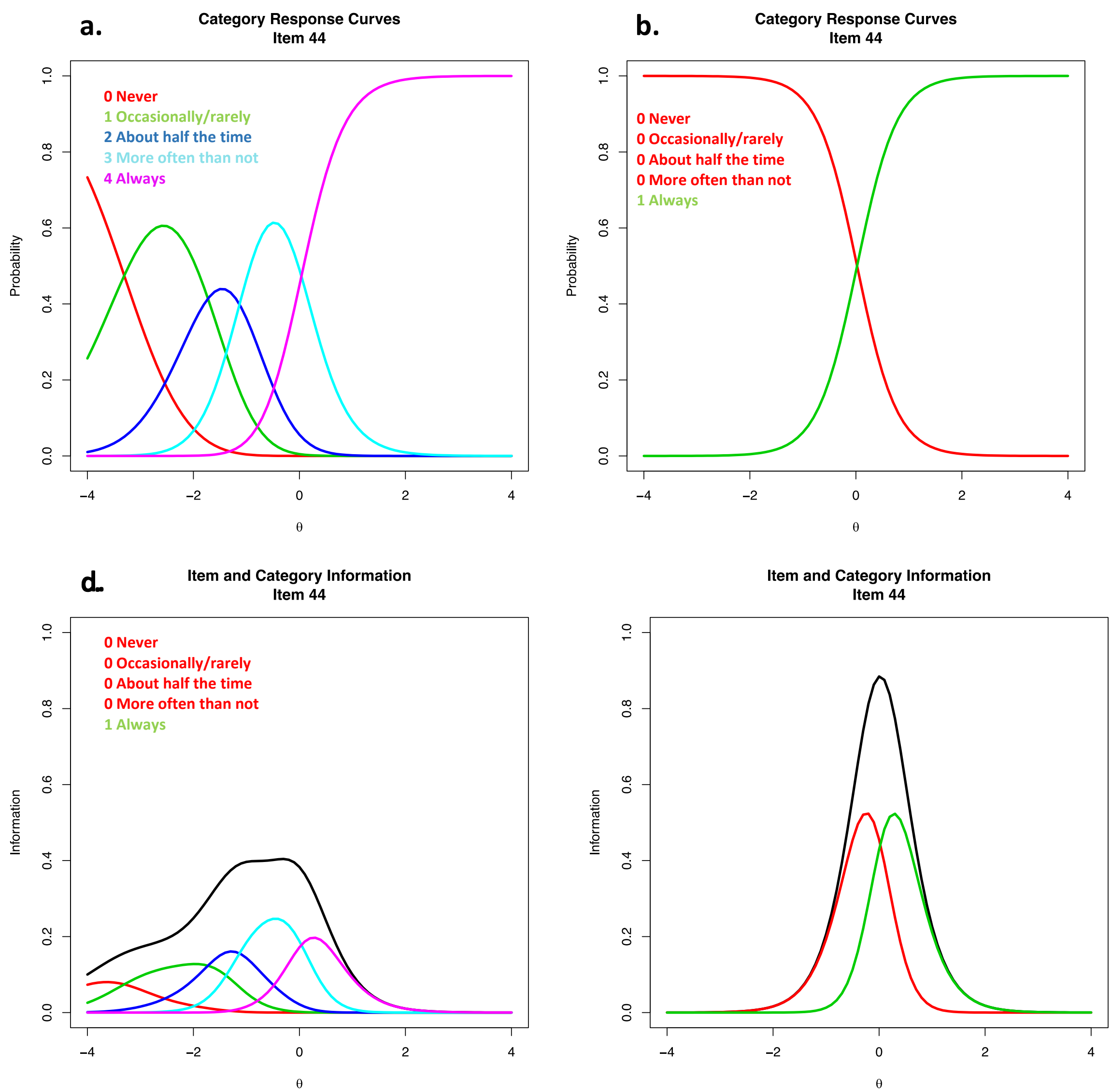


Figure 55. Factor 5 Item 44: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

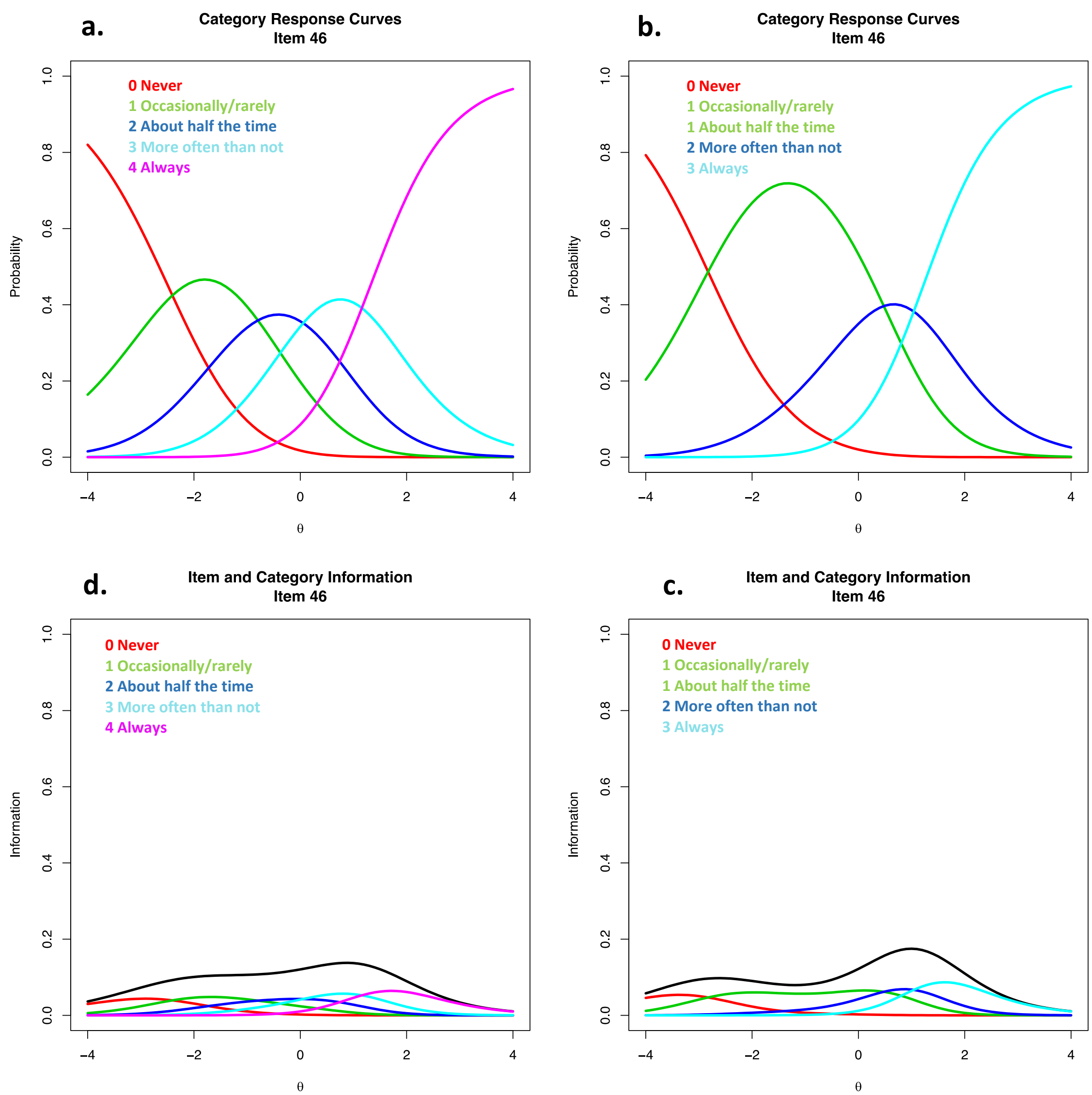


Figure 56. Factor 5 Item 46: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

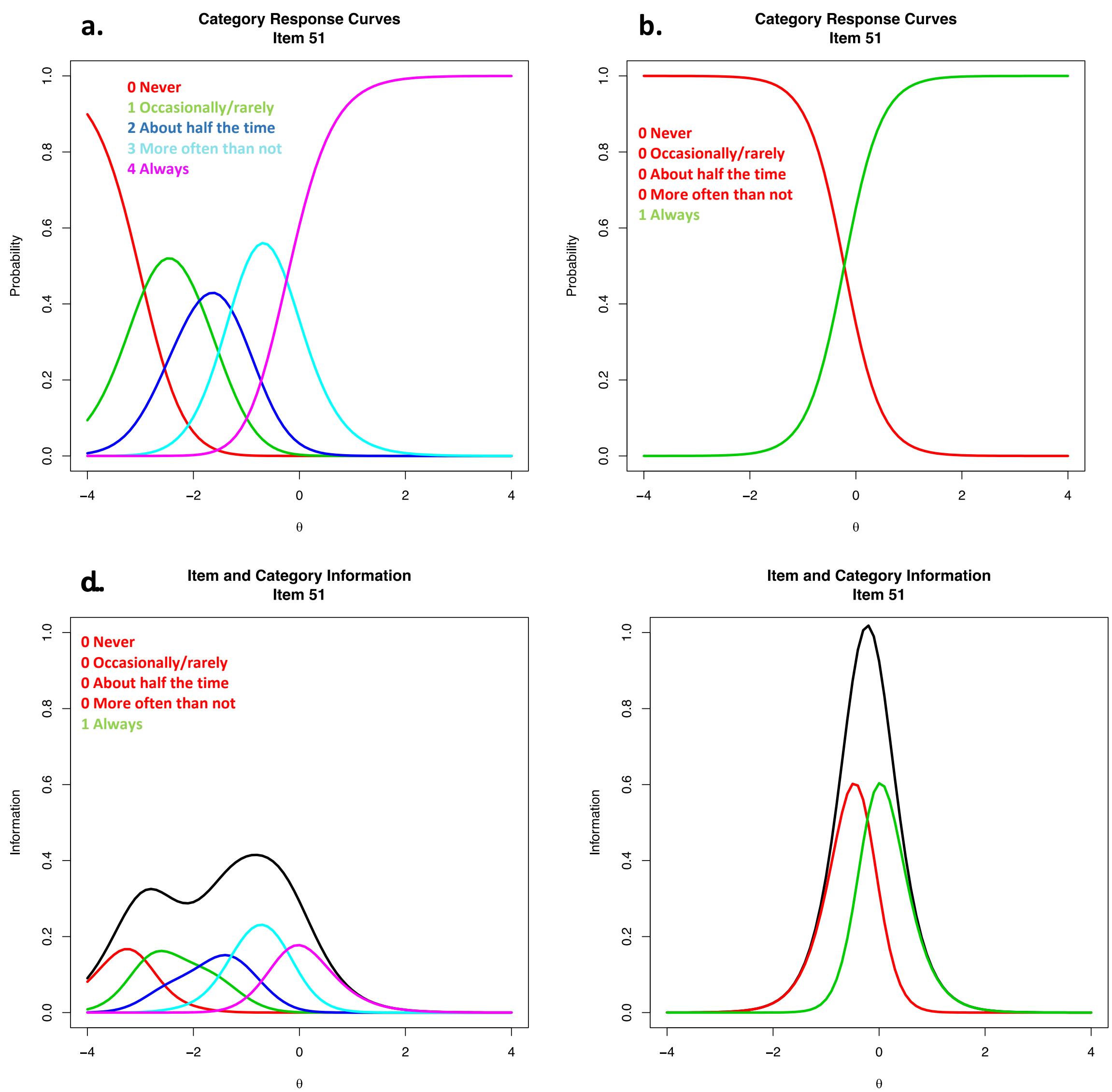


Figure 57. Factor 5 Item 51: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

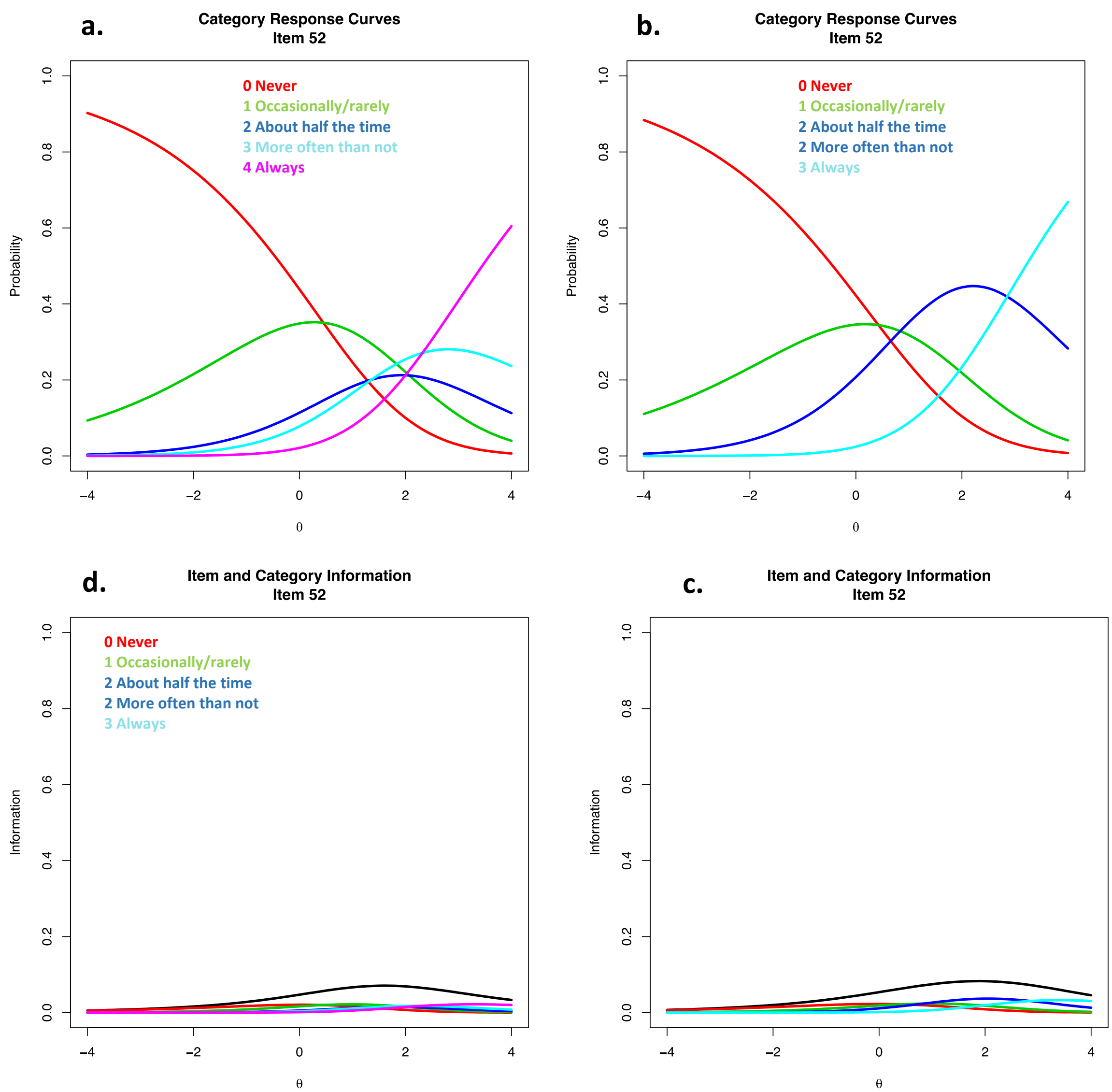


Figure 58. Factor 5 Item 52: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

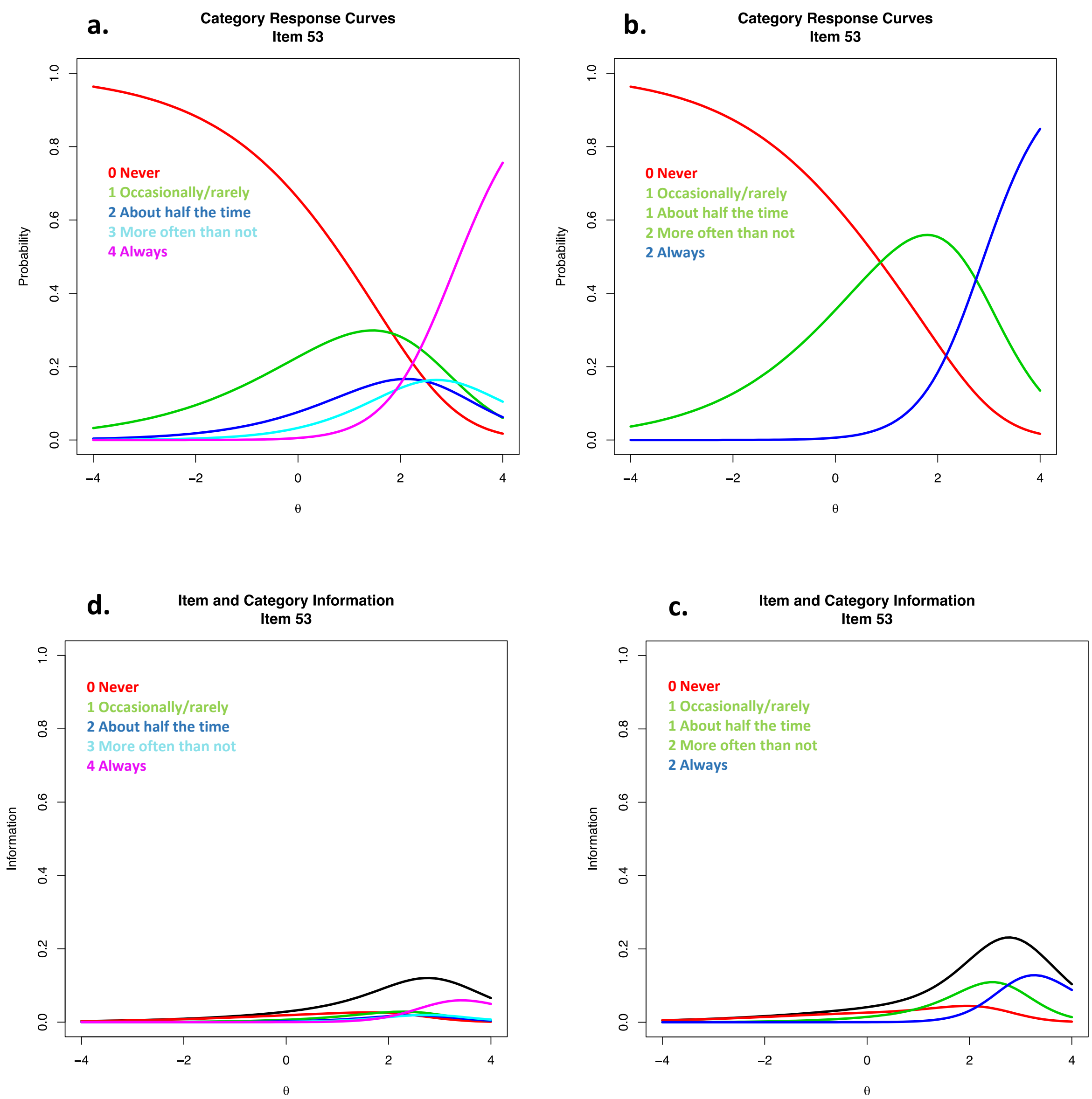


Figure 59. Factor 5 Item 53: Original Category Response Curves (a), rescored Category Response Curves (b), rescored Item and Category Information Functions (c), and original Item and Category Information Functions (d).

APPENDIX J: Test Characteristic Curves (TCCs) and Scatterplots of Original Versus Rescored Outcomes

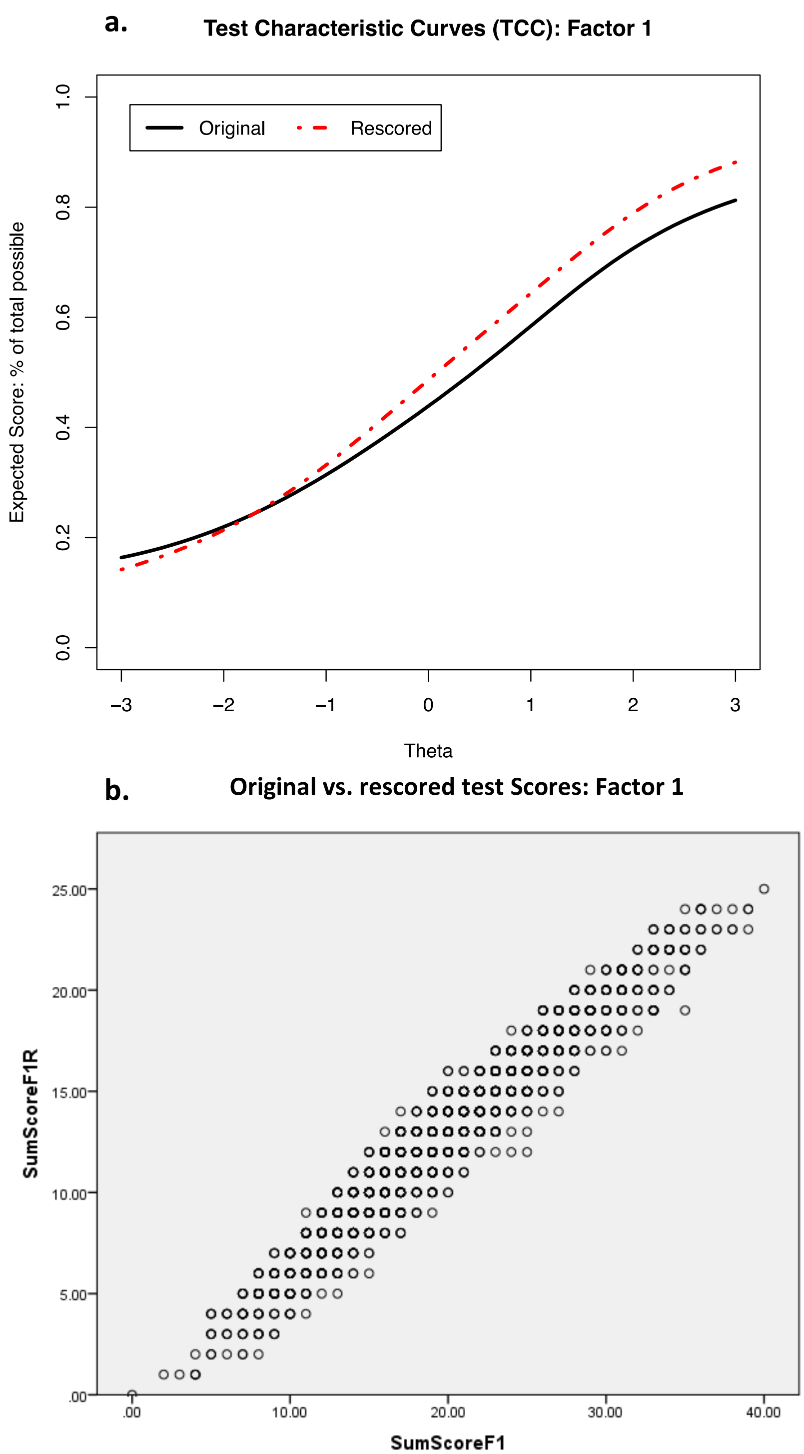


Figure 60. Rescored versus original Test Characteristic Curves (a) and scatterplot of original versus rescored test scores (b) for Factor 1.

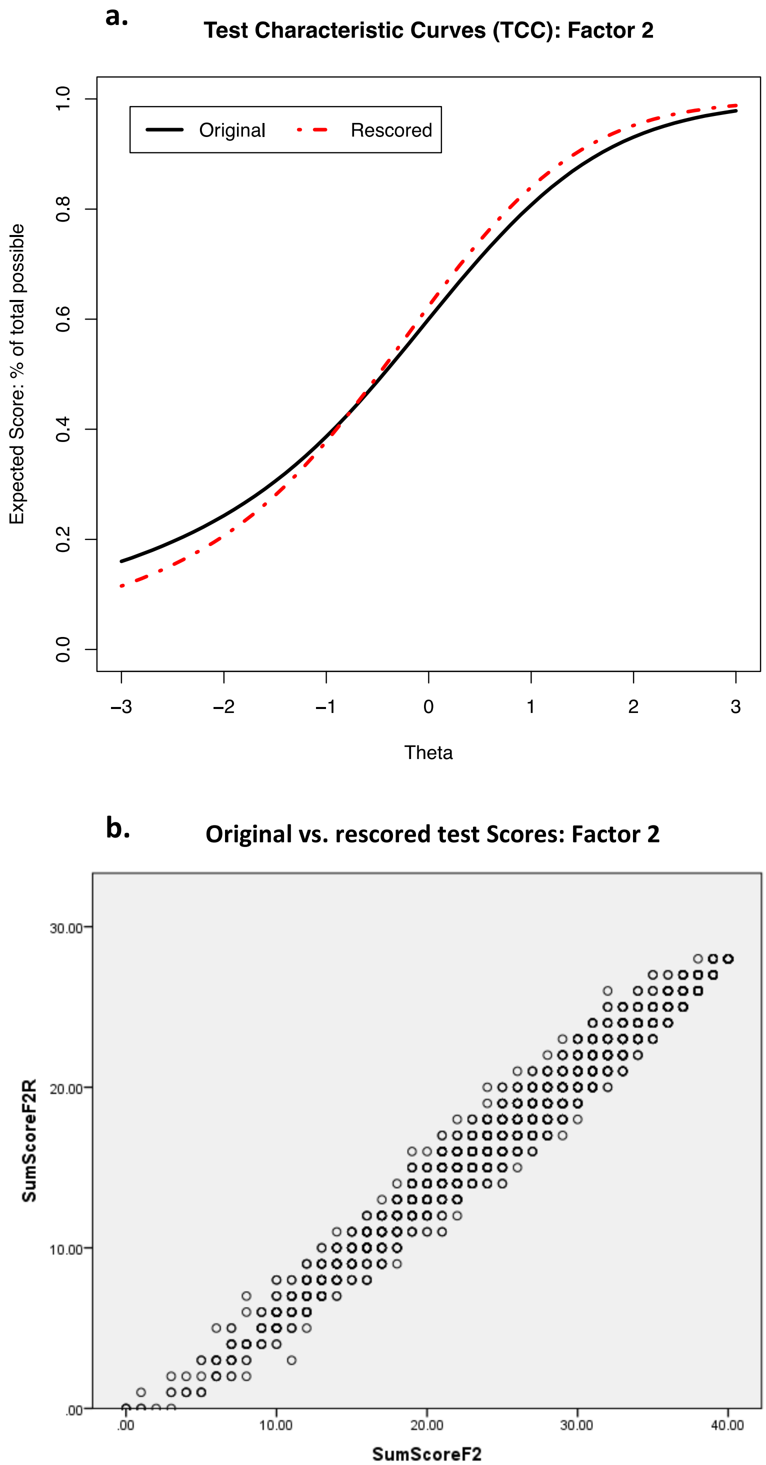


Figure 61. Rescored versus original Test Characteristic Curves (a) and scatterplot of original versus rescored test scores (b) for Factor 2.

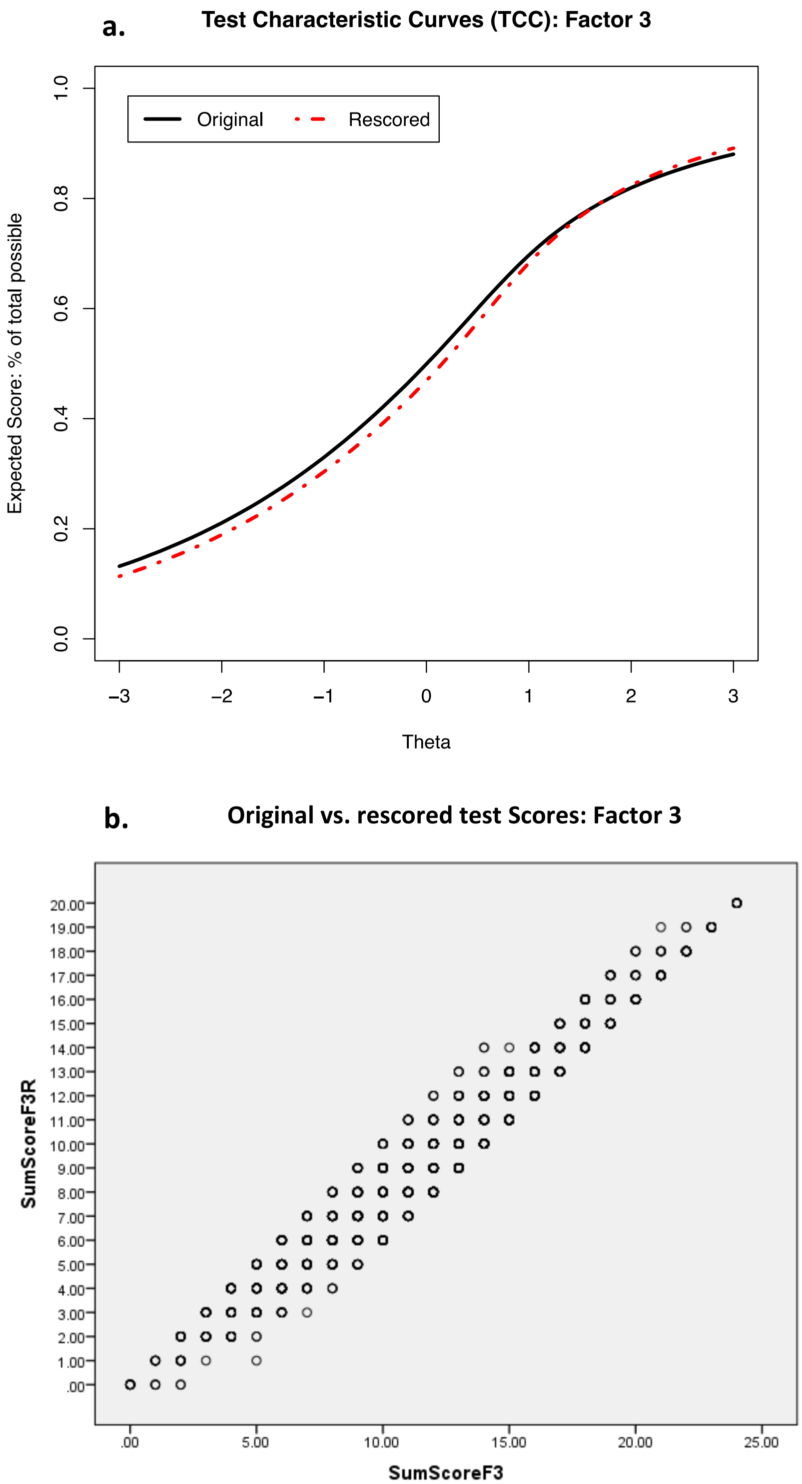


Figure 62. Rescored versus original Test Characteristic Curves (a) and scatterplot of original versus rescored test scores (b) for Factor 3.

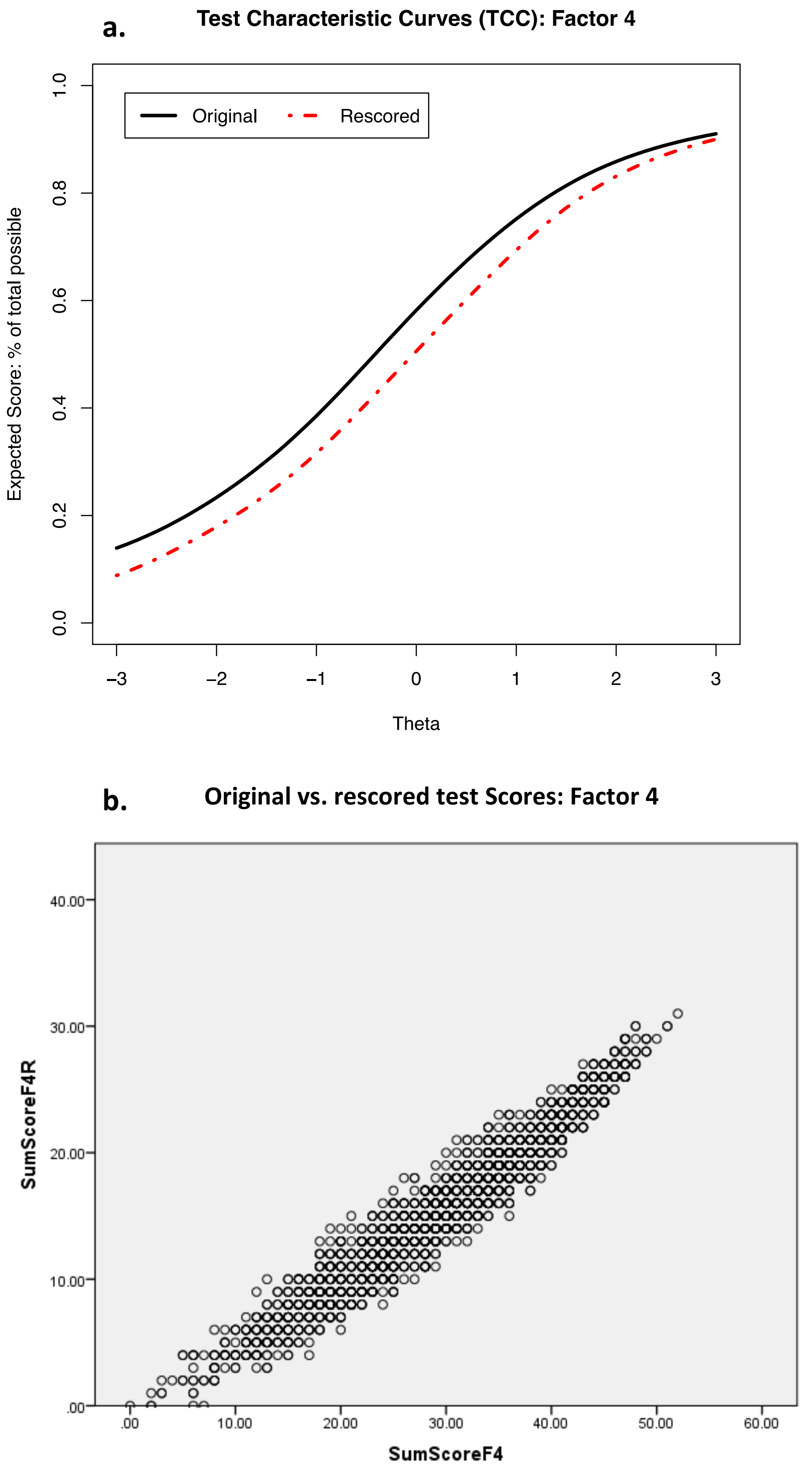


Figure 63. Rescored versus original Test Characteristic Curves (a), and scatterplot of original versus rescored test scores (b) for Factor 4.