A New Empirical Framework for the Study of Positive and Negative Affect

Monica K. Erbacher
Newfane, NY

M.A., University of Virginia, 2010
B.A., State University of New York (SUNY) at Fredonia, 2007

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Psychology

University of Virginia
August, 2013

Copyright

Abstract

Four assumptions are commonly made in affect research. Two pertain to the measurement of positive and negative affect (PA and NA): 1) PA and NA have similar, desirable measurement properties, and 2) PA and NA are adequately captured by the same response scale. Two pertain to individual differences in affect: 1) The factor structure of PA and NA is the same across individuals as it is within individuals, and 2) The correlation between PA and NA is the same across as within individuals. This dissertation project demonstrates the fallacy of these assumptions in two longitudinal data sets with different sample characteristics, item sets, periods of measurement, and response scales. Longitudinal item response models (IRMs) anchored across occasions revealed NA measures in both data sets poorly targeted respondents, produced less accurate person scores, and demonstrated more response scale parameter reversals compared to PA measures. IRMs of recoded data contrasting all possible ways of collapsing original response scales indicated a binary collapsed response scale was optimal for both NA measures, and a 5-category scale performed best for PA measures. Data were recoded according to these collapsed response scales and used in person-specific and occasion-specific exploratory factor analyses (EFAs) to challenge individual differences assumptions. EFAs exhibited substantial variation between individuals in factor loadings and factor correlations, much larger than any variation found between occasions. Results refute the four assumptions examined in two longitudinal data sets. A new framework for affect research is proposed, in which measurement and individual differences are major foci, and recommendations are made for researchers.

Dedication

To my wonderful parents, Herman and Michele Erbacher, and to the incredible woman I call sister, Megan, thank you for instilling me with a hunger for knowledge, and for your unconditional love, support, and encouragement. This achievement is for you.

Acknowledgements

First, I am grateful for the scientific skill, academic life lessons, and wealth of professional experience gained from my advisor, Dr. Karen Schmidt. Her guidance, support, and selfless dedication to student development is inspiring. You do so much above and beyond the call of duty. Thank you for everything.

Second, I want to acknowledge Dr. Steve Boker for his encouraging involvement in earlier projects and committee service during this work. I would like to thank Dr. Patrick Meyer for his committee service and generous professional support. I also want to acknowledge Dr. Timo von Oertzen for his committee service and statistical advice, and Dr. Ryne Estabrook for his generous feedback. Thank you all for your time and expertise.

This work was conducted at the University of Virginia and funded under the Quantitative Training Grant (Award T32AG020500) from the National Institute on Aging. I am grateful to Dr. John Nesselroade for allowing me to be funded by this grant, and for his invaluable feedback in early stages of this work. Thank you for your support.

Thanks also goes to the PI of the NDSHWB, Dr. Cindy Bergeman (University of Notre Dame), and the Maastricht project, Dr. Marieke Wichers (Maastricht University). Thank you both for sharing such rich longitudinal data with my colleagues and myself.

Last, I am privileged to work with colleagues who are a rare combination of brilliant and kind-hearted. Thank you all for the knowledge and joy you bring to academia. Special thanks goes to Eric Smith, the source of much appreciated sanity throughout this process, for his generous contributions of feedback, encouragement, and unconditional support. It means more than you can imagine. I love you.

Table of Contents

Chapter 1: Introduction to the Project

A New Empirical Framework for the Study of Positive and Negative Affect

For decades, affect researchers, specifically those who study positive and negative affect (PA and NA, respectively), have made a variety of assumptions about the psychometric properties of affect measures. This dissertation project takes a two-pronged approach to demonstrate the fallacy of these assumptions and the need for a new approach to affect research.

In the first prong, the Measurement Characteristics prong (Part 1), item response models (IRMs; see Embretson & Reise, 2000) are used to refute two major assumptions about the item and response scale characteristics of PA and NA measures: 1) PA and NA items have similar, desirable psychometric properties; and 2) PA and NA are measured equally well by the same Likert-type response scale. In the second prong, the Ergodicity prong (Part 2), exploratory factor analyses are employed as an alternative to the idiographic filter (see Nesselroade et al., 2007) to refute two major assumptions about the ergodicity (see Molenaar, 2004 for a formal definition, but note the term is used here to indicate similarity at the individual and sample levels) of PA and NA: 1) PA and NA have the same factor structure across as within individuals; and 2) The relationship among PA and NA is the same across individuals as within individuals.

Together, these two prongs demonstrate the shortcomings of conventional frameworks for the empirical study of affect. The Measurement Characteristics portion reveals that current measurement practices fall short of adequately measuring NA. The Ergodicity portion demonstrates that PA and NA factor structures and factor correlations

are not ergodic; thus nomothetic studies cannot be used to make inferences about the reported experience of affect at the individual level. Below, I review the importance of valid research on affect and elaborate on the motivation for each part of this dissertation project. First, a brief note on the dimensions of affect is given.

A Brief Note on Dimensions of Affect

Although empirical support exists for a variety of affect factors (for a review see Feldman-Barrett & Russell, 1999), such as Tense Arousal and Energetic Arousal (e.g., Rafeali, Rogers, & Revelle, 2007; Thayer, 1989), Pleasure and Arousal (Russell, 1980), and Valence and Arousal (Feldman, 1995), which are factor analytic rotations of PA and NA (Tellegen, Watson, & Clark, 1999), PA and NA are two of the most commonly investigated factors of affect (e.g., Crawford & Henry, 2004; Crocker, 1997; Hu & Gruber, 2008; Jackson & Bergeman, 2011; Kawata, 2006; Kercher, 1992; Ong, Zautra, & Reid, 2010; Robazza, Bortoli, Nocini, Moser, & Arslan, 2000). The measures used in the present work contain affect items from the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988), the Circumplex Model of Emotion (Larsen & Diener, 1992; Russell, Lewicka, Niit, 1989), and other adjectives selected specifically to measure PA and NA. Thus, PA and NA are examined in the present investigation.

Importance of Research on PA and NA

PA and NA are integral components of health and well-being research. In fact, Diener (2000) includes these facets of affect in his definition of subjective well-being; satisfaction with specific domains of life, overall life satisfaction, the presence of PA, and a lack of NA. Research connecting affect and facets of mental and physical health is

crucial to understanding how affect can contribute to, and potentially benefit, well-being over the lifespan.

Links have already been identified between NA and immune system functioning (Kiecolt-Glaser, McGuire, Robles, & Glaser, 2002), specifically inflammatory-related diseases such as stroke, coronary artery disease, congestive heart failure, arthritis, and osteoporosis (Ferucci et al., 1999). NA also appears to be connected with endocrine system functioning (Matsunaga et al., 2008). Empirical evidence suggests PA may influence physical health through the same pathways as NA (Antoni, LaPerriere, Scheiderman, & Fletcher, 1991; Kiecolt-Glaser et al., 2002), potentially buffering NA's detrimental effects.

PA and NA also have roles in a variety of physical health conditions, such as chronic pain. Higher NA is related to increased pain severity (Erbacher, Schmidt, & Schroeder, in prep; Roth, Geisser, Theisen-Goodvich, & Dixon, 2005; Staud, Price, Robinson, & Vierck, 2004; Zautra, Smith, Affleck, & Tennen, 2001) and increased catastrophizing about pain in chronic pain sufferers (Erbacher et al., in prep; Kratz, Davis, & Zautra, 2007; Roth et al., 2005). Emerging work indicates PA is also related to these components of chronic pain experience and may aid in counteracting influences of NA (Erbacher et al., in prep; Ong, Zautra, & Reid, 2010; Strand, Zautra, Thoresend, Odegard, Uhlig, & Finset, 2005; Zautra, Johnson, & Davis, 2005).

Finally, empirical evidence links PA and NA to reported mental health. In a sample of older adults (Montpetit, Bergeman, Deboeck, Tiberio, & Boker, 2010), NA and stress were significantly coupled, such that acceleration in NA predicted acceleration in

stress levels and vice versa over time. Similarly, NA and stress were positively related in a second sample of older adults; however, PA interacted with stress to offset this relationship. PA also mediated the effect of stress on NA one day later. These results were confirmed in a study of recent widows (Ong, Bergeman, & Bisconti, 2004). Dua (1993) similarly found positive associations of NA with stress and depression in university students, and inverse associations between the same constructs and PA. A study of 1,003 adults supported these associations (Crawford & Henry, 2004).

Conducting valid empirical investigations on the roles of PA and NA in mental and physical well-being across the lifespan is predicated on adequately measuring PA and NA. Without adequate measurement, scientific inquiry is fruitless.

The present project highlights two sets of assumptions made about the psychometric properties of PA and NA in the majority of the existing affect literature, and compiles evidence indicating these assumptions are incorrect. Adhering to these false assumptions will grievously slow progress in the field of affect research by producing biased findings that continue to support opposing theories of affect (e.g., conflicting theories on the bipolarity of affect, see Reich & Zautra, 2002; Zautra et al., 1997; conflicting theories on the factor structure of affect, see Crawford & Henry, 2004; Gaudreau, Sanchez, & Blondin, 2006). Below, each set of PA and NA assumptions is identified and evidence of their invalidity is offered. First, two assumptions about measurement characteristics of PA and NA item responses are addressed. Second, two assumptions about the ergodicity of PA and NA responses are examined.

Chapter 2: Measurement of PA and NA

False Assumption Set 1: Measurement Characteristics of PA and NA

The first set of assumptions revolve around the item and response scale characteristics of PA and NA measures: 1) PA and NA item sets have the same desirable measurement characteristics; and 2) The same rating scale adequately captures PA and NA. The following is a review of evidence indicating these assumptions are prevalent in current affect literature, as well as a review of empirical evidence refuting these assumptions.

Conventional FA: Evidence of Implied Assumptions in the Literature

The vast majority of the literature on PA and NA measurement takes a factor analytic approach. Both PA and NA items are often analyzed with the same types of models, frequently confirmatory or exploratory factor analysis (e.g., Gaudreau, Sanchez, & Blondin, 2006; Kercher, 1992; Leue & Beauducel, 2011). Also, both PA and NA items are most commonly administered using the same response rating scale, usually a 5- or 7-point Likert-type scale (see Likert, 1931), such as in the Positive and Negative Affect Schedule (Watson, Clark, & Tellegen, 1988; for administration examples see Crawford & Henry 2004; Crocker, 1997; Kercher, 1992). Measuring PA and NA with the same response scales and analyzing PA and NA responses with the same models implies that PA and NA are expected to have similar measurement properties. More specifically, such studies assume PA and NA have similarly adequate factor analytic characteristics, and both types of affect are measured equally well by the same response scale.

Taking a conventional factor analytic approach reveals two important differences between PA and NA. First, NA indicators tend to have less desirable (lower) common factor loadings and (higher) uniqueness factor loadings. For example, in a confirmatory test of correlated PA and NA factors with 1,003 adults (Crawford & Henry, 2004), PA items had a median standardized loading of 0.69 (M = 0.66, SD = 0.10), whereas NA items had a median loading of 0.62 (M = 0.59, SD = 0.12), resulting in higher uniqueness loadings for NA items than PA items. Although this difference is small, the same pattern has been observed in a variety of studies, including orthogonal PA and NA factors tested with 645 athletes, 10 to 17 years old (Crocker, 1997), correlated PA and NA factors tested with a Spanish sample of 708 adult women, 45 to 65 years old (Joiner, Sandin, Chorot, Lostao, & Marquina, 1997), one PA and two NA factors calibrated on 305 French Canadian athletes, 14 to 47 years old (Gaudreau, Sanchez, & Blondin, 2006), and one PA and two NA factors extracted using exploratory factor analysis with a sample of 83 adults, 60 to 80 years old, with a clinical diagnosis of generalized anxiety disorder (Beck et al., 2003). Also, the standard errors of NA parameters tend to be larger than those of PA parameters (e.g., Crawford & Henry, 2004). For a more comprehensive review of various factor solutions for the PANAS (Watson et al., 1988), see Leue and Beauducel (2011).

Second, the empirical findings on the dimensionality of NA are less consistent than those of PA, and the overall factor structure of affect shares this inconsistency. Empirical support exists for both one (e.g., Crawford & Henry, 2004; Crocker, 1997) and two (e.g., Beck et al., 2003; Gaudreau et al., 2006) NA factors, whereas most findings

support a single PA factor (e.g., Beck et al., 2003; Crocker, 1997).

It is unclear whether these differences are due to the nature of the underlying PA and NA constructs being measured, the quality of existing measures of PA and NA, both, or something completely separate, such as sample composition. Conventional factor analysis results do not provide much information on how the response scale for each affect item is used. Continuing to only use a factor analytic approach is unlikely to yield any more novel information. Rather, such endeavors will continue to produce conflicting findings similar to those already reported. A paradigm shift is necessary for making any future gains in information on affect measurement characteristics. Specifically, taking an Item Response Theory approach, in addition to the factor analytic approach commonly used, will reveal unique information about the measurement of PA and NA that is beneficial to future affect research.

The Need for an Item Response Theory Approach

Very few empirical investigations have employed techniques from Item Response Theory in affect research, presumably due to lack availability of open-source and non-proprietary software for much of the past 30 years. While factor analytic techniques provide valuable information about the measurement process, and while it should be noted that some of the simplest factor analysis models for binary data are transformable into some of the simplest dichotomous item response models (IRMs; see Kamata & Bauer 2008), IRMs can yield additional information about the psychometric properties of items and measures that greatly improve our knowledge of measurement and developmental processes (e.g., see Nesselroade & Schmidt McCollam, 2000).

Benefits of item response models (IRMs). Some of the most useful results of IRM analyses stem from a single property of this set of statistical techniques: Item and person parameters are estimated on the same scale. Item location, or beta, parameters quantify how difficult an item is to endorse (or answer correctly), representing an item's location on the latent dimension measured. Similarly, person location parameters, called thetas, represent the locations of individuals on the same latent dimension, indicating the amount of the ability, trait, or state the person exhibits.

Item beta scores and person theta scores are both estimated on the same log-odds unit (logit) scale, allowing for direct comparison of person and item locations on the latent dimension of interest (see Baker & Kim, 2004; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). If the distribution of item locations overlaps much of the distribution of person locations, then the measure targets individuals in the sample well. If all items are located much lower or higher than the individual on the estimated logit scale, analyses will fail to produce an accurate estimate of that individual's score. These results seem particularly beneficial when examining a construct that is commonly observed at low levels, such as the potentially underreported and/or rarely experienced NA. Results from IRMs will allow us to ask the question: Are any items that measure NA good at tapping into lower levels of the construct, or are the NA items we currently use too difficult for most respondents? That is, are current items capable of measuring self-reported NA as it is usually observed?

Results from IRMs can communicate even richer information when applied to longitudinal studies. From applications of single-occasion IRMs to longitudinal data, we

can determine how stable the psychometric properties of PA and NA items remain across multiple measurement occasions. Using anchoring methods to link IRMs to the same scale across occasions allows longitudinal measurement characteristics of a survey to be examined. Unfortunately, it appears there may be only one study to date that has applied IRMs to longitudinal affect data over many measurement occasions (Erbacher, Schmidt, Boker, & Bergeman, 2012).

Evidence Against Assumptions: A Longitudinal IRM Study

Previous work by Erbacher and colleagues (2012) includes a preliminary examination of these IRM statistics in a longitudinal study. As part of the Notre Dame Study of Health and Well-Being (NDSHWB; for more details see Jackson & Bergeman, 2011; Russel, Bergeman, Deboeck, Baird, Monpetit, & Ong, 2011), a sample of 53- to 91-year-olds responded to the PANAS (Watson et al., 1988) plus additional items, most taken from the Circumplex Model of Emotion (Larsen & Diener, 1992), once a day for 56 consecutive days. Three methods of linking IRMs of separate measurement occasions were compared to unlinked models to determine the best method of analyzing longitudinal data with IRMs.

In both the unlinked (cross-sectional) analyses and in the best-performing linked (longitudinal) models, the same differences between PA and NA emerged. First, all of the NA items were too difficult for most participants, despite researchers' attempts to add lower-intensity NA items, such as Fatigued. PA items were better matched to the sample. Second, individuals almost exclusively used the lower half of the 5-point Likert-type response scale when responding to NA items, with many of the items appearing to be

dichotomous or close to dichotomous, whereas PA responses spanned the entire response scale for every item. Third, category usage was unstable for almost all NA items across the measurement period, while category usage for almost all PA items remained stable across occasions.

Disparate category usage patterns for affect items have been confirmed in a sample of chronic pain sufferers (Schmidt, 2006) in which several methods of collapsing a 5-point scale for items on the PANAS were compared. The patterns with the best statistical performance included collapsing most NA items down to a 3-category response scale, while leaving most PA items on a 5-point scale, suggesting that even persons for whom NA is likely a more common experience (e.g., chronic pain sufferers) a 5-point scale may require more precision than respondents tend to provide.

Overview of Current Project: Measurement

These preliminary findings demonstrate the useful kinds of information that can be obtained via IRM analyses, along with the need for this type of longitudinal measurement evaluation in the affect literature. Part 1 of this dissertation project revisits findings from work with the NDSHWB in greater detail and tests their replicability in a second data set. This second data set includes responses to a smaller set of affect items made by a younger sample of female adults in Germany over a much shorter time window (5 days), with more frequent measurement occasions (10 times a day). Furthermore, this work examines category usage in both data sets to determine whether PA and NA items warrant different response scales in both samples. If results hold across data with two very different time scales, sets of items, and samples of individuals, it will

provide strong support that changes are needed in the affect field's approach to measuring PA and NA.

Many of the studies reviewed thus far take a traditional nomothetic approach to measurement, indicating the existence of a second set of assumptions that PA and NA are approximately the same across individuals as they are within individuals. The next chapter defines this set of assumptions, reveals further evidence of their existence in the affect literature, and compiles empirical support refuting these assumptions.

Chapter 3: PA, NA, and Individual Differences

False Assumption Set 2: Ergodicity of PA and NA

The second prong of the proposed project addresses two major assumptions about the ergodicity, or the similarity of between- and within-person structures and processes (see Molenaar, 2004), of PA and NA. This set includes: 1) The assumption that the factor structure of PA and NA is the same across as within individuals; and 2) The assumption that the relationship between PA and NA is the same across as within individuals.

Importance of the Individual

Although an idiographic focus has recently gained momentum in psychology (e.g. see Molenaar, 2004; Nesselroade, Gerstorf, Hardy, & Ram, 2007), the notion of such an approach has been present in the literature for several decades (e.g., see Cattell et al., 1947; Nesselroade & Ford, 1985), even in the affect literature, as evidenced by Zevon and Tellegen's (1982, pg.111) remark that "[…] the idiographic study of individuals, rather than being antithetical to scientific psychology, can provide information of value and relevance to nomothetic description." With the recent refocus on an idiographic approach (see Molenaar, 2004; Nesselroade et al., 2007), evidence is surfacing (and resurfacing) that indicates the appearance of PA and NA across individuals is different from that of PA and NA within individuals (Feldman, 1995; Lebo & Nesselroade, 1978; Nesselroade et al., 2007; Zevon & Tellegen, 1982). Readily aggregating over PA and NA within individuals in a sample has likely contributed to the ongoing debate over contradictory theories about the bipolarity (Erbacher, Schmidt, & Bergeman, under review) and the factor structure of PA and NA. If these assumptions about the ergodicity

of PA and NA are false, a nomothetic approach tells us nearly nothing about the process of affect for a given individual.

Each of these two assumptions is addressed separately in the remainder of this review. First, empirical work reviewed in the previous section, conducted under the assumption that the factor structure of PA and NA is the same across as within persons, is briefly summarized, followed by evidence exposing the fallacy of this assumption. Second, literature implying the relationship among PA and NA is the same across persons as within is reviewed, followed by findings refuting this assumption. Finally, the aim of the present project is identified.

Individual Differences in Factor Structure

Nomothetic FA: Evidence of assumptions in the literature. As reviewed in the previous chapter, a variety of affect factors have emerged from past empirical nomothetic investigations, such as tension and energy (Thayer, 1989), valence and arousal (Feldman, 1995), and PA and NA (Watson et al., 1988). Multiple factor solutions have also been discovered in nomothetic examinations of the PANAS (Watson et al., 1988; for a more comprehensive review see Leue & Beauducel, 2011), with support existing for two orthogonal PA and NA factors (Crocker, 1997; Kawata, 2006; Kercher, 1992), two correlated PA and NA factors (Crawford & Henry, 2004), and a three-factor solution (Gaudreau et al., 2006). In these nomothetic investigations, it is assumed the factor structure of PA and NA is the same for all individuals. Perhaps the instability observed in the factor structure of affect across samples is a result of aggregating over individuals with disparate affect factor structures.

Evidence against the ergodicity assumption. In fact, evidence of individual differences in affect factor structure already exists in studies that have employed an idiographic approach. For example, Lebo and Nesselroade (1978) examined the factor structure of 75 affect adjectives with *p*-technique factor analysis, with a sample of five 22- to 25-year old pregnant women over 15-weeks of daily affect assessments. Based on conventional factor retention criteria, anywhere from five to nine factors were extracted for participants. Analyses revealed substantial individual differences in the factor structure of affect. For example, Enthusiastic loaded on an Energy factor with a loading of 0.47 for one individual and 0.71 for another. For a third, Enthusiastic loaded on one of two Well-Being factors with a low-moderate loading of 0.49, and a loading of 0.93 for a fourth individual. Variation between individuals was observed for many of the 75 affect items, indicating important individual differences in affect factor structure.

Similarly, Zevon and Tellegen (1982) administered a 60-item mood checklist to 23 undergraduate students once daily for 90 consecutive days to test for individual differences in affect factor structure. Two factors were extracted from each person's longitudinal data and compared to PA and NA. For all but two participants, two factors reasonably congruent to PA and NA were extracted, although order of extraction varied by individual. Factor loadings varied widely between individuals. For example, the factor loading for the item Interested on the PA factor ranged from 0.56 to 0.83 across participants, and the item Scared had loadings on NA ranging from 0.17 to 0.85. Thus, although individuals may experience affect through a common set of factors, the manifestation of each of those latent constructs differs by individual.

Feldman (1995) attempted to replicate Zevon and Tellegen's (1982) approach with 24 psychology university students over 62 to 91 consecutive days. A circumplex of 16 affect adjectives was created and responses from each person were analyzed with *p*-technique factor analysis. Feldman (1995) found large individual differences in affect factor structure using principal axis factor analysis. The number of factors extracted for each participant varied from one to four. Feldman was able to identify one factor reasonably congruent to a valence dimension and one factor congruent to an arousal dimension for all but one participant, though the primary factor varied by individual. Examination of *p*-technique results revealed affect space plots were near perfect circumplexes for some individuals, ellipses for others, odd groupings of items for others, and even a line for one participant, indicating substantial individual differences in the factor structure of affect.

Further support for individual differences in affect factor structure comes from a reanalysis (Nesselroade et al., 2007) of the 100-occasion affect data for five pregnant women from Lebo and Nesselroade (1978). After identifying five factors present for at least four participants with *p*-technique factor analysis, with at least one manifest per factor consistent for all five participants, responses to a subset of items were analyzed by applying the idiographic filter to the common factor model. Results allowed for direct comparisons of participants on each of the common factors identified. For example, a common set of five indicators of a Well-being factor were found across all participants, with additional items loading on this factor that varied by individual. Even the loadings for the common set of five items varied widely by individual. Relaxed had the highest

loading on Well-being ($\lambda = 2.00$) for one woman, while for a second the highest loading was for Cheerful ($\lambda = 0.97$), and for a third all five items had comparatively low (0.38 to 0.57) loadings.

Clear empirical support exists for individual differences in the factor structure of affect. Despite evidence spanning the past 35 years, the majority of current research on affect factor structure continues to take a nomothetic approach, aggregating over individual differences. In this dissertation project, I seek to confirm the presence of individual differences in affect factor structure in two longitudinal data sets with different sets of items, different sample characteristics, and different time scales. Moreover, I aim to detect individual differences in both the relationships between the manifest and latent variables (i.e., loadings), as well as the latent variable associations, as discussed below.

Individual Differences in the PA-NA Relationship

Nomothetic studies: Evidence of implied assumptions in the literature. A variety of competing theories about the relationship between PA and NA measurements have resulted from a multitude of nomothetic investigations. These theories make contradictory claims over how PA and NA are related at any given time, across the adult lifespan, and in specific contexts.

For example, the Bivariate Model and the Bipolar Model (see Reich & Zautra, 2002) are two competing theories on the relationship between PA and NA at any given time. The Bivariate Model of affect postulates PA and NA are independent and uncorrelated continuums of emotion. Watson, Clark, and Tellegen (1988) developed the PANAS under the Bivariate Model with a sample of undergraduates, and the

independence of its PA and NA subscales have been validated in a variety of samples (e.g., older adults [Kercher, 1992]; adolescents [Crocker, 1997]; and undergraduates [Gable, Reis, & Elliot, 2000; Goldstein & Strube, 1994]).

The Bipolar Model posits PA and NA are opposite poles of the same continuum and thus are negatively correlated. Strong evidence for the Bipolar Model comes from a daily diary study (Ready et al., 2008), in which 49 participants rated their PA and NA (five adjectives each) for 28 consecutive days. Average PA across occasions was significantly correlated with average NA for both younger and older adults (r = -0.66, p < 0.05, and r = -0.61, p < 0.05, respectively). Although the authors collected longitudinal data, the correlation of PA and NA was conducted on mean scores aggregated over occasions. Negatively correlated PA and NA factors on the PANAS have been found in adult samples (r = -0.30, p < 0.001 [Crawford & Henry, 2004]; r = -.29 [Molloy, Pallant, & Kantas, 2001]), and adolescent samples (r = -.25 [Molloy et al., 2001]).

Additional conflicting findings come from empirical investigations of PA and NA across the adult lifespan. For example, proponents of Socioemotional Selectivity Theory (SST; Carstensen, Isaacowitz, & Charles, 1999) posit aging leads to a change in time perception and behavioral goals, resulting in a shift from bivariate to bipolar affect across the lifespan. Support for SST is found in studies on the effects of emotional stimuli on amygdala activation (Mather, Canli, & Whitfield et al., 2004), memory (Charles, Mather, & Carstensen, 2003) and attention (Mather & Carstensen, 2003), and self-report data in which the correlation of PA with NA was more strongly negative in older adults compared to younger adults (Ready et al., 2008; Ready, Robinson, & Weinberger, 2006).

Opposing evidence suggests PA and NA may become more positively associated as time perspective shifts, commonly in late adulthood. This work is centered around the phenomenon of poignancy; The presence of both positive and negative emotions brought about by experiencing something pleasant for the last time. Support for this phenomenon comes from a study of university students and older adults in the U.S. (Ersner-Hershfield, Mikels, Sullivan, & Carstensen, 2008), as well as a study of Chinese adults (Zhang, Ersner-Hershfield, & Fung, 2010).

Finally, Proponents of dynamic theories claim the relationship between PA and NA changes over much shorter periods of time, becoming increasingly bipolar in stressful situations, when individuals have less cognitive resources available and must use the simpler Bipolar Model to evaluate affect (e.g., Dynamic Model of Affect; Zautra et al., 1997; Dynamic Integration Theory; Labouvie-Vief et al., 2007). These theories were supported in a study of older adults who were asked to complete a public speaking task (Study 1 in Zautra et al., 2000). PA and NA were negligibly correlated before the public speaking task (r = .05) and more strongly correlated immediately after the task, when stress should be elevated (r = -.33). Similarly, Reich and Zautra (2002) found higher levels of arousal, in combination with high PA or high NA, were associated with a more extreme inverse relationship between PA and NA. This relationship between PA-NA bipolarity and stress has also been confirmed in older adults dealing with life stressors, such as physical disability or widowhood (Study 2 in Zautra et al., 2000).

In sum, there is empirical support for several conflicting models and theories of affect bipolarity. Inappropriate aggregation over individual differences in the relationship

between PA and NA may perpetuate continued support of conflicting models. If the PA-NA association differs largely between individuals, then the value of this association for a given sample is contingent on the characteristics of the individuals in the sample. For example, if the PA-NA correlation varies from large negative to zero for several individuals, and these individuals are selected in a single sample, the correlation calculated for that sample will likely be negative. Depending on the participants selected, the correlation between PA and NA for a second sample could be approximately zero. The presence of conflicting results demonstrates the need for more complex theories of PA and NA informed by examinations of these constructs at the individual level.

Evidence against the ergodicity assumption. There is increasing empirical support for individual differences like the ones hypothesized above in the correlation between PA and NA. In a study of 24 adults measured daily for 62 to 91 days, Feldman (1995) found individual differences in the correlation of PA and NA measured by the PANAS (Watson et al., 1988). Individual-level PA-NA correlation coefficients across all measurement occasions ranged from -.72 to .21 (M = .35, SD = .28). Similar variation in the association of PA and NA were found in a longitudinal structured interview study (Coifman, Bonnano, & Rafaeli, 2007), a study of Italian students and adults (Terracciano, McCrae, Hagemann, & Costa, 2003), and in five daily diary studies conducted with undergraduate participants (Rafeali, Rogers, & Revelle, 2007).

These individual differences were examined and extended by work with the NDSHWB (Erbacher et al., under review). Large variation between older adults was found in the correlation of levels of PA and NA and among their first and second

derivatives. Participants varied on three correlations: PA position (level) with NA position (*r* range = -.99 to .88), PA velocity with NA velocity (*r* range = -.92 to .55), and PA acceleration with NA acceleration (*r* range = -.99 to .66).

Extensive evidence for individual differences in the relationship between PA and NA exists. It is critical to examine this relationship at the individual level to create more adequately informed theories about the experience and bipolarity of PA and NA. Without a closer examination of the individual, the controversy of affect bipolarity fueled by group-based findings will continue to be unresolved and improperly informed.

Examining Individual Differences at Two Levels

There is evidence for individual differences in both affect factor structure and the relationship between PA and NA or similar factors. The combination of these results begs the question: How do we properly parse variance attributable to individual differences in factor structure from variance attributable to differences in factor correlations? Feldman (1995) has addressed this issue via separate *p*-technique analyses, with results supporting substantial individual differences in affect factor structure and in the PA-NA correlation. Proposed analyses aimed to test for variation in both factor structure and factor correlations using the Idiographic Filter (IF; Nesselroade et al., 2007). In the present project, an alternative analysis plan is carried out to explore individual differences in factor structure and in affect factor correlation separately, in two longitudinal data sets.

The aim of this work is to examine the four assumptions reviewed above in two very different longitudinal affect data sets. If the findings obtained refute assumptions in both data sets, they will provide strong evidence of the fallacy of these assumptions.

Chapter 4: Longitudinal Data Sets

Methods

Notre dame Study of Health and Well-Being (NDSHWB)

Participants. Two hundred eighty-eight participants, 53- to 91-years of age (M = 68.0, SD = 5.3, $Median$ = 58), were recruited through the Notre Dame Study of Health and Well-being (NDHWB; see Jackson & Bergeman, 2011; Russel, Bergeman, Deboeck, Baird, Monpetit, & Ong, 2011). Age data was missing for 2 (0.7%) participants. Over half (58%) the participants were female, 81% identified themselves as White, 12% as African American, 4% as Hispanic, and 3% as Asian. Income varied widely among participants, with 2.4% earning less than $7,500, 17.4% earning between $7,500 and $15,000, 20.8% earning between $15,000 and $25,000, 25.7% earning between $25,000 and $40,000, 22.6% earning between $40,000 and $75,000, 5.2% earning between $75,000 and $100,000, and 3.1% earning over $100,000 (2.8% of income responses were missing). Education also widely varied among participants, with 7.6% having completed a graduate, medical, or law degree, 6.3% a post college professional degree, 12.5% a college degree, 23.3% some college classes, 8.3% vocational education, 38.5% high school, 2.1% having completed 9th grade, and 0.7% school through 6th grade.

Measures. The Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) was developed to measure PA and NA as two orthogonal dimensions of affect with 20 emotional adjectives. Empirical findings have confirmed this two-factor structure in a variety of samples (e.g., Crawford & Henry, 2004; Crocker, 1997; Kawata, 2006; Kercher 1992). Thus, the present study will treat the PANAS as having two factors; PA

and NA. Participants completed the PANAS plus additional items once a day for 56 consecutive days. The PANAS includes 10 PA and 10 NA items. Participants responded to each adjective by indicating the extent to which they felt each emotion (Watson et al., 1988) on a 5-point Likert-type scale (1 = *Not at all* to 5 = *Extremely*).

Twenty-two items (12 PA; 10 NA), most taken from the Circumplex Model of Emotion (Larsen & Diener, 1992), were added to the PANAS adjectives. These items were posited to aid in representing a wider range of the affect domain, measuring PA and NA more comprehensively.

Two PA items, Passive[2] and Still[2], along with two NA items, Scared[1] and Lonely, were removed to avoid degrading measurement after preliminary IRM results indicated these items demonstrated severe lack of fit (infit and/or outfit mean square values above 1.7; Bond & Fox, 2007) to the PCM on at least 30 of the 56 occasions. Twenty PA adjectives and 18 NA adjectives were included in the final scales. These scales will be referred to as PA and NA, though it should be noted they are different from the original PANAS subscales. PA adjectives included Active[1], Calm[2], Alert[1], Attentive[1], Elated[2], Determined[1], Strong[1], Stimulated[2], Happy[2], Enthusiastic[1], Excited[1], Love, Proud[1], Joyful, Interested[1], Pleased[2], Content[2], Aroused[2], Inspired[1], and Euphoric[2]. NA adjectives included Afraid[1], Unhappy[2], Annoyed[2], Ashamed[1], Guilty[1], Angry, Sad[2], Hostile[1], Upset[1], Irritable[1], Depressed, Jittery[1], Drowsy[2], Fatigued, Sluggish[2], Worried, Nervous[1], and Distressed[1]. The final set of PA items evidenced high reliability with the present sample, both overall (Cronbach's $\alpha = .98$, Guttman's $\lambda_6 = .98$) and on a single occasion

[1]Indicates original PANAS items (Watson et al., 1988).
[2]Indicates items taken from the Circumplex Model of Emotion (Larsen & Diener, 1992).

in the middle of the measurement period (day 28; Cronbach's α = .98, Guttman's λ_6 = .99). The final set of NA items also showed very good reliability overall (Cronbach's α = .96, Guttman's λ_6 = .97) and on day 28 (Cronbach's α = .96, Guttman's λ_6 = .97).

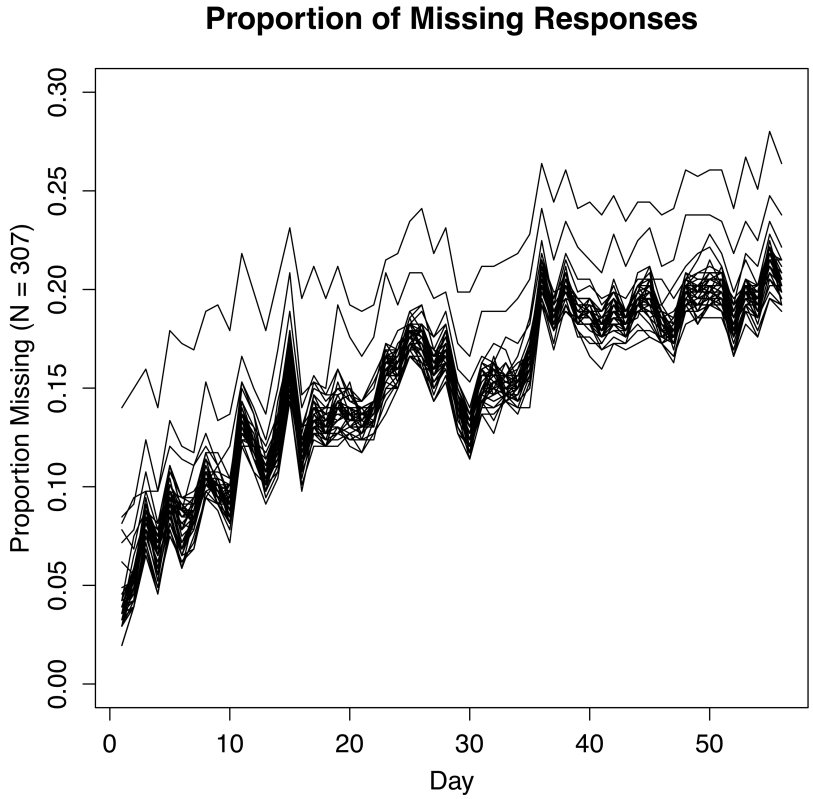

Figure 1. Proportion of missing responses by occasion for each item (each line) in the NDSHWB data.

Procedure and missing data. Participants were instructed to rate their experience of the 42 affect adjectives over the past day every evening for 56 consecutive days. The median number of missing responses across all items on the first occasion was 12 (*M* = 13.2) or 3.9% (*N* = 307). The number of missing responses rose somewhat steadily across occasions for all items (see Figure 1) to a median number of 51 (16.6%) on day 28 and a median number of 62.5 (*M* = 63.0) or 20.4% of responses missing on the last occasion.

The median number of missing responses across all occasions and items was 49 (*M* = 47.2) or 16.0% of responses.

Maastricht Project

Participants. Adult female twins and triplets (*N* = 579), monozygotic and dizygotic, 18 to 61 years of age (*M* = 27.8 years, *SD* = 7.9, *Median* = 26.0) were recruited. To avoid adding dependence between participants to the already lengthy proposed statistical analyses, and to retain a sample for cross-validation of the results found in this project as a part of future research plans, a random subset of unrelated participants was chosen for the present investigation. For every twin or triplet group, one individual was selected at random for inclusion in the present project analyses, thereby eliminating dependence between participants. The selected subset of adult women (*n* = 267) ranged in age from 18 to 46 years, (*M* = 27.2 years, *SD* = 7.4, *Median* = 25.0; age data was missing for three participants). About 20% of the selected participants had either a current or past depression diagnosis (*n* = 54). Income varied across participants, with 33.3% earning less than $20,000 annually, 7.1% earning between $20,000 and $39,999, 12.4% earning between $40,000 and $59,999, 8.2% earning between $60,000 and $79,999, 21.0% earning between $80,000 and $99,999, 14.6% earning between $100,000 and $149,999, and 2.6% earning $150,000 or more. Over one third of the sample (34.5%) was married, 27.7% were in a relationship, 20.2% were single, 15.7% were cohabitating with a partner, and 1.1% were divorced. Marital status was missing for 2 participants.

Measures. At each measurement occasion, 16 affect adjectives were administered to participants. Eight items were included to measure PA: Cheerful (Opgewkt), Satisfied

(Tevreden), Enthusiastic (Enthous), Energetic (Energiek), Pleased (Plezier), Alert (Duidelijk), Active (Actief), and Calm (Rustig). Eight items were administered to measure NA: Guilty (Schuldig), Unsure (Onzeker), Lonely (Eenzam), Anxious (Angstig), Gloomy (Somber), Suspicious (Wantrou), Angry (Boosgei), and Tired (Moe). Participants indicated how strongly they felt each component of affect on a 7-point Likert-type scale (1 = *Not at all* to 7 = *Very much*). PA items showed reasonable reliability with the present sample, both overall (Cronbach's $\alpha = .75$, Guttman's $\lambda_6 = .79$) and on a single occasion in the middle of the measurement period (beep 20; Cronbach's $\alpha = .75$, Guttman's $\lambda_6 = .79$). NA items also demonstrated sufficient reliability overall (Cronbach's $\alpha = .70$, Guttman's $\lambda_6 = .75$) and at beep 20 (Cronbach's $\alpha = .76$, Guttman's $\lambda_6 = .82$).

Procedure and missing data. Participants were given a wristwatch beeper to alert them 10 times throughout the day when they should complete the affect assessment. The waking hours of the day were split into ten, 90-minute non-overlapping blocks. In each block, a time was selected at random for the participant to complete the assessment.

There was no visual evidence of time-related trends in the proportion of missing responses on each item across occasions (see Figure 2). The median number of missing responses across all items on the first occasion ($M = 5.6$, *Median* = 4 or 1.5%) was similar to the median number of missing responses on occasion 25 ($M = 3.2$, *Median* = 2 or 0.7%) and on occasion 50 ($M = 2.1$, *Median* = 1 or 0.4%). Across all occasions and items, the median number of missing responses was 2 ($M = 3.2$) or 0.7%.

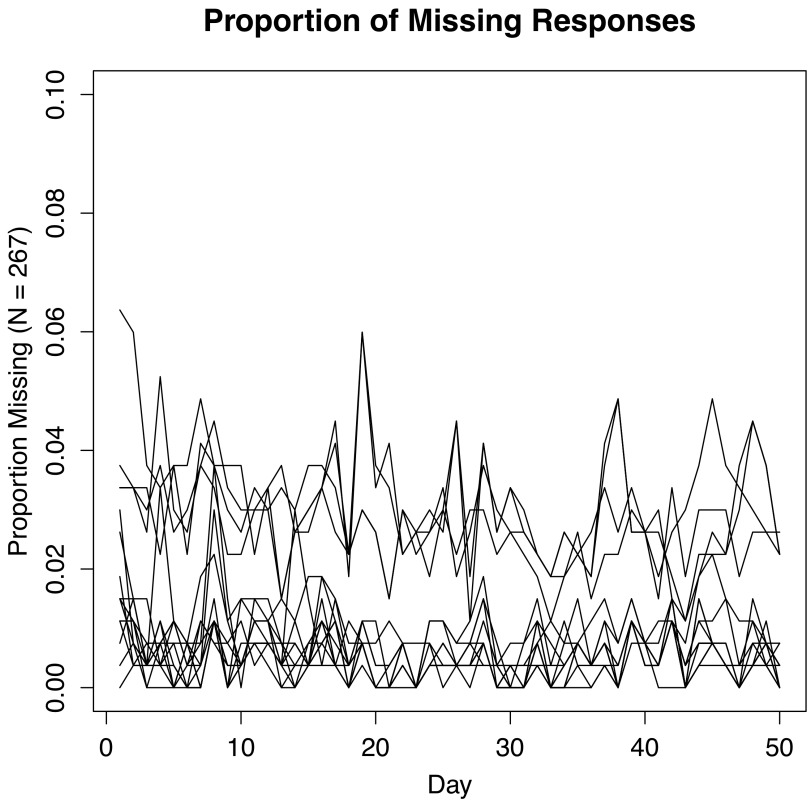

Figure 2. Proportion of missing responses by occasion for each item (each line): Maastricht data.

Chapter 5: Testing Measurement Assumptions

Data Analysis

In this section, data analyses used to test the measurement assumptions commonly made about PA and NA are discussed, followed by the results obtained from these analyses. First, IRMs are reviewed, followed by IRM estimation, and the application of IRMs to longitudinal data via anchoring. Next, executed psychometric analyses are described in detail, including anchored IRM analyses and comparisons of collapsed response patterns. Analyses of the individual variability assumptions and their resultant findings are discussed in subsequent sections.

Introduction to Item Response Models (IRMs)

Item response models (IRMs) were largely introduced into psychology by Birnbaum's contribution to Lord and Novick's (1968) textbook on statistical test theories. IRMs are latent linear, often logistic models in which the probability of a particular response, such as a correct response or a specific category of a rating scale, is modeled as a function of a person's position on the latent dimension being measured and the difficulty of the item (or the response category) to which the person is responding. These models emerged as part of a new theory of measurement, Item Response Theory, building on previous measurement theories, mainly Classical Test Theory (CTT; Hambleton & Swaminathan, 1985; Embretson & Reise, 2000).

Several IRMs fall into the category of Rasch-based models (e.g., see Fischer & Molenaar, 1995). The simplest Rasch-based IRM is called the Rasch model (Rasch, 1960), and was developed for dichotomous data. Location parameters are estimated for

each person and item on the latent dimension measured from individuals' responses to a set of dichotomous items. One of the most useful properties of IRMs, including the Rasch Model, is that person location (theta) parameters are estimated on the same equal interval scale of measurement, a scale of log-odds units or logits, as item location (beta) parameters, thereby allowing direct comparisons between items and persons.

This valuable property of IRMs allows for the determination of which individuals are measured most accurately. Individuals with one or more items close to them on the latent dimension will be measured more accurately than individuals with no items close to them on the latent dimension. For these latter individuals, each item will either be too easy or too difficult to endorse, and thus responses that follow model predictions are less informative. Comparing the distribution of persons in an IRM to the distribution of items permits the identification of gaps where more items are needed. In the proposed project, we can determine if additional items are needed for measuring PA and/or NA at specific locations on the latent dimensions being measured.

Polytomous: The Partial Credit Model. Building on the idea of the Rasch model, Masters (1982) developed the Partial Credit Model (PCM) for use with polytomous items. Instead of a single difficulty parameter for each item, a set of thresholds is estimated. Each threshold acts as a difficulty parameter for every pair of adjacent response categories; that is, each threshold parameter indicates the point on the logit scale at which two adjacent categories are equally likely to be chosen or observed.

The equation for the PCM is as follows:

$$P(x_{in} = X \mid \theta_n) = \frac{\exp(\sum_{j=0}^{x}(\theta_n - \delta_{ij}))}{\sum_{r=0}^{m_i}\exp(\sum_{j=0}^{r}(\theta_n - \delta_{ij}))}, \quad (1)$$

where *j* is the threshold index from zero to one less the total number of possible response categories, and *r* and *x* are the current threshold. The index upper bound, *m,* refers to the full set of thresholds for each item. Instead of modeling the probability of a correct response, or a response coded as 1, the PCM equation models the probability of choosing a specific category. The equation models the probability of individual *n* responding to item *i* with category *X* as a function of the individual's theta score and the item's threshold scores, summed up to the current category and divided by the sum over all categories.

The PCM equation describes a set of curves for each item, with a probability curve existing for each response category. These curves are called category response curves (CRCs). As with IRFs for the Rasch model, each CRC indicates the probability of choosing that response category for various values of theta. CRCs can be plotted with item thresholds centered around either zero or the item's difficulty. For ease of explanation, assume that the example CRCs examined in this section are centered around the item's location which happens to be zero.

The CRCs for an example item with three possible responses under the PCM are plotted in Figure 3. The curves indicate how likely each category is to be chosen as a response, given an individual's location on the latent dimension (his or her theta score).

For example, a person with a theta score equal to 2.0 (right-most gray dashed lines in Figure 4) is equally likely to respond to the example item with the second (middle curve) and third (right curve) categories. Note the threshold parameter between the second and third categories is also 2.0. An individual with a lower score on the latent dimension, in this case a theta score of 1.0 (left-most gray dashed line), is most likely to respond to the item pictured with the middle response category. The same person would be less likely to respond with the third category, and least likely to respond with the first category.

The PCM is one of the more flexible polytomous IRMs. Another common polytomous IRM is the Rating Scale Model (RSM), introduced by Andrich (1978). The RSM is slightly more rigid in its assumptions about response scales. Under the RSM, spacing between categories can differ within an item (i.e., 1 and 2 can be farther apart than 2 and 3); however the set of distances between categories is assumed to be the same for all items. The RSM may be too strict for some affect items. For example, on a 5-point Likert-type scale (1 = *Not at all*, 2 = *A little*, 3 = *Somewhat*, 4 = *A lot*, and 5 = *Extremely*), the distances between response categories when responding to the NA item "Hostile" is probably much different than distances between categories for "Fatigued". The distance between response categories 3 and 4 may be much larger for Hostile than for Fatigued. Thus, a more flexible Rasch-based IRM, such as the PCM, is necessary for the present project.

Although the PCM is flexible in terms of distances between response categories on the latent dimension being measured, two principles are assumed in PCM analyses (see Embretson & Reise, 2000): 1) the data is unidimensional; and 2) observations are

locally independent (i.e., the only relationship between observations is that caused by the underlying dimension being measured). Therefore, each row in the data set (each individual) is independent of one another and each column in the data set (each item) is independent of one another, after controlling for the underlying dimension measured.

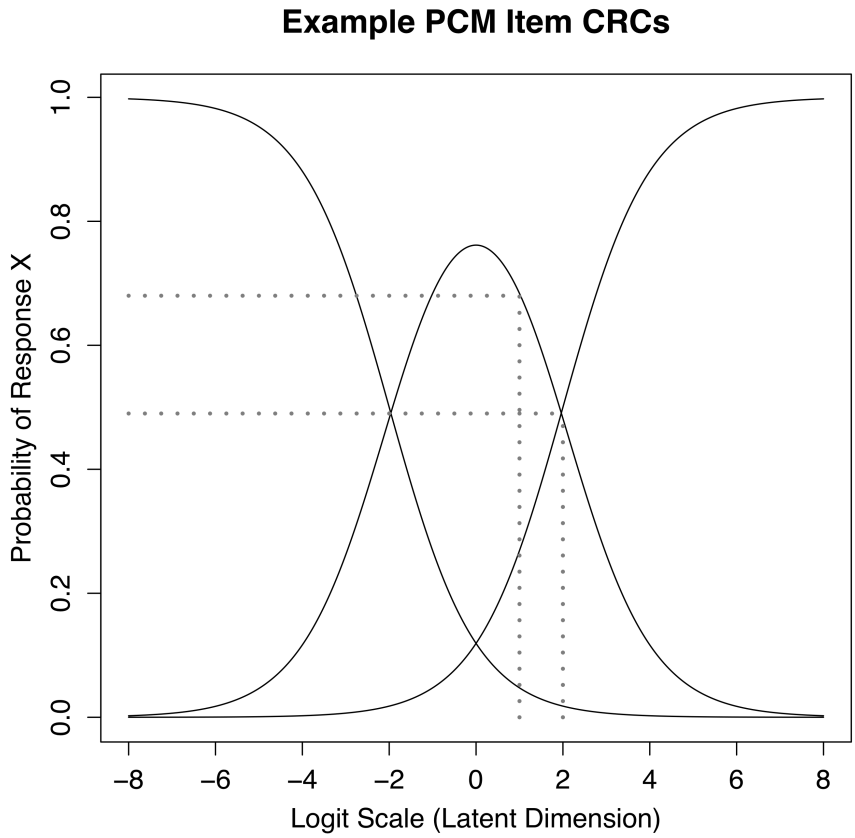

Figure 3. PCM CRCs for an example polytomous item.

This local independence assumption becomes a problem when attempting to apply the PCM to longitudinal data. Including more than one occasion of measurement for each person and each item results in either multiple rows for the same person or multiple columns for the same item in the data set being analyzed. Thus, either rows or columns

would no longer be completely independent after controlling for the underlying latent dimension. To apply the PCM to longitudinal data, anchoring techniques must be used.

Longitudinal IRM Analysis: Anchoring. Although some multi-occasion IRMs exist (e.g., Embretson, 1991; Jannarone, 2010), for most of these models it is only computationally feasible to analyze two or three occasions. To circumvent computational limitations, anchoring may be used to link many IRMs to the same logit scale.

Anchoring is accomplished by fixing item parameter values to be constant across all single-occasion analyses. Anchor values can be obtained using a variety of methods. For example, the estimated item parameters from the first measurement occasion could be used as anchors for all PCM analyses on all other occasions. In this example, if the PA item "Content" had a threshold set of -1.5, -0.5, 0.5, and 1.5 estimated from data on the first measurement occasion, these values would become the anchor values for "Content" in all other analyses. In each single-occasion PCM analysis, Content would have the same thresholds, the fixed values of -1.5, -0.5, 0.5, and 1.5. By repeating this process for all items, parameters from each PCM are forced onto the same logit scale. Recall that under the Birnbaum (1968) procedure, item and person parameters are standardized by the mean and standard deviation of the item parameters, resulting in a mean item parameter of zero and a standard deviation of 1. By fixing item parameters to be the same for a single item in each model, we are also fixing the mean and standard deviation of the item parameters for each model, in turn fixing the location of the logit scale for each model to be equivalent. Thus, estimated person parameters are also on this same logit scale across occasions, allowing for longitudinal analysis of person scores.

Although many different methods of obtaining anchor values can be employed, evidence suggests the best method involves obtaining a mean set of thresholds across separate-occasion IRMs for each item, using these mean values as anchors (Erbacher et al., 2012). In the present project, this mean anchor method is employed to analyze two longitudinal data sets of PA and NA responses with the PCM.

IRM Analyses

Each data set of longitudinal affect responses was analyzed by the process outlined below, with PA items analyzed separately from NA items in each data set. The following major steps were completed: (1) Mean parameter values across all occasions were computed for each item; (2) These mean values were used as anchors in PCM analyses linked across all occasions, and longitudinal measurement characteristics of the items were evaluated; (3) Separate single-occasion PCM analyses were used to examine the stability of item measurement properties over time; and (4) To evaluate response scale performance for PA and NA responses, all possible ways of collapsing response categories were compared to the original response scales employed in administration. Procedures for each of these steps are described in detail below.

Four sets of items were analyzed separately according to the four steps described below: PA items from the NDSHWB; NA items from the NDSHWB; PA items from the Maastricht project; and NA items from the Maastricht project. For brevity, procedures are detailed as though they were applied only to PA items in the NDSHWB data set.

Longitudinal PCM analyses. First, responses from all occasions were entered into one PCM, such that administrations of the same item on different occasions were

treated as separate items. For example, in this model, the item Happy administered on the first occasion was analyzed as one item and Happy administered on the second occasion was analyzed as a different item, independent from the first. From this model including responses from all occasions, each item received *t* sets of estimated deltas (thresholds), where *t* is the total number of measurement occasions (*t* = 56 in the NDSHWB data and *t* = 50 in the Maastricht data). A mean set of deltas was calculated for each item from all *t* sets of deltas.

Second, these sets of mean delta parameters were used as anchors to link PCM analyses across occasions. In this step, each occasion of responses was analyzed with a separate PCM. The deltas for a given item were fixed at the obtained mean delta values for that item in all single-occasion PCM analyses. Suppose the set of mean deltas for the item Happy consisted of the values -2.5, -1.5, -0.5, and 0.5. In the PCM of PA responses on the first occasion, and in all other single-occasion PCMs, the delta parameters for the item Happy were not estimated, rather they were fixed to -2.5, -1.5, -0.5, and 0.5. These constraints forced item parameters across all occasions onto the same scale, allowing for the examination of longitudinal psychometric properties of items.

Once the longitudinal psychometric properties of the items had been investigated, the cross-sectional measurement characteristics of the same items were obtained for comparison. In this third step, responses from each occasion of measurement were analyzed with a separate PCM, with all deltas freely estimated and no anchors used. Thus, the deltas estimated for the item Happy in the PCM of responses on the first occasion were permitted to differ from the deltas estimated for Happy in all other single-

occasion PCMs. This procedure resulted in parameters for a given item estimated on different logit scales by occasion. Although this prevented direct comparison of parameter values, the rank order of item betas and deltas were comparable across occasions, providing information on the temporal stability of measurement characteristics.

Collapsed response scales. Finally, a fourth set of analyses was carried out to determine the ideal set of response categories for each group of items. All possible ways of collapsing a 5-category (for NDSHWB data) and a 7-category (for Maastricht data) Likert-type response scale were tested against the original response scale in each study.

Collapsed response scales were constrained to be the same for all items of the same type of affect, administered in the same study (e.g., a single collapsed pattern was used for all PA responses in the NDSHWB data set). If the five response categories for Happy were collapsed down to three categories, such that the lowest two categories were recoded as "1," the middle category as "2," and the highest two categories as "3," then responses to all PA items in the same data set were also recoded as 1, 1, 2, 3, 3, respectively. This constraint prevented employing different response scales for different items, a task that would likely cause an unreasonable cognitive load for respondents if employed in future administrations.

To test collapsed response patterns against the original response scale, data recoded under each collapsed pattern were analyzed with cross-sectional PCMs. Fit statistics, reliability and separability estimates, and parameter-summed score correlations for the collapsed and original response scales were compared to identify the response

scale with the best overall psychometric properties for each type of affect in each of the two data sets. Best-performing scales were compared across data sets and affect types. Such comparisons provided information on the psychometric differences between PA and NA, as well as differences in psychometric properties of affect items administered over varying time scales. Importantly, data were recoded according to the best-performing collapsed response scales before being used to test the ergodicity of PA and NA.

Estimation. Joint Maximum Likelihood (JML) estimation (Birnbaum, 1968; Wright & Panchapakesan, 1969) was used when conducting all PCM analyses. A brief review of Birnbaum's JML estimation procedure is given below (see Baker & Kim, 2004).

In JML estimation, item threshold and person theta parameters are iteratively updated to maximize the likelihood of the observed data given the model. The probability of the observed responses modeled by the PCM equation can be converted to the following likelihood equation (adapted from Baker & Kim, 2004):

$$P(U|\theta,\delta) = \prod_{i=1}^{I} \prod_{n=1}^{N} \prod_{j=1}^{J_i} P_{inj}^{y_{inj}}, \quad (2)$$

where $P(U|\theta,\delta)$ is the probability of the observed data, given the estimated theta and delta parameters, i is the index for items, n is the index for persons, and j is the index for categories on the i^{th} item. The probability function P_{inj} is the probability of an observed response, given by the PCM equation (see Equation 1), and y_{inj} contains a dichotomous recoding of the original data, such that y is equal to 1 if an observed response is equal to category j and y is equal to 0 if the observed response is not equal to j. This recoding is

conducted for all *J* categories. Thus, when indexing over categories, if a response is equal to the current category, the probability of the response is calculated according to the PCM. If the response is not equal to the current category, *y* becomes 0 and the probability raised to the 0^{th} power becomes 1, effectively dropping out of the multiplication.

In JML estimation, the log of this function is maximized. In order to find the maximum of the log-likelihood equation, we apply the Newton-Raphson formula (see Equation 3 below), a function of the first and second partial derivatives of the log-likelihood equation with respect to each parameter. The formula can be written as:

$$A_{x+1} = A_x - B_x^{-1} F_x \ , \tag{3}$$

where *x* indexes the iteration, *A* is the matrix of parameter estimates, *B* is the matrix of partial second order derivatives of the log-likelihood function, and *F* is the matrix of partial first order derivatives of the log-likelihood function. The assumptions of local independence and unidimensionality cause the matrix of partial second derivatives to be a sparse diagonal matrix, allowing diagonal elements to be evaluated separately.

At each iteration of this estimation procedure, item and person parameters are placed on the same scale via Birnbaum's three-step paradigm. First, item parameters are fixed either at chosen starting values or at values estimated in the previous iteration, and the person parameters are optimized. Second, person parameters are fixed at these most recent estimated values and item parameters are optimized. Third, both person and item parameters are standardized by subtracting the mean of the item parameters from each and dividing each by the standard deviation of the item parameters. Thus, both item and person parameters are standardized onto the same logit scale.

Other common estimation procedures include Conditional Maximum Likelihood, in which total test scores, rather than theta parameters, are used in the calculation of the response pattern likelihood (CML; see Linacre, 2012) and Marginal Maximum Likelihood, in which expected population response pattern frequencies are used, rather than observed frequencies, and the EM algorithm is used in parameter optimization (MML; see Linacre, 2012). CML is less flexible than JML with missing data, and MML requires distributional specifications about persons that would likely be problematic for NA thetas. Although JML can yield biased estimates, particularly standard errors, for surveys with very few items that span a wide range of the logit scale (e.g., 5 items over an 8 logit item parameter range; Linacre, 2012). IRM statistical software packages (e.g., WINSTEPS developed by Linacre, 2012) provide adequate methods of correcting for this bias. The shortest measures in the present data sets are the Maastricht affect measure, with 8 items spanning approximately 5 logits, resulting in bias near zero in the executed analyses. Thus, JML was chosen to estimate IRM parameters in the present project.

Summary of IRM analyses. To review, executed IRM analyses include: 1) PCM analyses anchored across occasions to examine longitudinal psychometric properties of affect items; 2) single-occasion PCM analyses to examine the stability of psychometric properties of affect items over time; and 3) single-occasion PCM analyses comparing all possible patterns of collapsing categories to determine the best-performing response scale for affect items. Analyses were conducted for PA and NA in each data set separately, resulting in four sets of analyses: PA items in the NDSHWB data; NA items in the NDSHWB data; PA items in the Maastricht data; and NA items in the Maastricht data.

Chapter 6: Measurement Assumption Results

IRM Results

Results appear in the same order as their corresponding analyses in the previous section. First, output from longitudinal PCM analyses is examined, followed by results of cross-sectional PCM analyses indicating the psychometric stability (or lack thereof) of PA and NA, and, last, comparisons between statistics from cross-sectional PCMs of all possible collapsed response scales are made.

Longitudinal Measurement Characteristics: PA versus NA

The first measurement assumption commonly made about affect examined here is the assumption that PA and NA have similar desirable psychometric properties. PA and NA are most often analyzed with the same statistical model, such as confirmatory factor analysis, and differences in psychometric properties, including correlated errors and factor loading magnitudes, tend to receive very little acknowledgment. To challenge this assumption, longitudinal measurement characteristics of PA and NA items are explored below. Specifically, the (mis)match between items and participants on the latent dimension, category usage, and measures of fit, error, and reliability are examined.

Matching items to participants. Person-item maps and information curves were examined to determine how well each affect measure targeted the sample to which it was administered. Person-item maps provide a direct visual comparison of person locations to item locations, allowing for the identification of gaps where additional items are needed. Item information indicates the slope of the expected value function for an item across the entire domain of possible theta values. The steeper the slope at a given value of theta, the

larger a difference in response probabilities there is for two given values of theta. Thus, the steeper the slope of the expected value function at a given theta value, the better we are able to distinguish between individuals with different theta values, and the more information we gain at that theta value. Fisher information for an item at a given theta value is calculated using the following formula (Linacre, 2005).

$$I(\theta) = \sum_{k=0}^{m} (k - E(x_{in} | \theta_n))^2 P(k), \quad (4)$$

where k indexes over the categories of the item, $P(k)$ is the probability of observing category k for the item under the PCM, and $E(x|\theta_n)$ is the expected value of the observed response from person n to item i, defined as

$$E(x_{in}) = \sum_{k=0}^{m} kP(k) \ . \quad (5)$$

Calculating information for a given item across all theta values yields an information curve indicating which interval(s) of theta an item targets well. The sum of the item information curves within a single survey is equal to the information curve for the entire survey. Person-item maps and item and survey information curves are examined below.

NDSHWB. The person-item maps for PA items on days 15 and 35 and for NA items on the same days are located in Figures 4 and 5, respectively. In each person-item map, the line through the middle of the plot represents the logit scale on which item and person parameters are estimated. Participants are plotted by theta parameter to the left of this line, and items are plotted by beta parameter to the right. If all individuals are being targeted well, items will span the entire distribution of thetas without any major gaps.

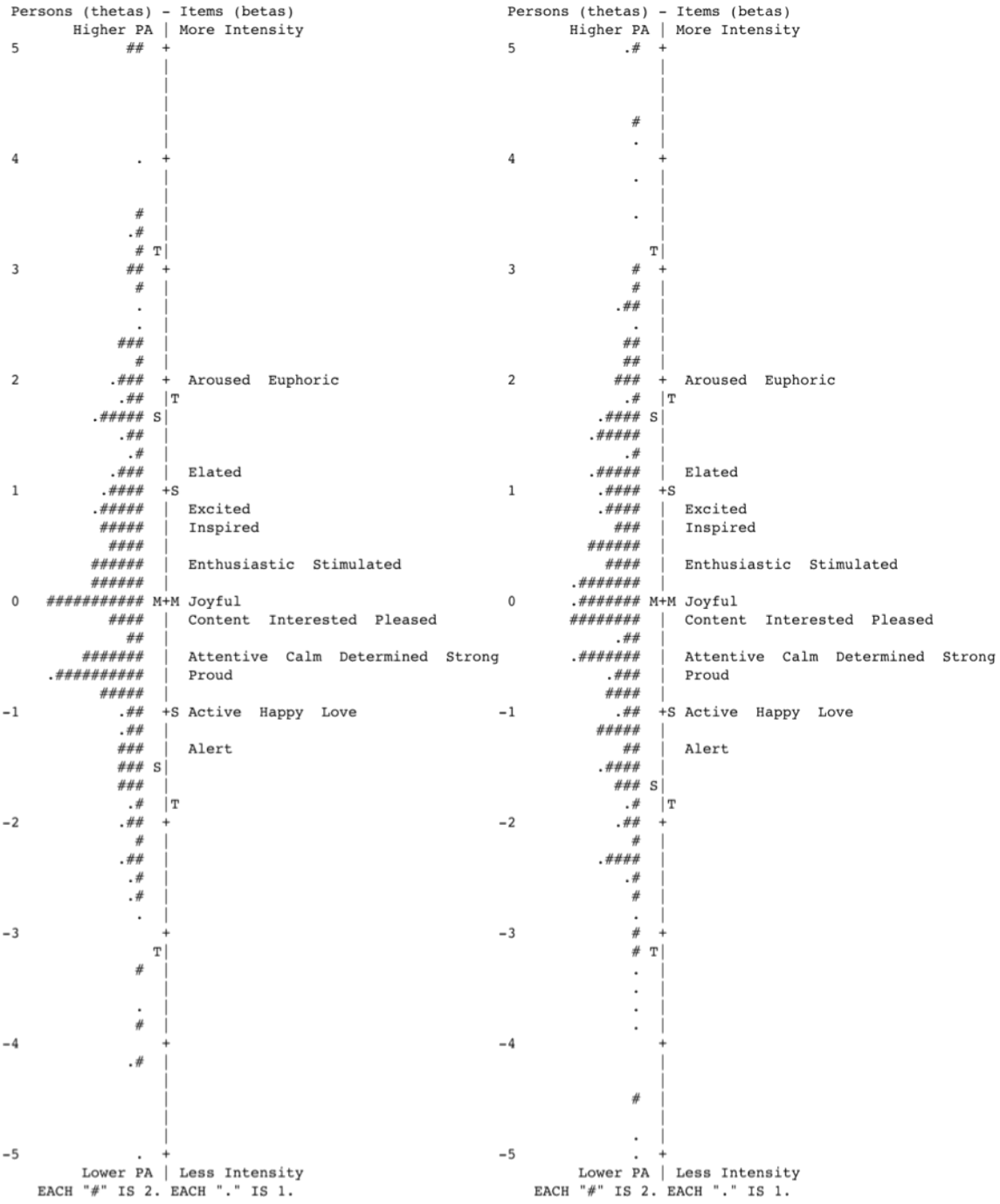

Figure 4. Person-item maps for PA items in the NDSHWB from longitudinally anchored PCMs on Days 15 and 35.

PA items in the NDSHWB function very well. More items are needed at each of the extremes of the latent dimension in order to span the entire person distribution, and an additional item with a higher beta than Elated and a lower beta than Aroused would be beneficial; however, the vast majority of the participants have at least one item well-matched to their own theta scores on both measurement occasions. Recall the beta for any given item is fixed across occasions in these longitudinal PCMs, thus the distribution of participants may change over occasions but the distribution of items is constant.

NA items do not perform nearly as well. Most of the NA items are located much higher on the latent NA dimension than any of the participants. For over 50% of participants on day 15 and day 35, the best-matched item has a beta at least one logit away from their thetas. Thus, the vast majority of the sample is not being targeted well.

These findings are reflected in the item and survey information curves for PA and NA items in the NDSHWB data, plotted in Figure 6. The PA item information curves have peaks that collectively span much of the estimated theta parameter continuum, from approximately -3 logits to +3 logits, resulting in a survey (test or scale) information curve with a wide peak.

For NA, the item and survey information curves have much steeper, narrower peaks than those of PA items. The 20 PA items in the NDSHWB study provide adequate information about persons with a wider range of theta scores, whereas the 18 NA items examined provide more information about a more limited segment of the sample.

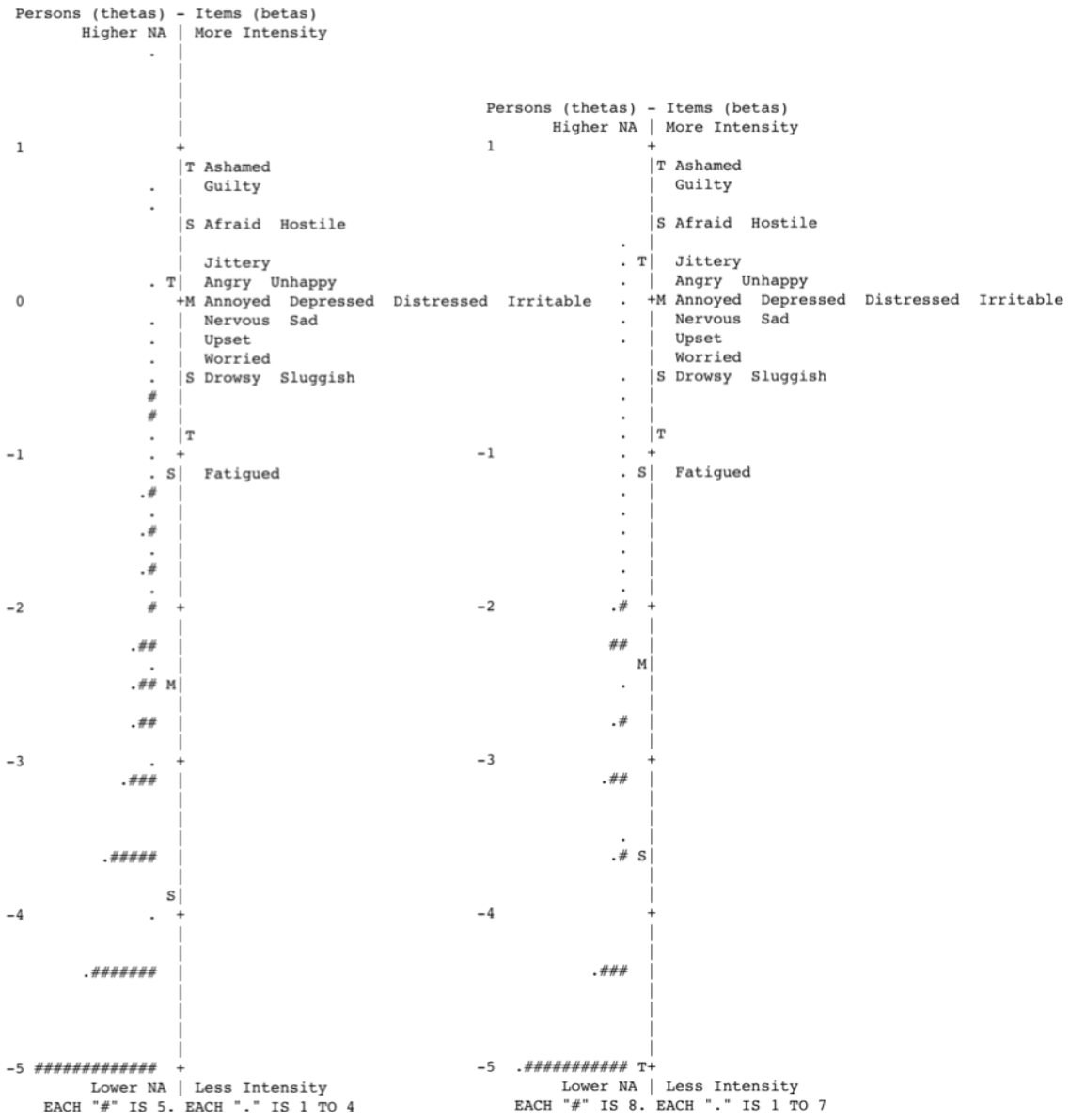

Figure 5. Person-item maps for NA items in the NDSHWB from longitudinally anchored PCMs on Days 15 and 35.

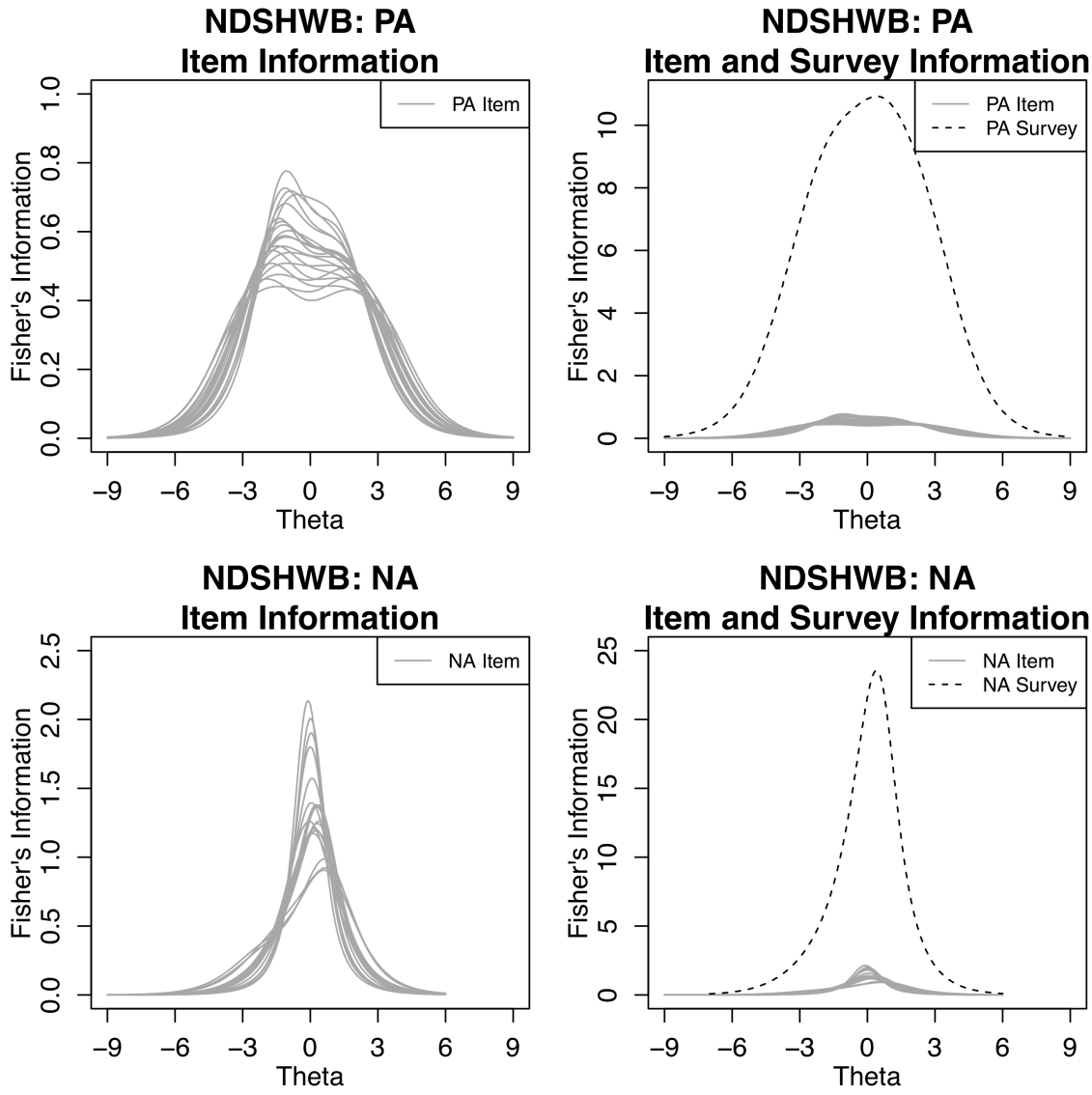

Figure 6. Item and survey information curves for PA and NA items in the NDSHWB from longitudinally anchored PCMs on Day 15.

Maastricht. Similar patterns exist in the results from the PCM analyses of the Maastricht data, with an obvious mismatch between NA items and participants. Only 8 PA items were administered, thus a few gaps exist between items in the person-item maps in Figure 7. For example, adding an item between Alert and Calm and another between Enthusiastic and Active would strengthen the overall match between items and

participants on the latent dimension. Adding items at either extreme of the latent dimension would also help to target more participants well. Despite these smaller gaps, PA items target much more of the sample well than NA items.

As in the prior data set, the NA items administered in the Maastricht study barely overlap with the distribution of participant locations in the person-item maps depicted in Figure 8. The item with the lowest location parameter, Tired, is still approximately half a logit above the mean location of the person distribution on both occasions. For over 35% of the participants on the 12th beep and over 25% of the participants on the 39th beep, the item targeting them the most is at least one logit away on the latent NA dimension. As in the NDSHWB data, the NA items administered in the Maastricht study target participants very poorly compared to the PA items.

Figure 7. Person-item maps for PA items in the Maastricht data from longitudinally anchored PCMs on beeps 12 and 39.

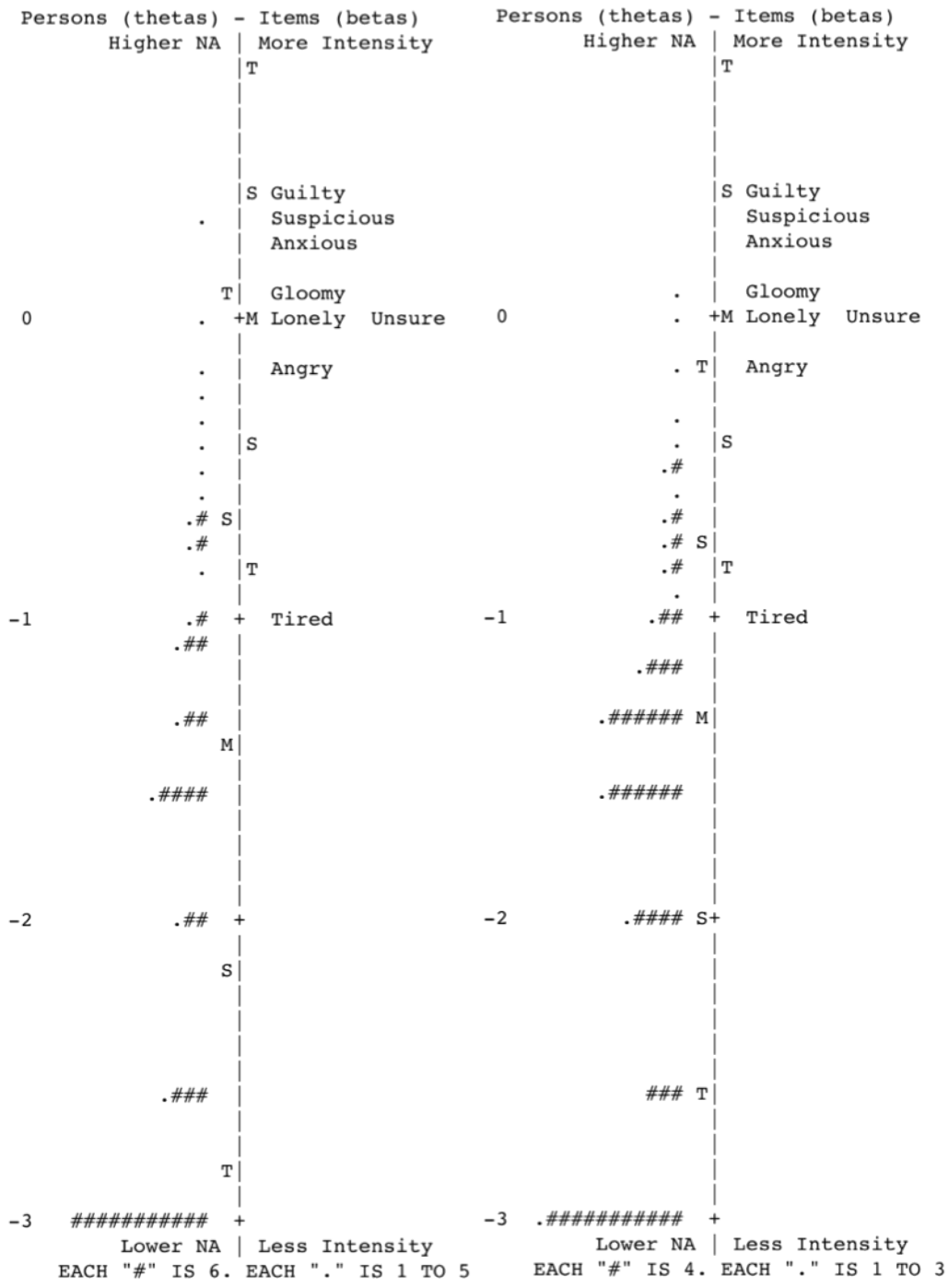

Figure 8. Person-item maps for NA items in the Maastricht data from longitudinally anchored PCMs on beeps 12 and 39.

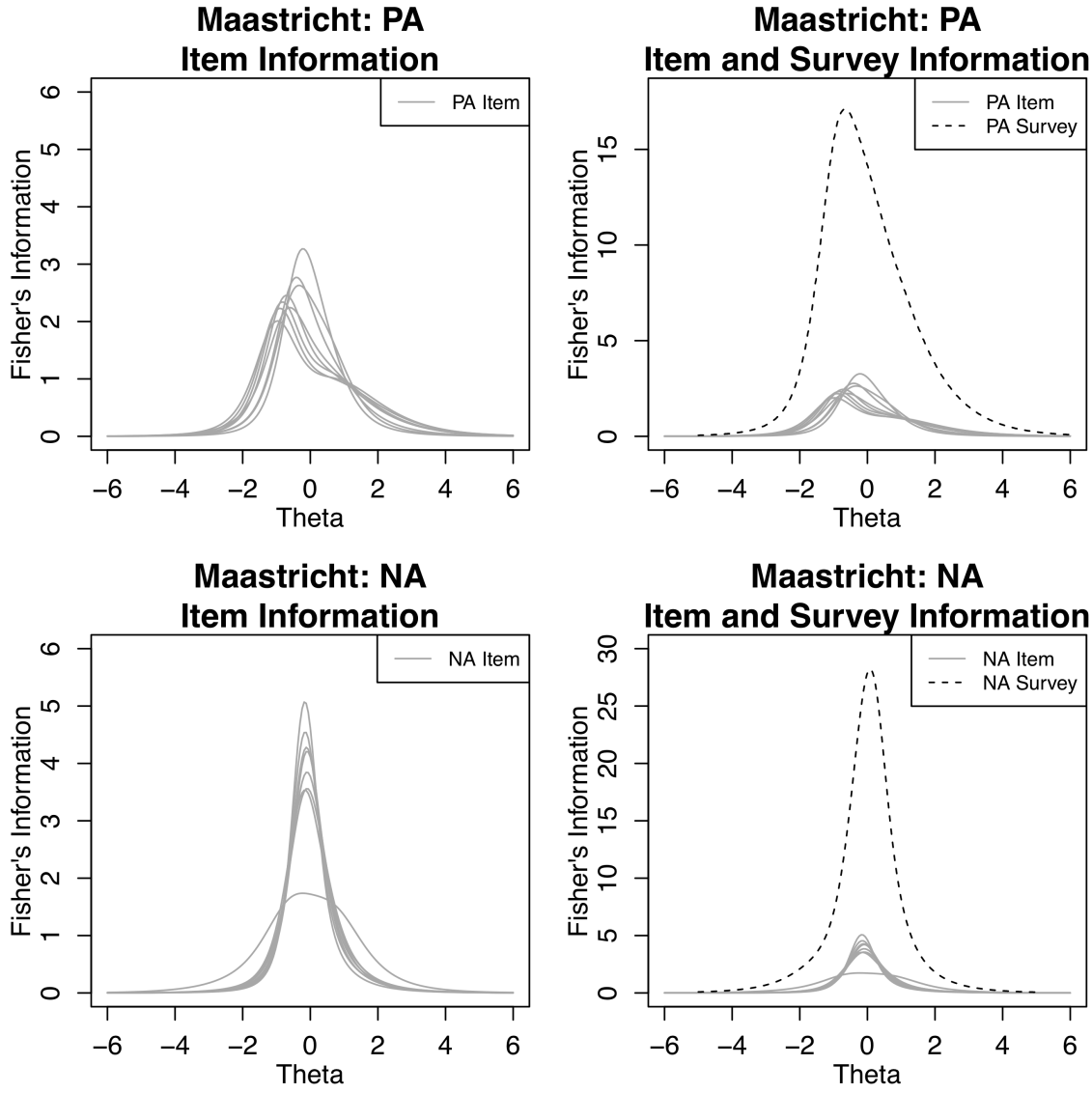

Figure 9. Item and survey information curves for PA and NA items in the Maastricht data from longitudinally anchored PCMs on occasion 12.

The same pattern was evident in the item and survey information curves for the Maastricht affect items (see Figure 9). The peaks of the PA items spanned a somewhat wider interval of the logit scale than those of NA items, resulting in a taller, but narrower, survey information curve for NA. Notably, the difference between PA and NA was not as large here as it was in the NDSHWB data, although it followed the same trend favoring

PA items. One NA item, Tired, had a desirably wider information curve. Tired had the lowest location parameter of all Maastricht NA items and thus better targeted participants.

Response scale use. To evaluate response scale use for PA items and NA items separately, histograms of response distributions, category response curves (CRCs), and threshold reversals were examined. Note that as a consequence of fixing item parameters across occasions, CRCs were also fixed across occasions, thus the CRC from only one occasion was examined for each of the items selected as examples below.

NDSHWB. Figure 10 illustrates the response distribution for two PA items, Happy and Calm, and two NA items, Sad and Jittery, in the NDSHWB data, along with their corresponding CRCs. PA response distributions much more closely resembled a normal distribution than NA responses. Each of the five response categories was endorsed by at least some participants in the NDSHWB for PA items and very little skew, if any, was detectable visually. Categories four and five were sometimes not endorsed by any participants for the NA items in Figure 10, and the response distributions for both items on both measurement occasions were highly skewed. Participants mainly used the lower half of the response scale for NA items.

The NA items also had much less desirable CRCs. Ideal CRCs look similar to those of the PA items pictured. Each category has some interval of the theta continuum for which it is the most likely response category to be observed. If we observed a response to the item Happy in the fourth category, we would predict the respondent has a theta score somewhere between 0.5 and approximately 2.5. Conversely, if we observed a

response in the fourth category to the item Jittery, we would be less certain of our theta score prediction, because there is no interval of theta for which the fourth category is the most likely response. The peak of the fourth category is dropped below the rest. Almost all of the NA items had at least one of these dropped peaks, whereas almost none of the PA items exhibited this problem. The 5-category response scale administered worked better for measuring PA than for NA.

Categories with dropped peaks, peaks that are never above all other category curve peaks, can result in the misordering of the delta parameters for an item, a phenomenon often referred to as disordered thresholds or disordered deltas. Three NA items had disordered deltas: Afraid, Ashamed, and Guilty. None of the PA items exhibited disordered deltas. This phenomenon may indicate a variety of issues with the measure and/or the model employed. At present, it is sufficient evidence against the assumption that PA and NA measures have similar desirable properties to identify the presence of disordered deltas for most NA items and few PA items.

Maastricht. The response distributions for PA and NA items in the Maastricht data, depicted in Figure 11, further supported the conclusions drawn from the NDSHWB data. PA response distributions all look similar to the example items displayed, Cheerful and Calm, across all occasions, with occasions 12 and 39 included in the figure. Responses to PA items were slightly skewed, but were much closer to being normally distributed than responses to NA items. As in the previous data, most participants were using only the lower half of the response scale for NA items, and the majority of respondents were endorsing the lowest category (1).

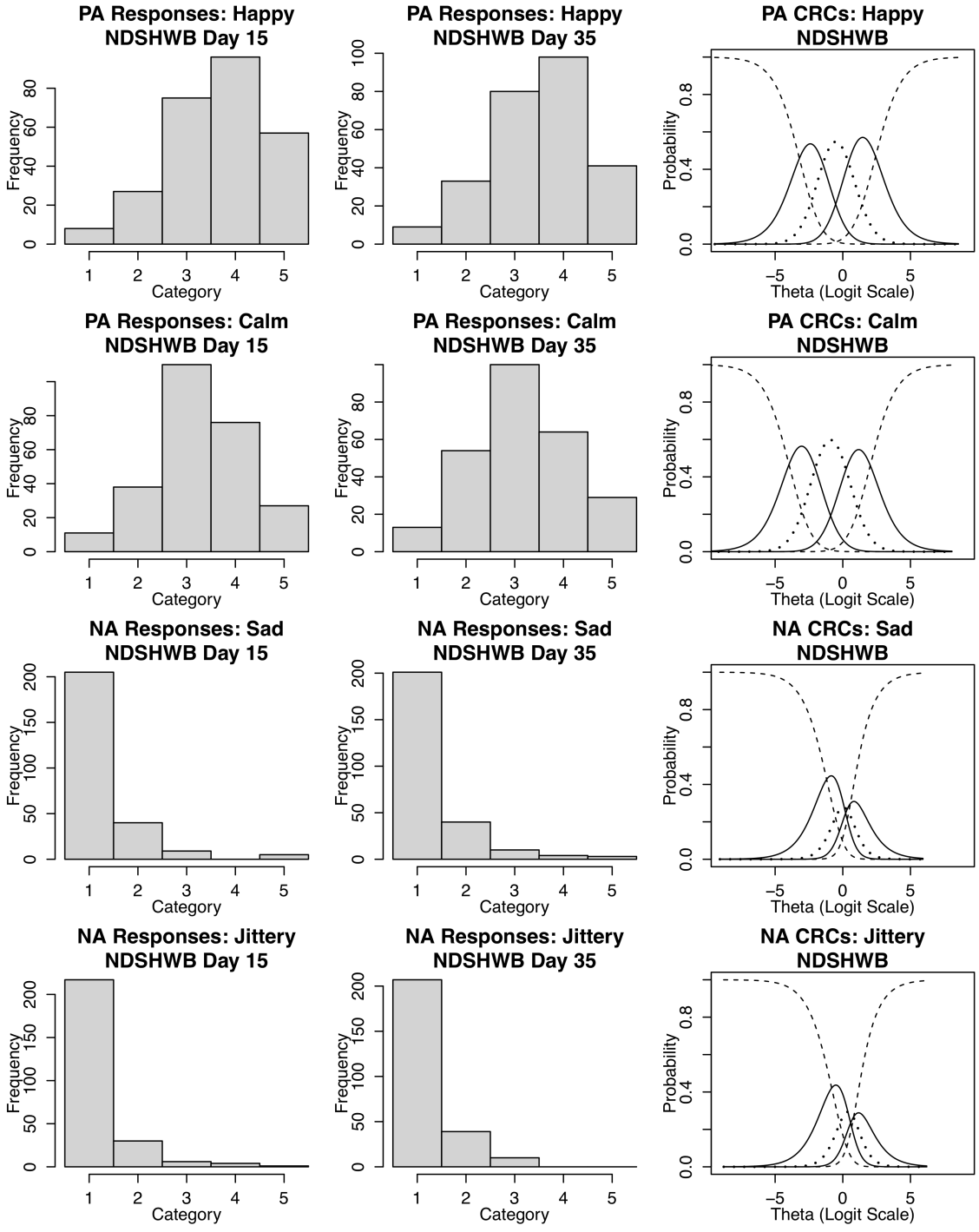

Figure 10. Response distributions and CRCs for two PA and two NA items from the NDSHWB.

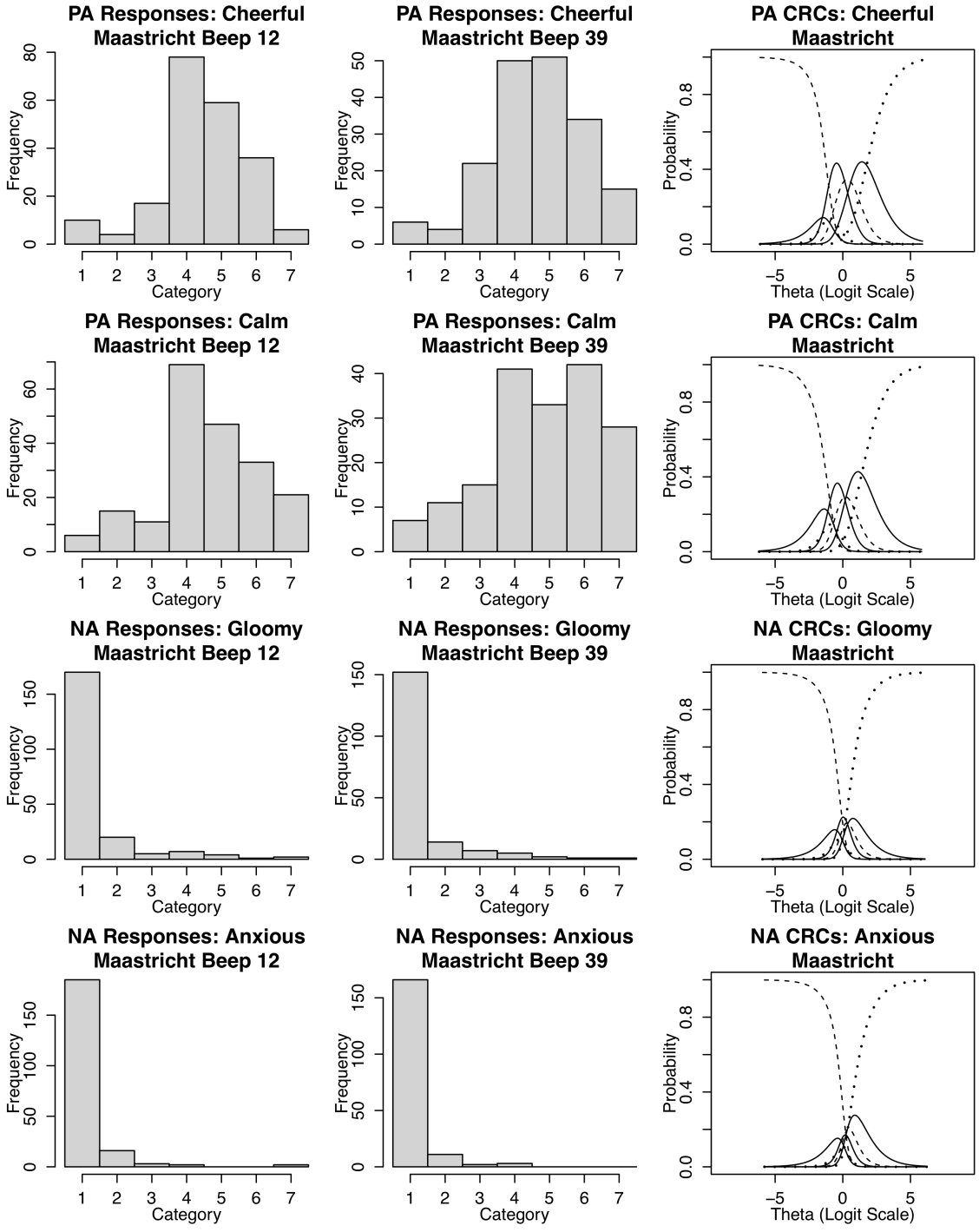

Figure 11. Response distributions and CRCs for two PA and two NA items from the Maastricht data.

The CRCs for the Maastricht affect items told a somewhat different story. Dropped peaks and disordered deltas were not only present for every NA item, but for every PA items as well. PA items tended to have approximately four or five clear peaks, appearing similar to Cheerful and Calm in Figure 11, whereas NA items tended to have two or three, similar to Gloomy and Anxious. Administering a response scale with too many categories and consequently expecting more precision than participants may use when evaluating affect could induce noise in the observed data, resulting in the dropped peaks and disordered deltas observed in the Maastricht data.

Regardless of the cause, it is clear administering the affect items in the Maastricht study with a 7-point response scale is problematic for measuring PA and NA. In these longitudinal anchored PCMs, every PA and NA item evidenced disordered deltas. Although these results suggest the psychometric properties of PA and NA measures may be somewhat similar to one another, such problematic characteristics also indicate previous work in the affect measurement literature may not be explicit enough about unfavorable psychometric characteristics of affect measures.

Fit statistics and error. Fit statistics and measures of error were evaluated to determine how well responses to PA and NA items fit the PCM, including the standard error (SE) of theta, and two chi-square statistics called infit and outfit mean squares. If PA and NA responses show differing magnitudes of misfit to the PCM, it is added support for the difference in psychometric properties between PA and NA measures.

Measures of fit. Item and person infit and outfit mean square statistics were examined to determine how well the PCM fit PA and NA responses for each item and

participant. Both mean-square fit indices are chi-square based statistics, with ideal values of 1. Infit mean square is calculated by Equation 6 below.

$$Infit\ MNSQ_i = \frac{\sum_{n=1}^{N}(x_{in} - E(x_{in}))^2}{\sum_{n=1}^{N}\sum_{k=0}^{m}(k - E(x_{in}))^2 P(k)} \quad (6)$$

For each person, n is an index over items and i indicates the person being evaluated, and for each item, n indexes persons and i indicates the item being evaluated. In the numerator, the difference between the observed response, x_i, and the expected response $E(x_{in})$ from Equation 5, or the residual associated with the observed response, is squared. This squared residual is compared to the expected residual, given the PCM, calculated by summing the squared difference between the expected value of the response and each possible category, k, multiplied by the probability of observing each category, $P(k)$, defined by the PCM in Equation 1. Thus, the infit mean square statistic is a ratio of the squared observed residual to the squared expected residual, indicating whether there is more or less misfit than expected given the model. Infit mean square is an inlier-sensitive statistic (Linacre, 2012): it is most sensitive to misfitting responses from well-matched person-item pairs.

Outfit mean square is very similar to infit mean square, with an additional division by the number of responses (persons or items, see Equation 7 below).

$$Outfit\ MNSQ_i = \frac{\sum_{n=1}^{N} \frac{(x_{in} - E(x_{in}))^2}{\sum_{k=0}^{m} (k - E(x_{in}))^2 P(k)}}{N} \quad (7)$$

Again, x_{in} denotes the observed response, $E(x_{in})$ represents the expected response, and $P(k)$ is the probability of observing category k given the item and person parameters associated with the observed response. For person i, summations are taken over all items, indexed by n. For each item, the indices are reversed, with i referring to items and n referring to persons. Outfit mean square is an outlier-sensitive statistic, meaning it is most sensitive to misfitting responses from poorly matched person-item pairs (e.g., an item with a beta parameter higher or lower than the theta parameter of the responder).

Several guidelines for infit and outfit mean square statistics have been developed (e.g., Bond & Fox, 2007, but see Smith, Schumacker, Bush, 1998 for further discussion of mean square fit). In the present investigation, Linacre's (2012) guidelines will take precedence, with mean square values above 2.0 taken as indicators of degraded measurement caused by misfit, and mean square values below 0.5 taken as indicators of overfitting. In accord with most guidelines, misfit will be considered a more serious problem than overfit.

NDSHWB. The proportion of items and persons with unacceptably high or low fit statistics on each occasion are plotted in Figure 12. As is often the case, a higher proportion of participants than of items showed infit and/or outfit statistics outside of

reasonable bounds. PA items exhibited a greater total number of instances of unacceptably high outfit across all occasions than NA items. The most problematic items were Calm, with outfit mean square above 2.0 on 26 of the 56 measurement occasions, and Arousal, with outfit mean square above 2.0 for 16 of the days. The most problematic NA item was Guilty, with outfit mean square above 2.0 for 11 of the 56 occasions. Very few instances of high item infit mean squares occurred for either type of affect.

NA items exhibited a higher proportion of outfit mean square values below 0.5, indicating overfit. This may partially have been an artifact of the mismatch between participants and NA items. All NA items were far away from the majority of participants on the logit scale. Thus, most participants were expected to respond to NA items with very low categories, resulting in expected responses very close to the observed data. Most participants did respond to NA items with the lowest category, causing residuals to be even smaller than expected under the PCM.

Similarly, higher proportions of the sample had outfit mean squares below 0.5 when choosing NA responses, while more participants had infit mean squares below 0.5 when evaluating PA. When responding to PA items, slightly higher proportions of the sample also evidenced unacceptably high infit and outfit mean squares (i.e., above 2.0). In summary, although NA responses were associated with overfit in terms of outfit mean square, PA responses were associated with higher proportions of persons and items with all other types of fit, including infit mean squares at both high and low extremes, and unacceptably high outfit mean squares.

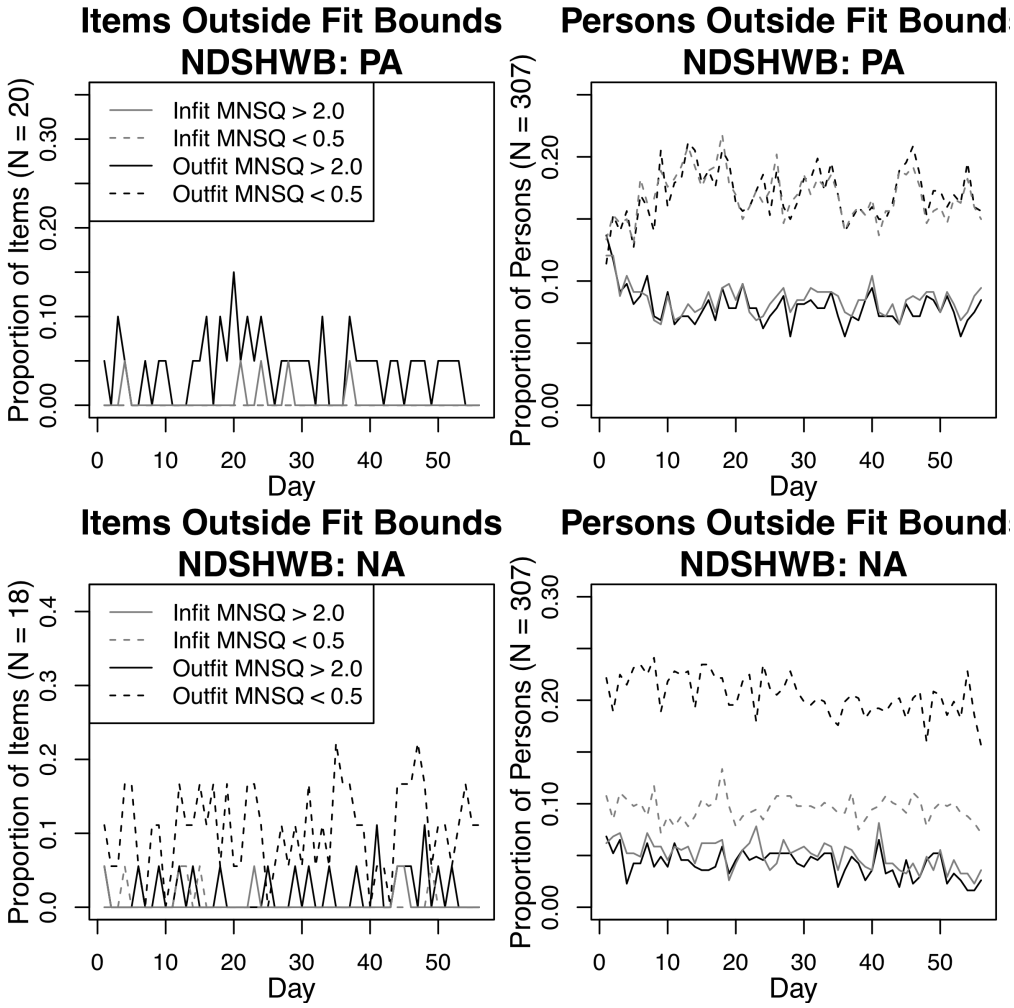

Figure 12. Items and persons with unacceptable fit statistics in the NDSHWB data.

Maastricht. The major differences between PA and NA person and item fit statistics in the Maastricht data occurred in unacceptably low mean square values (see Figure 13). Responses to NA items were associated with larger numbers of instances of overfitting, particularly in terms of outfit mean squares for both persons and items. The proportions of unacceptably high person infit mean squares, person outfit mean squares, and item infit mean squares were comparable across PA and NA. One PA item

contributed to higher proportions of unacceptably high item outfit mean squares, Active exhibited high outfit values on 18 of the 50 occasions.

Summary. The main difference between PA and NA in both data sets occurred in low outfit values. In both data sets, when examining fit associated with NA responses compared to PA responses, larger proportions of the item and person samples had outfit values indicating overfit to the PCM. This is likely at least partially an artifact of the severe mismatch between NA items and participants on the estimated NA logit scale.

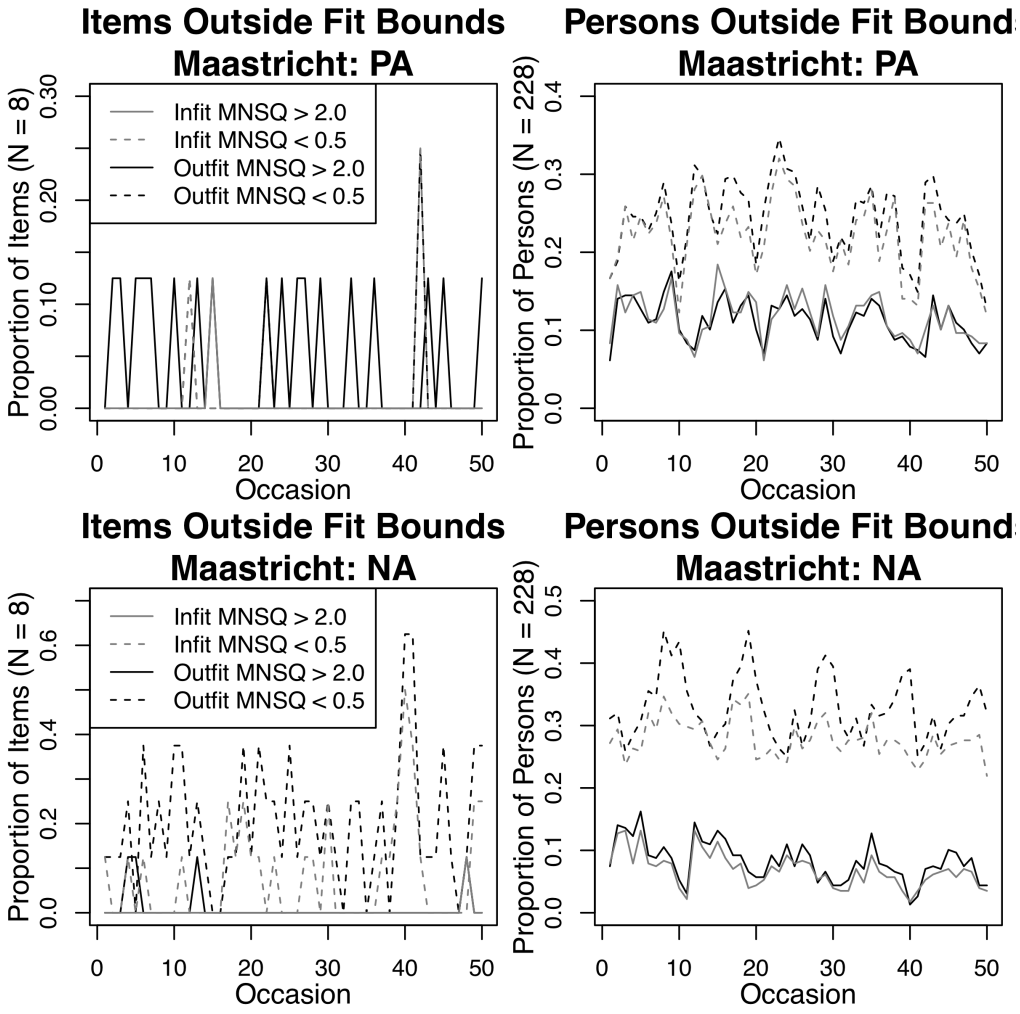

Figure 13. Items and persons with unacceptable fit statistics in the Maastricht data.

Measures of error. The standard error (SE) of theta and beta estimates provides an indication of the estimated parameters' precision by item and person, calculated in Equation 8.

$$SE_i = \frac{1}{\sqrt{\sum_{n=1}^{N}\sum_{k=0}^{m}(kP(k) - \sum_{k=0}^{m}kP(k))^2}} \quad (8)$$

The SE of an individual's theta estimate, with individual indexed by *i,* is summed over responses to all items, indexed by *n*. When calculating the SE of an item's beta estimate, indices change accordingly, such that *i* represents items and *n* indexes persons. In these longitudinal, anchored models, only SEs of thetas were examined here. Beta and delta parameters were fixed, and thus standard errors of item parameters were not a focus of this investigation.

NDSHWB. SEs tended to be much smaller for PA thetas than for NA thetas (see Figure 14 for the distributions of PA and NA theta SEs). The poor targeting of participants by NA items likely caused this. Participants with larger NA theta SEs, near 2.0, had low NA theta estimates, around -3.5, far from the targeted logit range of all NA items. Unsurprisingly, NA theta SEs were significantly correlated with NA theta estimates, $r = -.95$, $t(17190) = -390.88$, $p < .0001$. This correlation was weaker for PA theta estimates and SEs ($r = .25$, $t(17190) = 33.37$, $p < .0001$). in terms of person parameter accuracy, the PA measure in the NDSHWB data performed better than the NA measure.

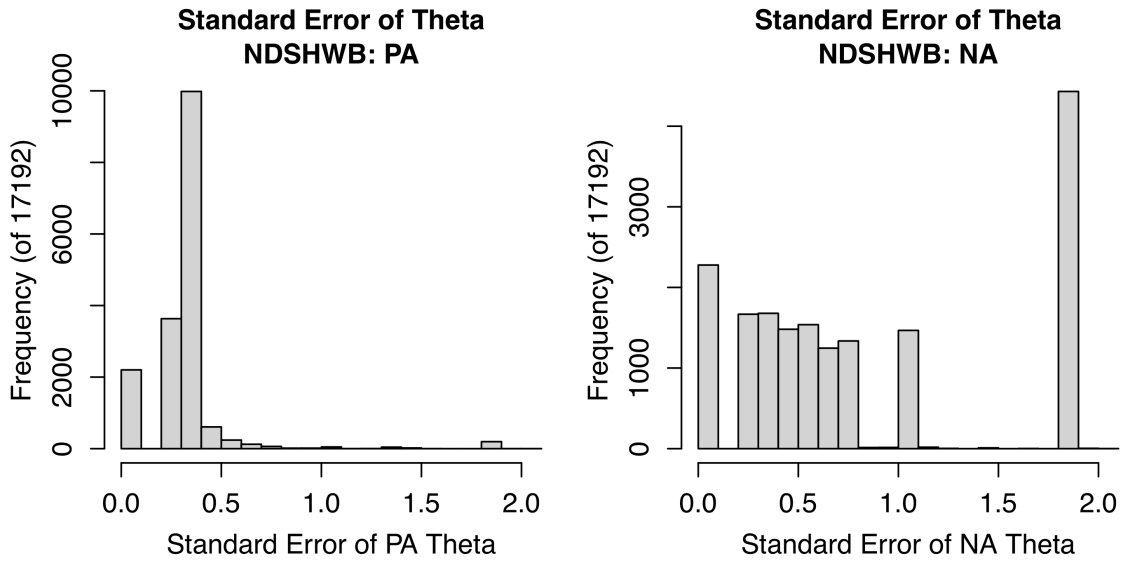

Figure 14. Standard errors of PA and NA thetas in the NDSHWB data.

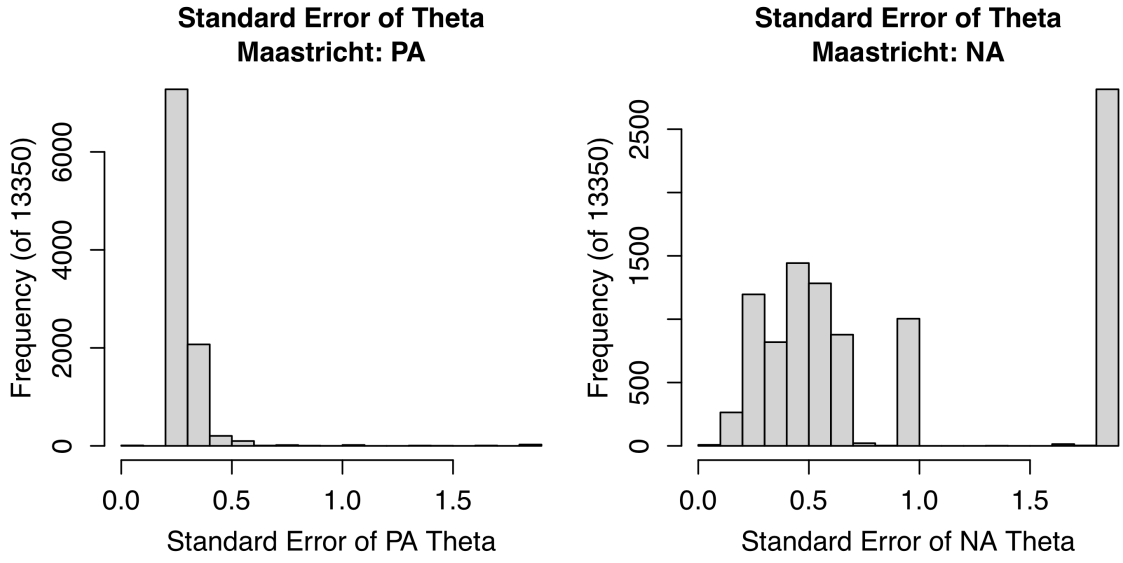

Figure 15. Standard errors of PA and NA thetas in the Maastricht data.

Maastricht. The SEs of the thetas estimated from the Maastricht data followed the same pattern as those of the NDSHWB data. The distribution of SEs for PA thetas over all occasions spanned a lower, smaller range than the distribution of SEs for NA thetas (see Figure 15). NA theta estimates and SEs were highly negatively correlated, r = -.98,

$t(9749) = 439.97$, $p < .0001$, indicating larger SEs for persons with low estimated NA scores, further away from most NA items. The correlation of PA theta estimates with SEs was moderate, $r = .57$, $t(9749) = 68.15$, $p < .0001$, but weaker in magnitude than that of NA. Administered PA items resulted in more accurate theta estimates than NA items.

Reliability and separability. To detect potential differences in PA and NA measurement characteristics, item reliability, person separability, and correlations between summed scores and parameter estimates were examined.

Reliability of the set of items analyzed in a PCM was calculated in WINSTEPS in accordance with Equation 9 below

$$Reliability = \frac{s_{\beta}^2 - s_{error}^2}{s_{\beta}^2} , \tag{9}$$

where

$$s_{error}^2 = \frac{\sum_{i=1}^{N} SE_{\beta i}^2}{N} . \tag{10}$$

Reliability of a set of items was calculated as the variance of the item parameters minus error variance, divided by item parameter variance, where error variance was the sum of squared beta parameter standard errors divided by the number of items in the set. The reliability calculated in WINSTEPS indicated the reproducibility of the item hierarchy obtained through the PCM and was largely uninfluenced by survey length and model fit (Linacre, 2012).

Separation coefficients were calculated as the following function of reliability.

$$Separation = \sqrt{\frac{Reliability}{1 - Reliability}} \tag{11}$$

Squaring this separation coefficient resulted in a signal-to-noise ratio. This coefficient provided an indication of true score variance, the variance of theta estimates less error variance, in RMSE units (Linacre, 2012). One separability coefficient was calculated for the set of items in an analysis, and another for the sample of respondents. Here, item reliability and person separability were examined to avoid redundancy.

NDSHWB. Reliability across occasions was higher for PA than for NA, although both sets of items had high reliability coefficients across all occasions. PA items had a reliability coefficient of .99 on every measurement occasion. NA items had a median reliability coefficient of .90 across occasions (M = .90, SD = .02). The separation coefficient for participants was much higher for the PCMs of PA responses than for those of NA responses. The median PA separation coefficient across occasions was 4.23 (M = 2.24, SD = 0.15), equivalent to a signal-to-noise ratio of 17.89. The median separation coefficient for participants in the PCMs of NA responses was 1.26 (M = 1.26, SD = 0.13), indicative of a signal-to-noise ratio of 1.59. Overall, PA items had higher reliability and participants had much higher separability coefficients when responding to PA items than to NA items.

Maastricht. The PA items in the Maastricht data had a median reliability across occasions of .97 (M = .97, SD = .01). NA items had a similar median reliability at 0.94 (M = 0.94, SD = 0.01). The difference between PA and NA in person separability

coefficients was larger. Median person separation was 1.95 across occasions for PCMs of PA responses (M = 1.95, SD = 0.14). PCMs of NA responses resulted in much lower person separability coefficients across occasions, with a median of 0.49 (M = 0.49, SD = 0.13). Thus, the signal-to-noise ratio corresponding to the median separability coefficient for PA models was approximately 3.80 and 0.24 for PA and NA, respectively, indicating more noise than signal was present in NA responses. Although the difference between PA and NA in item reliability was trivial, the difference in person separation coefficients was substantial. The separation coefficient for NA models in the Maastricht data was particularly problematic, suggesting the data contained more noise than signal.

Summary of Longitudinal Measurement Characteristics. In both longitudinal data sets analyzed with anchored PCMs, substantial differences between PA and NA measurement characteristics emerged. First, response distributions for PA items were more similar to a normal distribution and tended to include observed responses in all categories. Responses to NA items were highly skewed, and only the lower half of the scale was used. Second, PA items targeted participants much better than NA items, resulting in narrower information curves and outfit statistics lower than acceptable bounds for NA items. Finally, PA items had lower SEs, and PCMs of PA responses produced higher person separation coefficients than NA items and models in both longitudinal data sets. In sum these longitudinal IRM results substantially aid in refuting the assumption that PA and NA measures have similar psychometric properties.

The following section will further counter this assumption by examining differences in the temporal stability of PA and NA measurement characteristics.

Temporal Stability of Measurement Characteristics: PA versus NA

To allow detection of further differences between measures, results from single-occasion PCMs were examined to determine the temporal stability of the psychometric properties of PA and NA measures. Recall that parameters from a PCM are estimated on a scale unique to that model. This property prevents the direct comparison of parameter estimates between two unanchored, single-occasion PCMs; however, the rank order of parameter values from two separate models can be contrasted. The consistency of the rank ordered item betas across occasions, coupled with the stability of category deltas over time, for PA was compared to that of NA in each longitudinal data set below.

Item location order. When the PCM fits the data well, item difficulty parameters are inversely correlated with item summed scores, the sum of all responses to each item. Ideally, the rank order of items by difficulty parameter remains constant across single-occasion PCMs. Although person scores may change from one occasion to the next, changes in individual response trajectories for a given item across occasions should approximately cancel each other out, resulting in fairly constant item summed scores. Thus, items should have similar item difficulty parameters, or at least similar rank order difficulties, across occasions.

The rank order of item difficulties for PA and NA items in each data set is depicted in Figure 16. In both data sets, there appeared to be fewer changes in rank for PA than NA items betas. PA item difficulties appeared to be more stable over measurement occasions than NA item difficulties. Given the highly skewed distribution of NA items on each occasion, with most NA items having beta parameters very close to

one another at the high extreme of the logit scale, it was not surprising that NA item betas changed rank order more frequently than PA item betas. Note the NA item in each data set with the lowest item location, Fatigued in the NDSHWB data and Tired in the Maastricht data, had the most stable item location parameters. Finding more NA items that target the lower end of the latent NA dimension may help to improve the temporal stability of NA measurement characteristic stability.

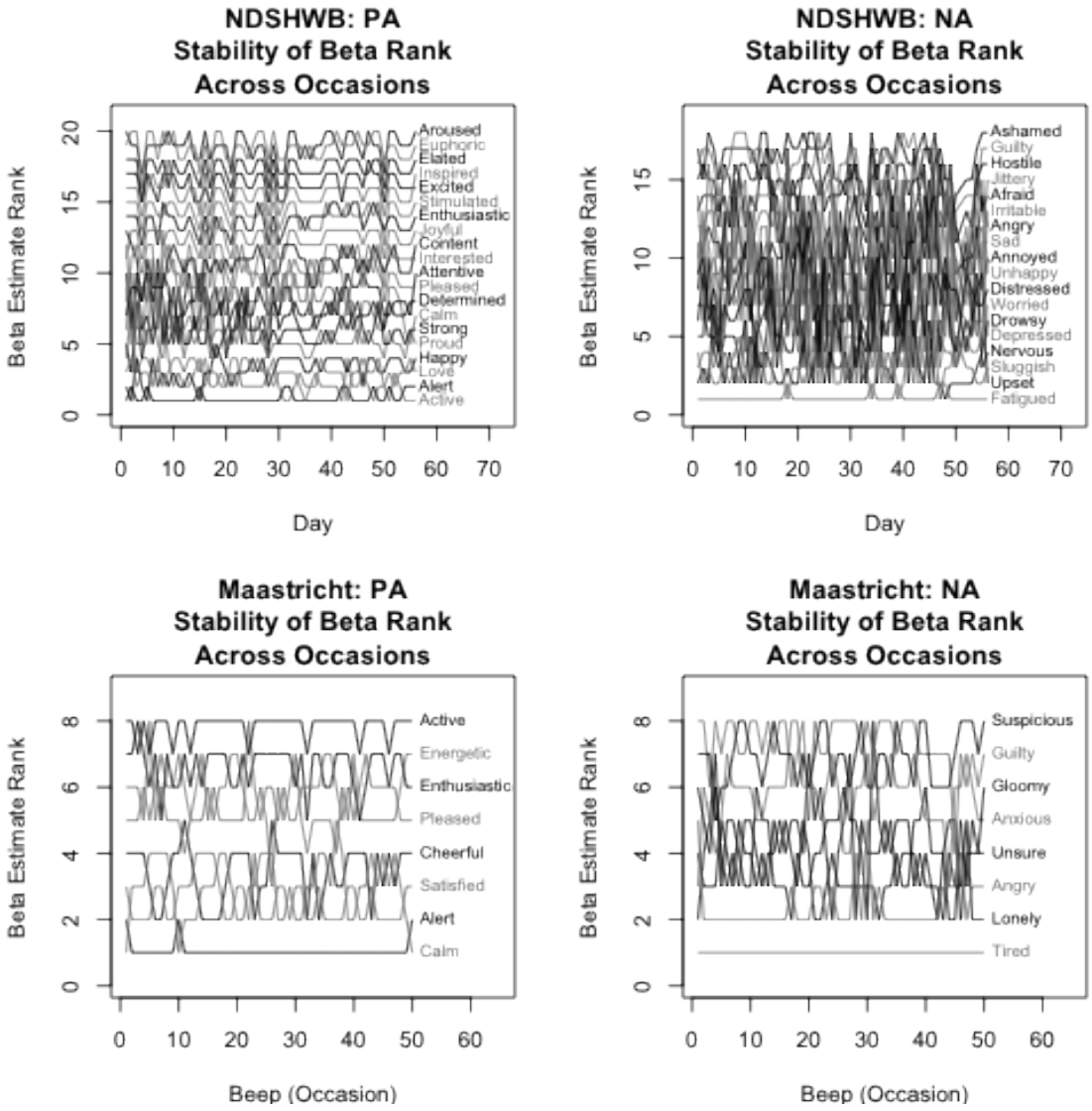

Figure 16. Item order by ranked beta estimates across occasions.

Delta order. A similar pattern was evident in sets of item deltas across occasions. Under the PCM, delta parameters, the thresholds between categories, should fall in appropriate order. The delta between categories 1 and 2 (denoted delta 1-2) should be lower than the delta between categories 2 and 3 (delta 2-3), and so on. When deltas are out of proper order, they are identified as disordered or as delta reversals. These reversals happen for a number of reasons, many of which are discussed by Andrich (2013). For example, disordered deltas could indicate a severe deviation from the Guttman-like structure loosely assumed by the PCM. Reversals could also indicate a poorly constructed response scale or some other substantive cause that must be investigated further. The presence of disordered thresholds with constant order across measurement occasions (e.g., deltas in the order delta 2-3, delta 1-2, delta 3-4, delta 4-5 observed on all occasions for a given item) indicates a problem with either the model or response scale that is stable over time. Disordered thresholds that have inconsistent order across occasions are more problematic, suggesting model misfit or response scale problems that are not stable over time. For each item, threshold order was examined across occasions.

In most applications of IRT, if reversals are identified, further investigation must be carried out to determine the source of the reversal, if the source can be detected. In the present investigation, the aim was to identify differences in measurement characteristic temporal stability between PA and NA measures. Thus, the source of reversals was not of as much interest as the detection of the reversals. Differences in the frequency and temporal stability of delta reversals between PA and NA measures would further expose

the fallacy of the assumption that PA and NA measures have similar desirable psychometric properties.

NDSHWB. Two PA and two NA items were selected from each data set and the rank order of delta estimates for each item was plotted across occasions in Figures 17 and 18 for the NDSHWB data and the Maastricht data, respectively.

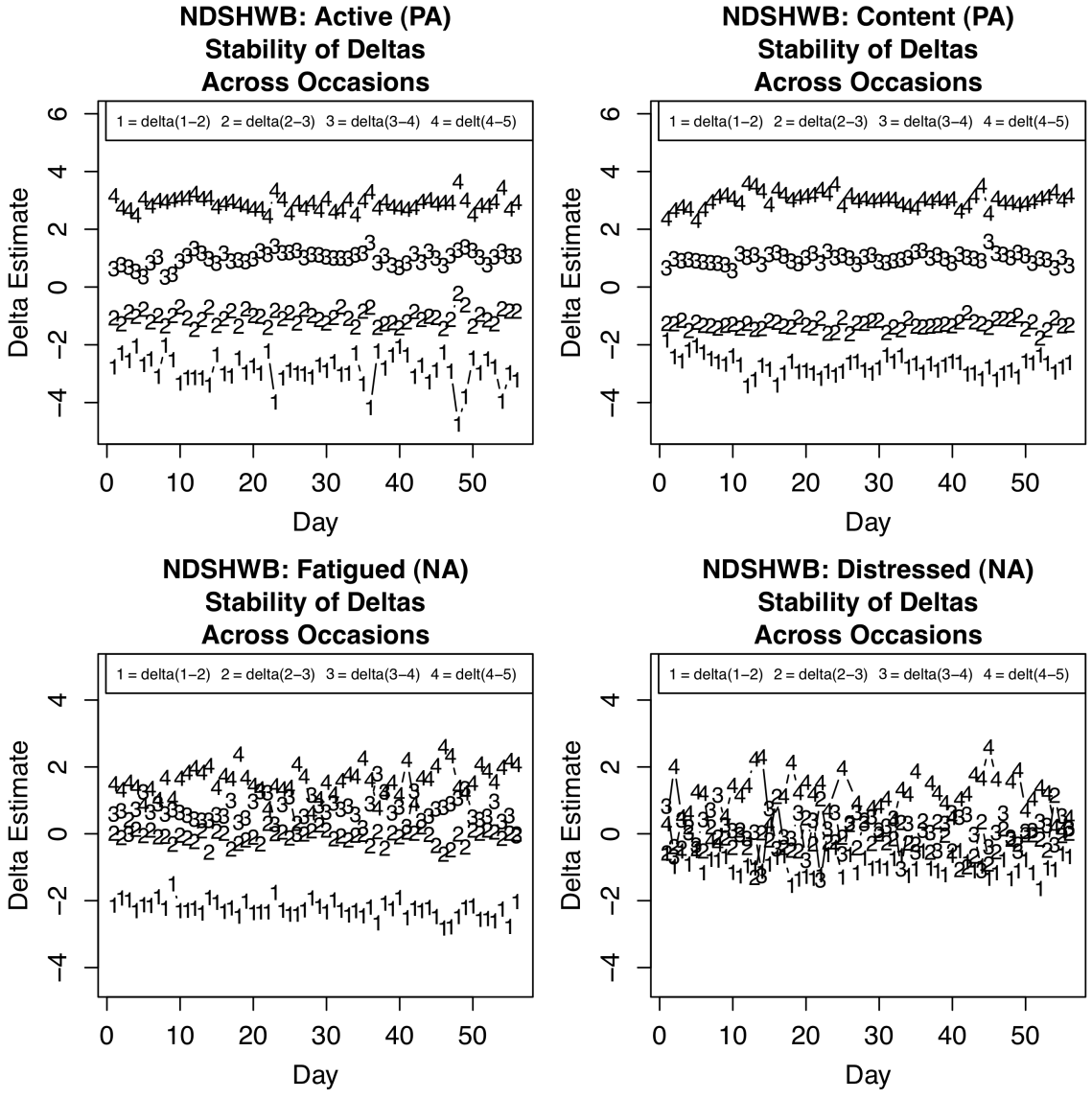

Figure 17. Rank order of delta estimates across occasions.

NDSHWB items were administered with a 5-point response scale, resulting in four estimated delta parameters for each item on each occasion. For the selected PA items, Active and Content, delta parameters remained in proper order across all measurement occasions: not a single reversal was present. These two PA items were representative of the vast majority of PA items administered in the NDSHWB data, as were the NA items selected here of the rest of the NA items in the data. The two NA items pictured, Fatigued and Distressed, had much less stable delta parameters, with reversals appearing to occur between varying categories across measurement occasions. These reversals did not appear to be consistent across occasions.

Maastricht. The deltas estimated from the Maastricht data, particularly those of PA items, exhibited a surprisingly different pattern than NDSHWB deltas. Again, selected items were representative of the rest of the items administered. With the 7-point scale given for items in the Maastricht data, the six deltas estimated for each item were not properly ordered in either of the item sets. Reversals were evident on most, if not all, occasions for both NA and PA items. It might have been the case that administering seven categories was expecting more precision from participants than they typically use when evaluating affect. Alternatively, specific categories might have been problematic. For example, the first delta was never the lowest for Cheerful and Active in Figure 19. Perhaps participants were unwilling to report low levels of PA adjectives unless they felt they endorsed enough other PA items highly, causing the first delta (delta 1-2) to be more difficult to pass than deltas between higher categories. The reversals for NA deltas were not nearly as systematic, indicating temporally inconsistent response scale problems.

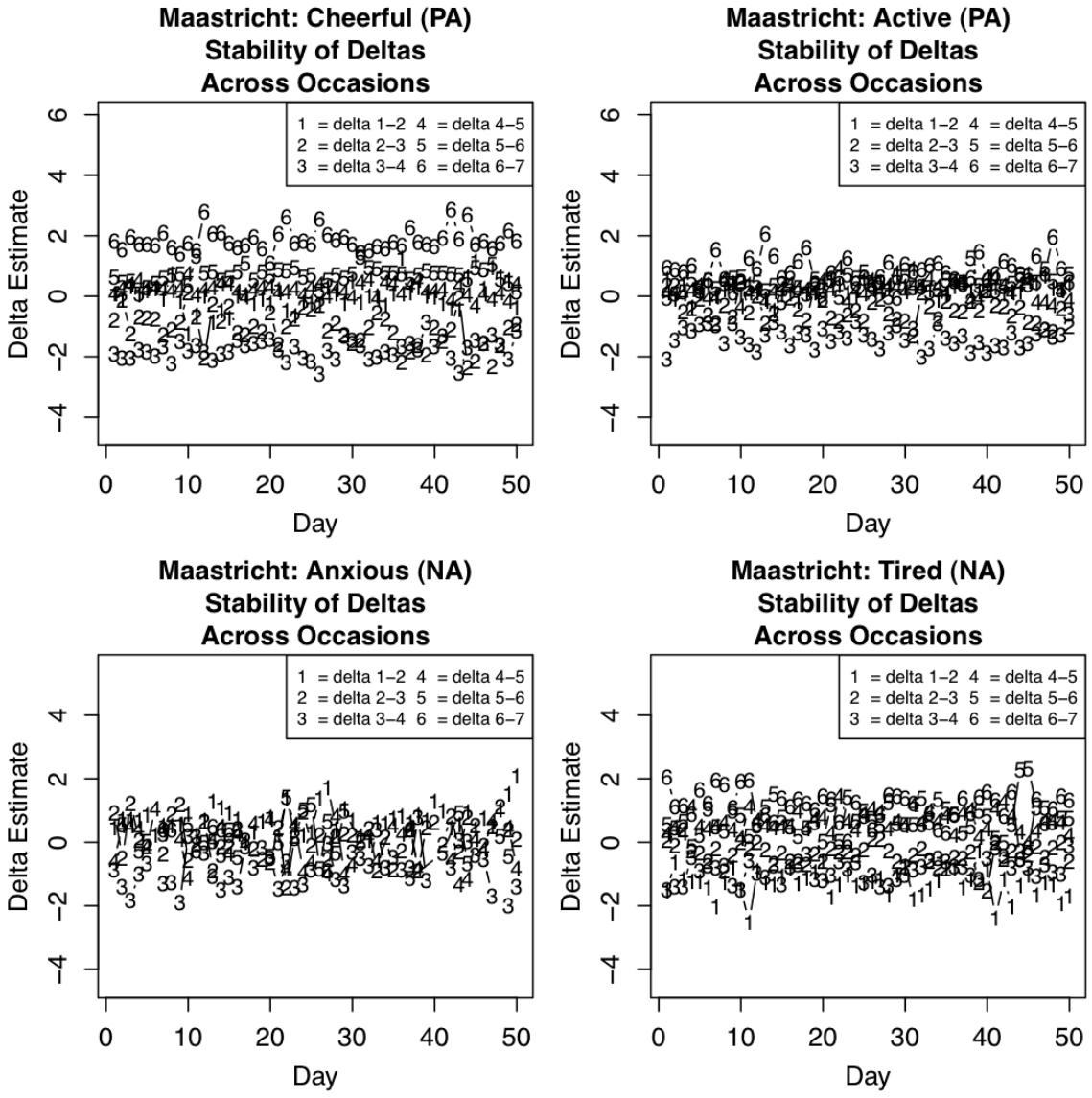

Figure 18. Rank order of delta estimates across occasions.

Summary of temporal stability results. Item betas and deltas were much less stable over time for NA items than for PA items in the NDSHWB data. In the Maastricht data, items measuring each type of affect exhibited disordered deltas, but disordering was less consistent over time for NA deltas. Both sets of results aided in refuting the assumption that PA and NA measures have similar, desirable psychometric properties. In

the NDSHWB data, the PA measure administered had much more temporally stable item parameters than the NA measure. In the Maastricht data, both the PA and NA measures exhibited undesirable delta reversals, which indicated a need for improvement; however, these reversals were more consistent across occasions for PA items than for NA items.

Collapsed Response Scales: PA versus NA

To challenge the assumption that PA and NA are captured adequately by the same response scale, all possible collapsed response scales were examined for PA and NA items separately, in each dataset. The 5-point response scale employed in the NDSHWB could be collapsed a total of 15 unique ways, including the original, and the 7-point scale in the Maastricht project yielded a total of 63 possible response scales (see Appendix A). To identify the best-performing collapsed response scale for each type of affect in each data set, a variety of precision statistics (e.g., standard errors), fit statistics, and other item and person performance criteria were evaluated.

Initial response scale elimination. After investigating preliminary results, several collapsed response scales were eliminated from further evaluation, due to one of two major problems. First, a very small number of the collapsed response scales resulted in a large number of inestimable item parameters. Perfect low (e.g., all responses to a given item are 1, meaning *Not at all*) and high item summed scores can cause problems for item parameter estimation. A small handful of the collapsed response scales resulted in the vast majority of responses to a given item being recoded as the highest or lowest category, particularly for NA items. For example, as illustrated earlier in Figure 11, most individuals in the NDSHWB used only the lower half of the original 5-point response

when rating levels of NA. Thus, under the collapsed response scale denoted 11112, in which the lowest four categories were collapsed into a single new category, all responses to a NA item may be recoded as 1, hindering item parameter estimation.

The Rasch modeling program WINSTEPS (Linacre, 2012) identifies items with inestimable parameters. Exploration of inestimable parameters suggested this problem only existed for a few collapsed response scales when applied to NA items in each data set (see Figure 19) and was absent for PA items. Specifically, response scales that collapsed the lower half or more of the response scale into a single category produced the most inestimable parameters. Based on Figure 19, the empirical cut-off of 5% was used to eliminate collapsed patterns with more than a handful of inestimable theta parameters across all participants and occasions. Three of the 15 response scales were removed from further consideration for NA items in the NDSHWB data (11112, 11122, and 11123) and seven of the 63 response scales were removed for NA items in the Maastricht data (1111112, 1111122, 1111123, 1111222, 1111223, 1111233, and 1111234).

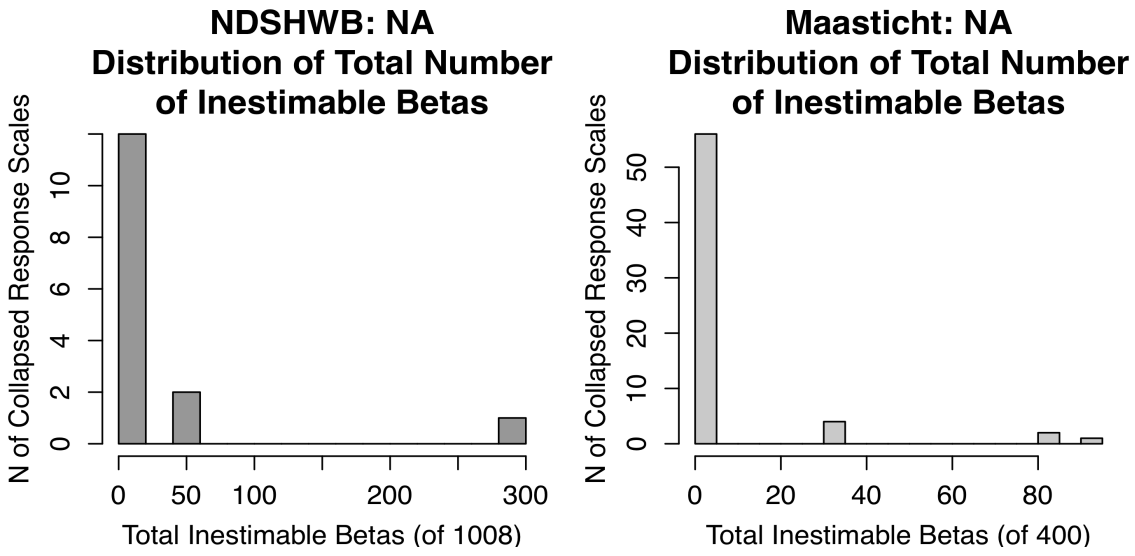

Figure 19. Distribution of inestimable thetas for collapsed response scales for NA items.

The second problem created by several collapsed response scales was disordered deltas. While the underlying hypothesis behind the PCM is that deltas monotonically increase, assuming items are not reversed scored and higher response categories indicate more of the latent dimension measured, no constraints are made in model implementation to prevent estimated deltas from violating this hypothesized order (for a more detailed examination of delta values and structure in Rasch-based models see Andrich, 2013). Disordered delta estimates can indicate a variety of psychometric and substantive phenomena, such as poor choice of response scale, misuse of response scale, model misfit, or incorrect assumptions about the nature of the assessed construct. Given the goal of the present investigation, to identify best-performing collapsed response scales for each type of affect, eliminating response scales that produced unreasonable frequencies of disordered deltas across items and occasions was more important than determining the source of the disorder. Thus, investigating the cause of disordered delta parameters is left for future work.

Longitudinal applications of the PCM are rare, and the debate on the appropriateness of collapsing response scales continues to thrive in the IRM literature (e.g., see Adams, Wu, & Wilson, 2012; Andrich, 2013). Thus, guidelines for deciding how many disordered deltas are (un)reasonable in longitudinal applications of the PCM are difficult to find, if any exist. An empirically derived cut-off was used in the present investigation. Results from collapsed response scale analyses, including fit statistics, measures of error and precision, reliability, and separability statistics, indicated the original 5-point scale performed well for PA items administered to the NDSHWB

sample. This original 5-point scale produced disordered thresholds 1.3 percent of the time, across all measurement occasions and items, or 15 instances out of 1120 total possible instances (20 items x 56 occasions = 1120 item-occasion pairs). The original response scale for NA items in the NDSHWB data and for PA and NA items in the Maastricht data produced unacceptably high proportions of disordered deltas (19.7%, 99.8%, and 90.8%, respectively). The more reasonable proportion of disordered deltas produced by the original 5-point scale of NDSHWB PA items was rounded up from 1.3% to 2%. Given the exploratory nature of this project, collapsed response scales with less than 5% of disordered deltas were examined; however, only response scales with less than 2% were considered for recoding the data for use in the exploratory factor analyses conducted in Chapter 7.

Histograms depicting the distribution of disordered deltas frequencies for all collapsed response scales for PA and NA items from both data sets are shown in Figure 20. For the NDSHWB data, none of the 15 collapsed response scales were removed from further consideration for PA items, and six of the remaining 12 scales were eliminated for NA items. Similarly, 40 of 63 collapsed response scales were eliminated for PA items in the Maastricht data, and 49 of the remaining 56 scales were removed for NA items.

After removing collapsed response scales with substantial numbers of inestimable parameters and/or disordered deltas, all 15 possible scales remained in consideration for PA items in the NDSHWB data, whereas only six remained for NA items. In the Maastricht data, 23 of the original 63 scales remained in consideration for PA items, and seven of the 63 scales remained for NA items. In the evaluation below, the original 5- or

7-point scales used in administration in the present projects were additionally included on any visualization of scale performance.

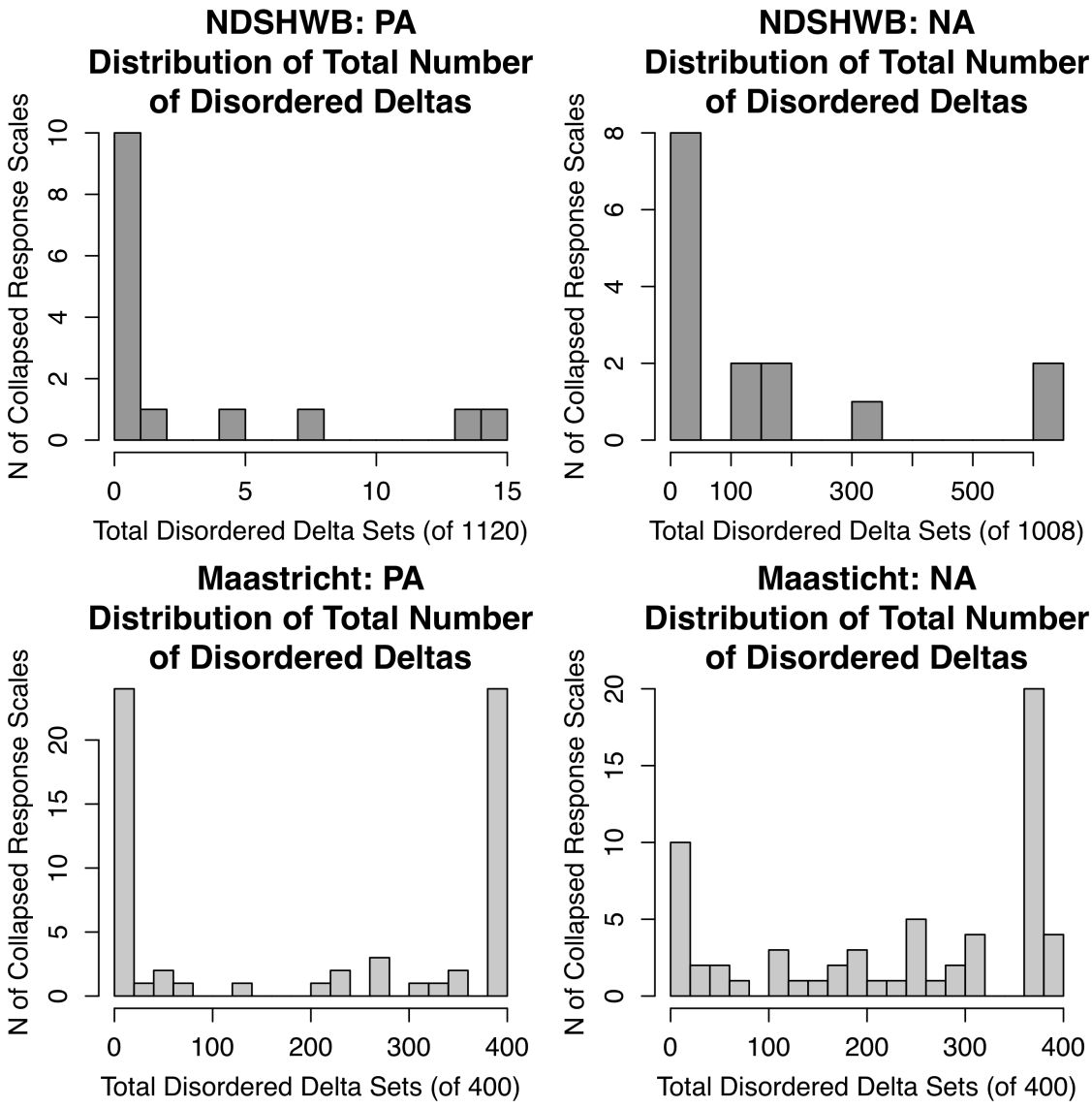

Figure 20. Distribution of disordered deltas for collapsed response scales.

PCM statistics. Several statistics, including person and item mean square fit statistics, separability, and reliability, have been agreed upon by many Rasch modelers (e.g., Green & Frantom, 2002; Lamoureux et al., 2008; Nilsson et al., 2005) to be results

of interest and were thus explored in this examination of alternative response scales. Additional results, the standard deviation of infit mean square and the ratio of separability to this standard deviation (*SD*), from a more extensive list of statistics specific to response scale evaluation (Stone, 1998) were also explored. Below, an overview of response scale performance is given and best-performing response scales are identified. A more detailed look at performance on each PCM statistic follows, beginning with fit mean squares, followed by separability, separability-infit *SD* ratio, and reliability, ending with correlations among summed scores and estimated parameters. In keeping with the structure established above, NDSHWB results are examined first and Maastricht results are examined second within each section.

Overview of ideal response scales. The best 3 performing response scales in each data set are listed in order from most to least ideal in Table 1 for each statistic examined under PA and NA separately. Best-performing response scales for each statistic were chosen visually based on three criteria: 1) Proximity of median to ideal value, 2) proximity of mean to ideal value, and 3) Smaller standard deviation. Two patterns were particularly noticeable in Table 1. First, for most of the eights statistics examined, the best-performing response scales for PA responses in both data sets had four or five categories, and the best-performing response scales for NA responses in both data sets had two or three categories. Second, the best-performing response scales for NA responses in both data sets, specifically the most frequent 2-category scales observed in each data set (12222 for NDSHWB and 1222222 for Maastricht), followed a very similar pattern. Both response scales collapsed all but the lowest category together, resulting in a

binary response scale. The performance of these two response scales, along with the best-performing response scales for PA responses, on the statistics listed in Table 1 are examined in more detail below.

Table 1

Top-Performing Response Scales by Data Set, Affect Type, and Statistic

	ND				**MAAS**			
	PA		**NA**		**PA**		**NA**	
Statistic	**Person**	**Item**	**Person**	**Item**	**Person**	**Item**	**Person**	**Item**
Infit MNSQ	11112 11122 11123	11112 11122 11222	12222 11222 12223	11222 12222 12223	1222222 1112222 1122222	1111222 1111122 1112222	1222222 1112222 1122222	***
Infit MNSQ *SD*	12223 11122 11112	11112 11122 11222	12222 12223 12233	11222 12222 12223	1222222 1222223 1122222	1111222 1222222 1112222	1222222 1122222 1112222	1112222 1122222 1122223
Outfit MNSQ	12345 12334 11234	11112 11123 11234	12233 12222 12223	11222 12222 12223	1112233 1222222 1222333	1222222 1122222 1112222	1222333 1222233 1222223	1222233 1222333 1122223
Reliability	*12345* *11234* *12234*	***	12222 12223 12233	12222 12223 12233	*1122345* *1222345* *1122334*	***	1222222 1222223 1222233	1222222 1222333 1122222
Separation	12345 12234 12334	12344 12345 12233	12222 12223 12233	12222 12223 12233	*1222345* *1122345* *1222334*	1122345 1222345 1112234	1222222 1222223 1222233	1222222 1222333 1122222
Separation Infit *SD* Ratio	12233 12223 12334	11122 11112 11222	12222 12223 12233	12222 11222 12223	1122223 1222223 1112223	1111222 1111122 1112222	1222222 1122222 1222333	1222222 1222333 1122222
Item-Total Correlation	12345 11234 12234	N/A	12223 12222 12233	N/A	1222344 1112233 1222334	N/A	1222222 1222223 1222233	N/A
Sum-Beta (or Theta) Correlation	***	11234 11123 12345	12222 12233 11222	12222 12233 12223	***	1111223 1112234 1122345	1222222 1122222 1222333	1222222 1222223 1122222

Note. Italics indicate the three response scales listed performed equally well. ***Indicates most (>75% of) response scales performed equally well (no highlights on corresponding plots).

Fit statistics. Mean square fit statistics were examined for each response scale. In addition to infit and outfit mean square statistics, the standard deviation of infit mean

square values was included in response scale evaluation as recommended by Stone (1998).

NDSHWB. In Figure 21, the means and medians of these statistics across all PCMs, along with standard error bars and standard deviation markers, are plotted by collapsed response scale for PA and NA responses in the NDSHWB. In all collapsed response scale plots, the three best-performing scales from Table 1 are highlighted in gray. Infit mean square values and infit standard deviation favored response scales with fewer categories for both PA and NA. Outfit mean squares favored response scales with more categories for PA responses, and a mix of more and fewer categories for NA.

Note the administered 5-point response scale for NA items had excellent outfit mean square statistics, despite the very poor performance it exhibited on other statistics and delta order. This result might have been an artifact of the substantial mismatch between NA item locations and participants on the estimated logit scale. Recall similar outfit mean square performance was observed with the administered 5-point scale when evaluating longitudinal PCM results, despite the disordered deltas and poor sample targeting associated with the original response scale.

Thus, when considering mean square statistics as performance measures, it is important to remember what these statistics indicate. Infit and outfit mean squares are both chi-square based statistics, indicating the extent to which deviations of observed responses from expected values are larger or smaller than expected, given model parameters. An item that is much higher on the measured latent dimension than any respondents will have a very low expected value for each respondent. If participants

endorse low categories for that item, the outfit mean square statistic for that item will be near the ideal value of one. It will not reveal whether the item is much too high on the latent dimension for the entire sample, nor give any indication of whether item or person parameter estimate SEs are low or high, and it will not expose delta reversals. Mean square fit statistics only expose how well the model fits the data for each item and person in terms of deviations from expected scores and must be evaluated with this in mind.

Maastricht. Mean square fit statistics for PA and NA response scales in the Maastricht data are plotted in Figures 22 and 23, respectively. Item and person infit and infit standard deviation favored response scales with fewer categories for both PA and NA. Outfit mean square values, particularly person outfit, favored response scales with more categories for both PA and NA responses.

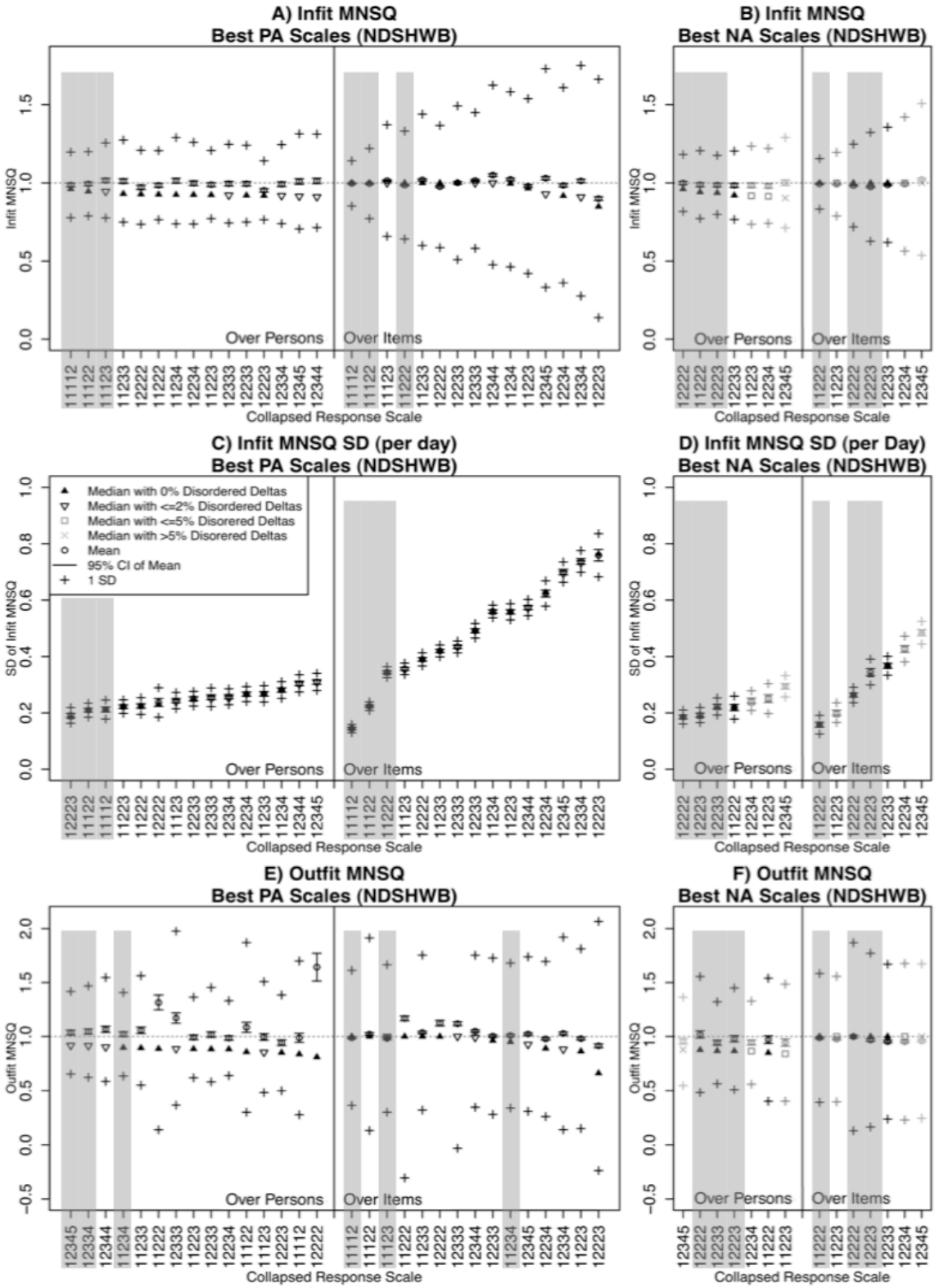

Figure 21. Mean square fit of items and persons by collapsed response scale: NDSHWB.

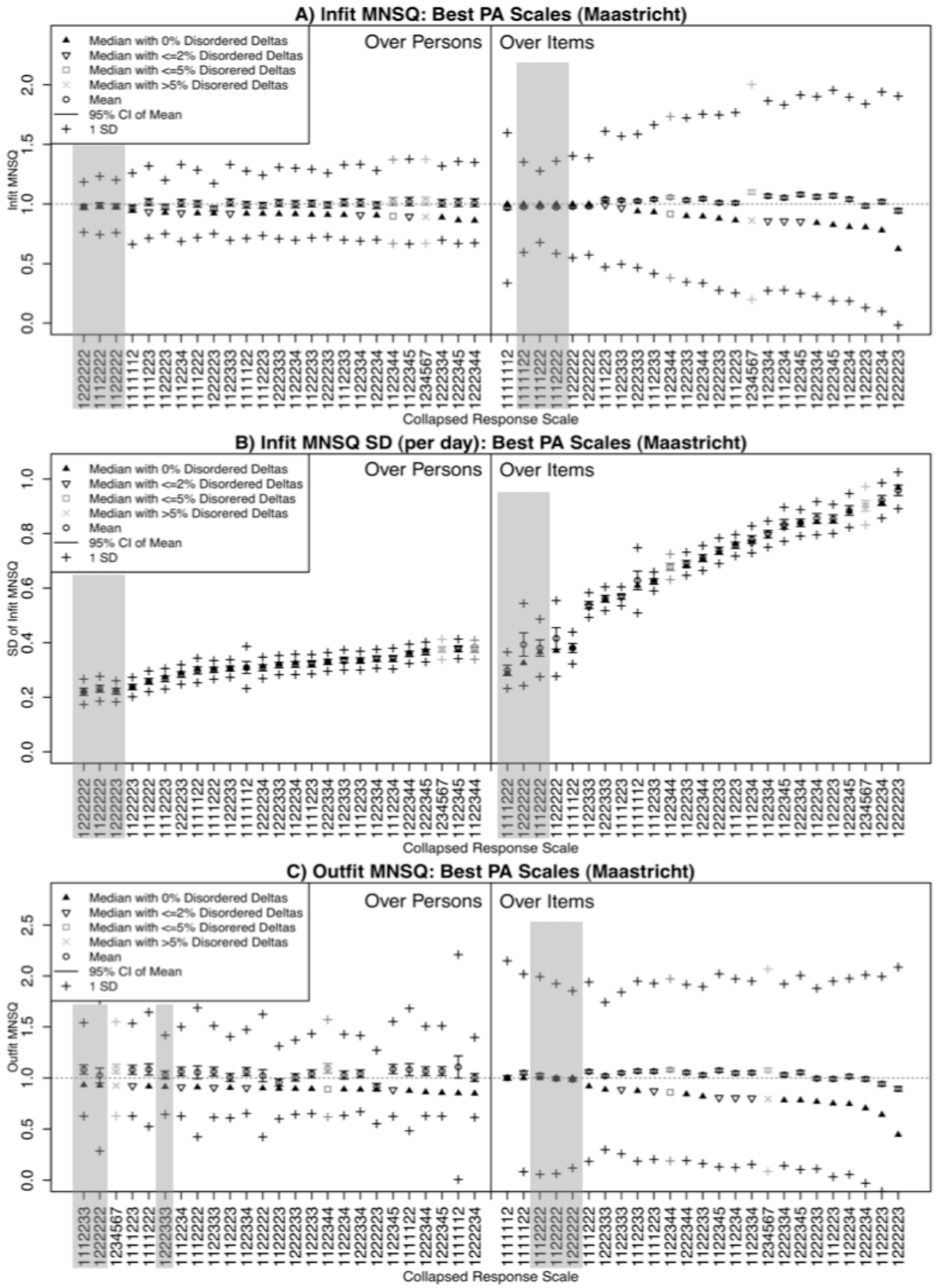

Figure 22. PA collapsed response scale performance on mean square fit: Maastricht data.

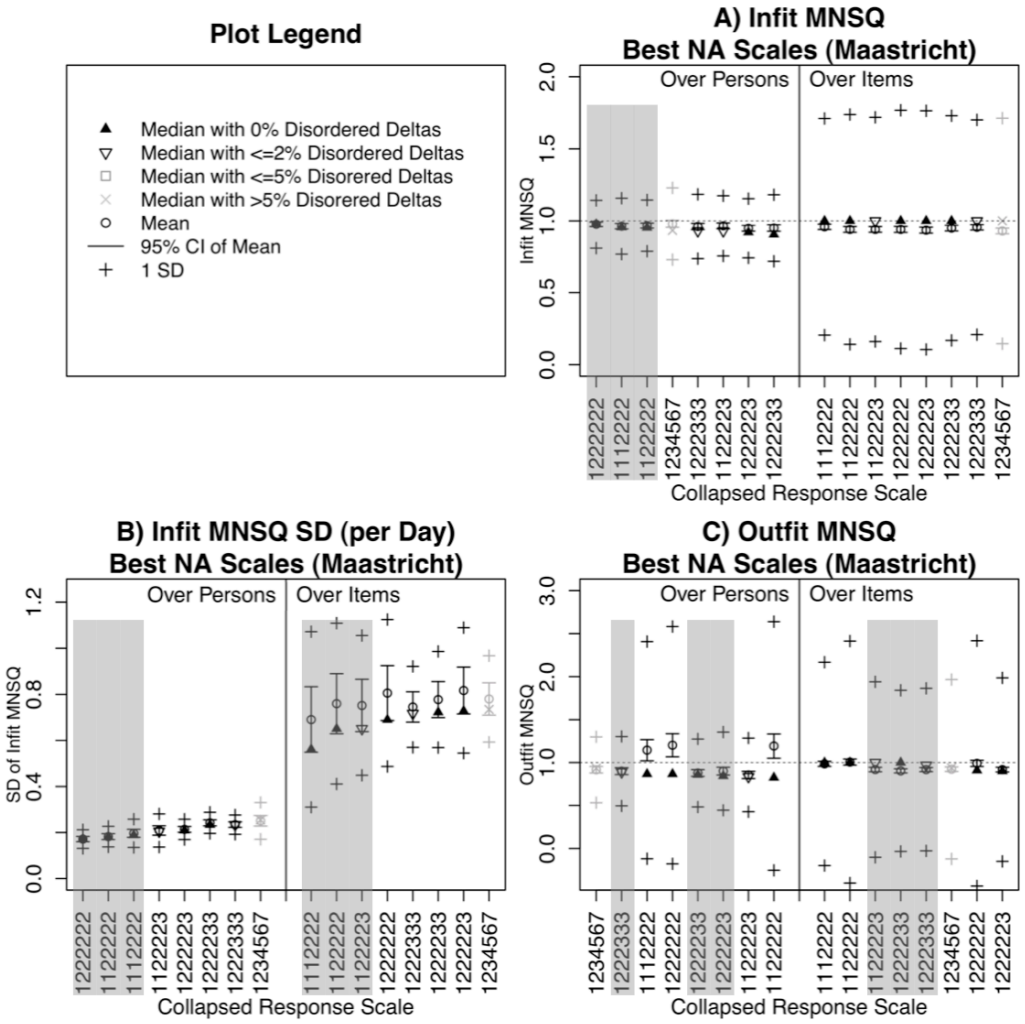

Figure 23. NA collapsed response scale performance on mean square fit: Maastricht data.

Reliability, separability, and separability-infit ratio. Response scales were also evaluated on item and person reliability, separability, and the ratio of separability to infit mean squares *SD*, as suggested by Stone (1998).

NDSHWB. These statistics are plotted by collapsed response scale for both PA and NA persons and items in the NDSHWB in Figure 24, with best-performing scales highlighted in gray. Response scales with three to five categories performed best for PA person and most item statistics. For NA responses, the response scale in which the top

four categories of the original 5-point scale were collapsed together performed best for all person and item statistics. According to reliability and separability statistics, the ideal collapsed response scale for NA responses was a binary scale, while the ideal collapsed response scale for PA had more than two response categories, likely four or five.

Maastricht. The pattern of results obtained from the Maastricht data mimicked those obtained from the NDSHWB data (see Figure 25 for PA and Figure 26 for NA results). Note that item reliability estimates for PA responses in both data sets did not differentiate between response scales well. Thus, no scales were highlighted on these plots. Reliability and separability statistics for both persons and items favored 4- and 5-category collapsed response scales for PA responses, and a 2-point collapsed response scale for NA responses. In addition, the specific 2-point scale favored for NA responses was similar in structure to the 2-point scale with the best performance for NA responses in the NDSHWB data. In both data sets, the ideal response scale in terms of reliability and separability statistics collapsed all categories but the lowest one together, resulting in a binary collapsed response scale.

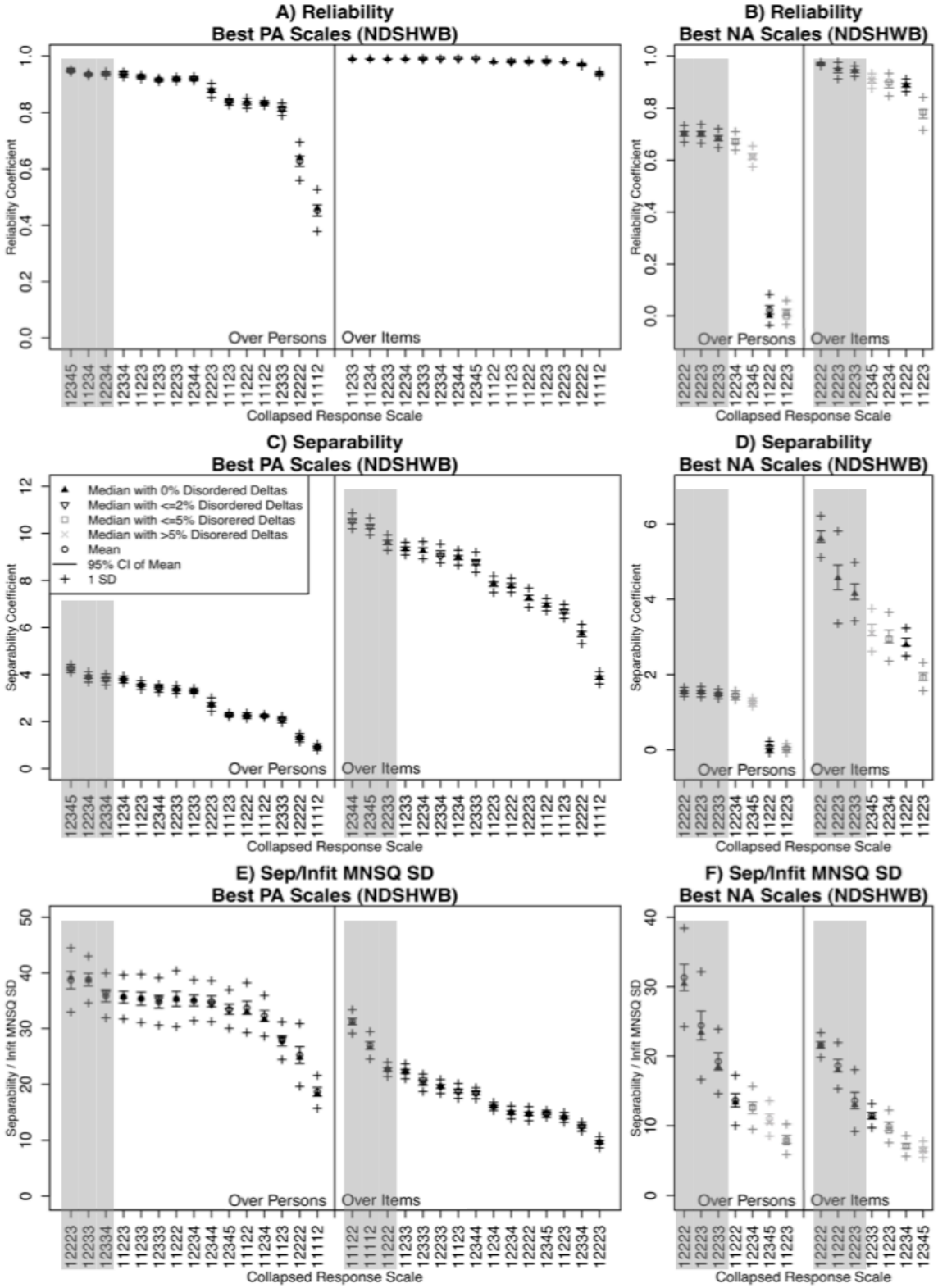

Figure 24. Reliability and separability by collapsed response scale: NDSHWB.

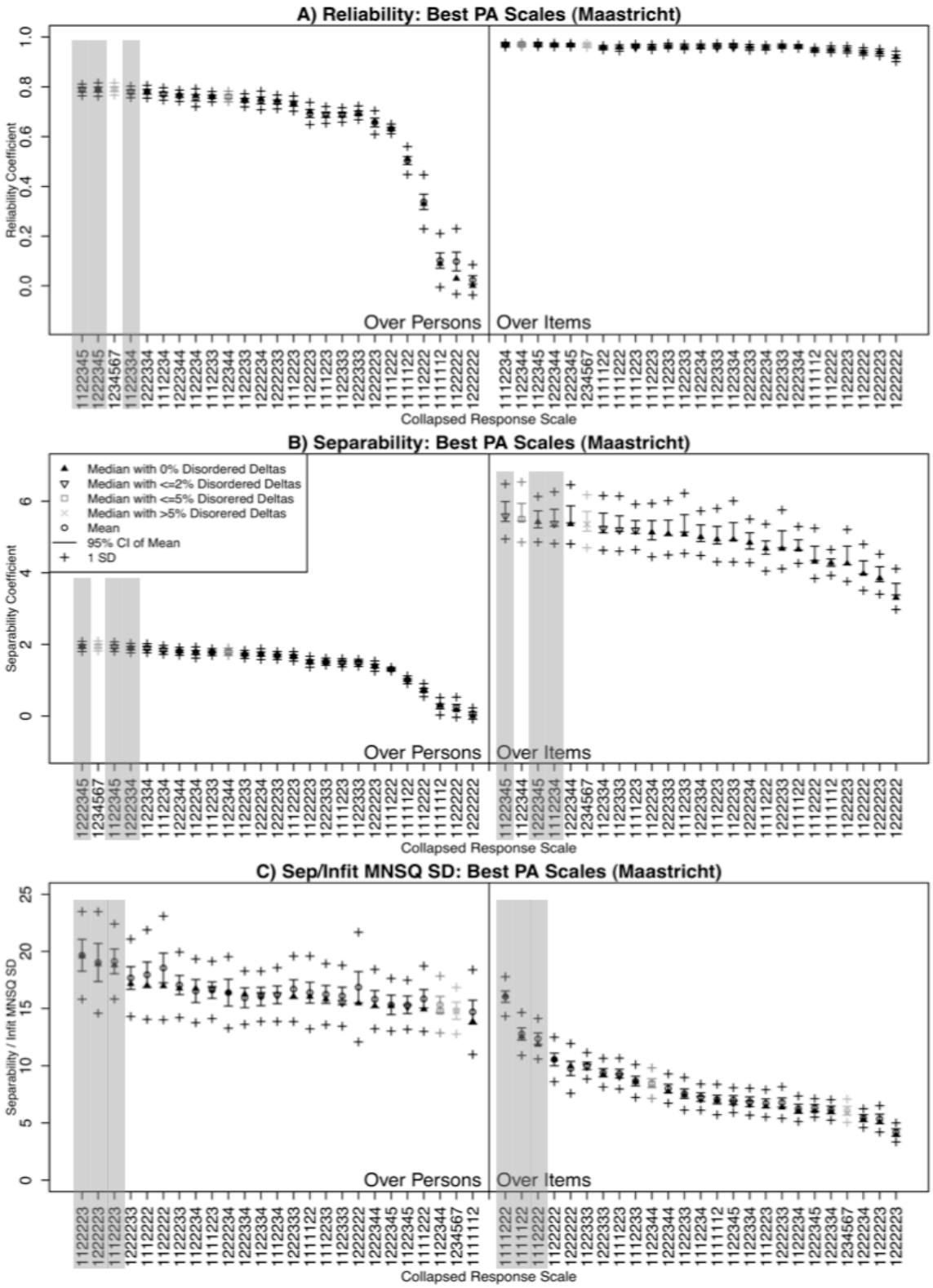

Figure 25. Reliability and separability by collapsed response scale: Maastricht PA data.

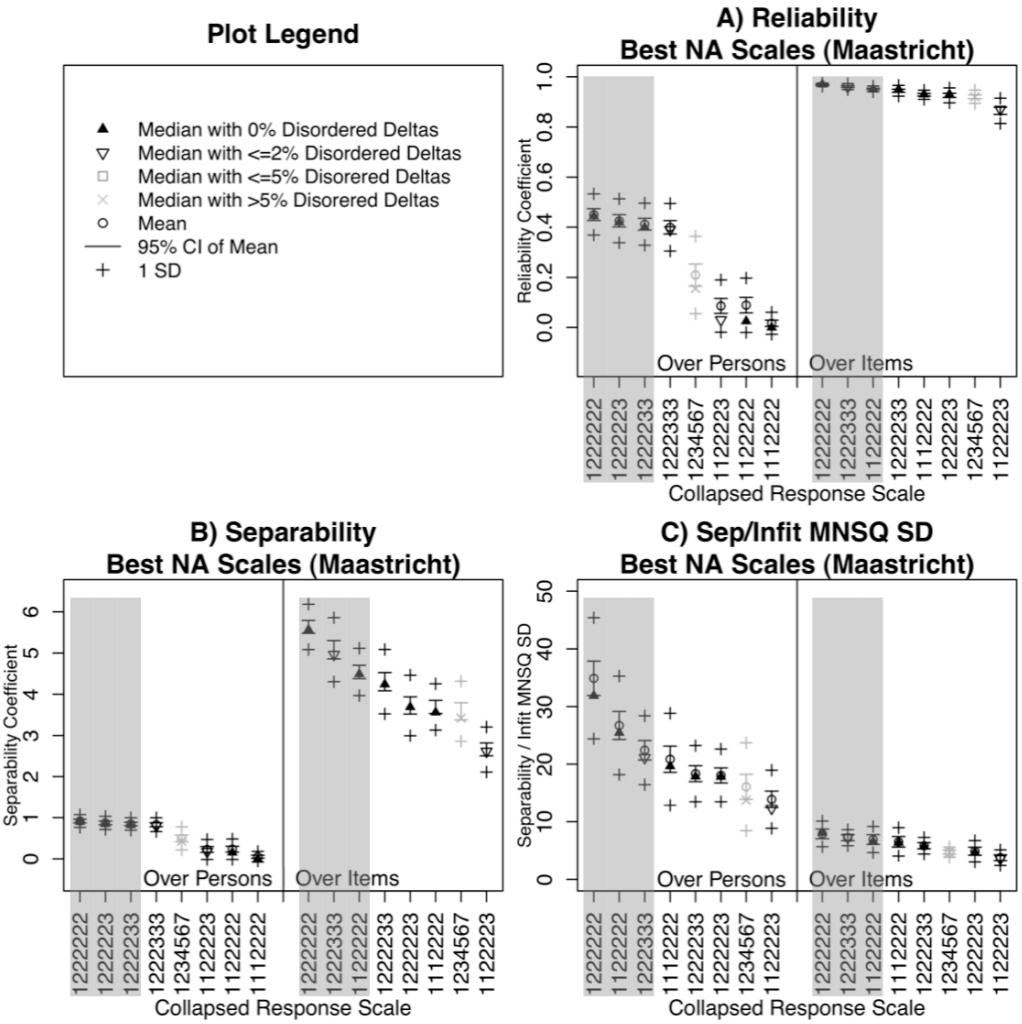

Figure 26. Reliability and separability by collapsed response scale: Maastricht NA data.

Item-total and parameter-summed score correlations. Three sets of correlations were examined to conclude the evaluation of collapsed response scales. First, the correlation between item (or person) responses and total scores was examined for each person and item, with high positive values being desirable. Then, the correlation between item summed scores and estimated item locations, or betas, was compared, along with the correlation between person summed scores and estimated thetas.

Summed scores are a sufficient statistic in Rasch-based models, such as the PCM. Thus, participants with the same summed scores have the same theta estimates. If an item is very high on the measured latent dimension and has a beta location parameter far above most of the theta parameters for the respondents, most individuals will endorse low categories on the item, resulting in a low summed score. Thus, the ideal value for the correlation between item summed scores and betas is -1. Items with very high betas should have low summed scores, and items with low betas should have high summed scores. The ideal value for the correlation between person summed scores and theta estimates is +1. Individuals with higher theta parameters should have higher levels of the construct measured, and have higher summed scores as a result.

NDSHWB. Summaries of these three correlation coefficients for each collapsed response scale in the NDSHWB data are plotted in Figure 27, with best-performing response scales highlighted in gray. Note that the correlation between item summed scores and beta estimates did not differentiate well between PA response scales, and thus no PA scales are highlighted. For PA response scales, all three correlations favored response scales with more categories, including the original 5-point scale administered, although the correlations involving item summed scores had near ideal values for almost all, if not all, collapsed response scales. Correlation coefficients for NA response scales exhibited larger differences between collapsed response scales. The binary scale that performed best on measures of reliability and separability also performed best on these three correlation coefficients.

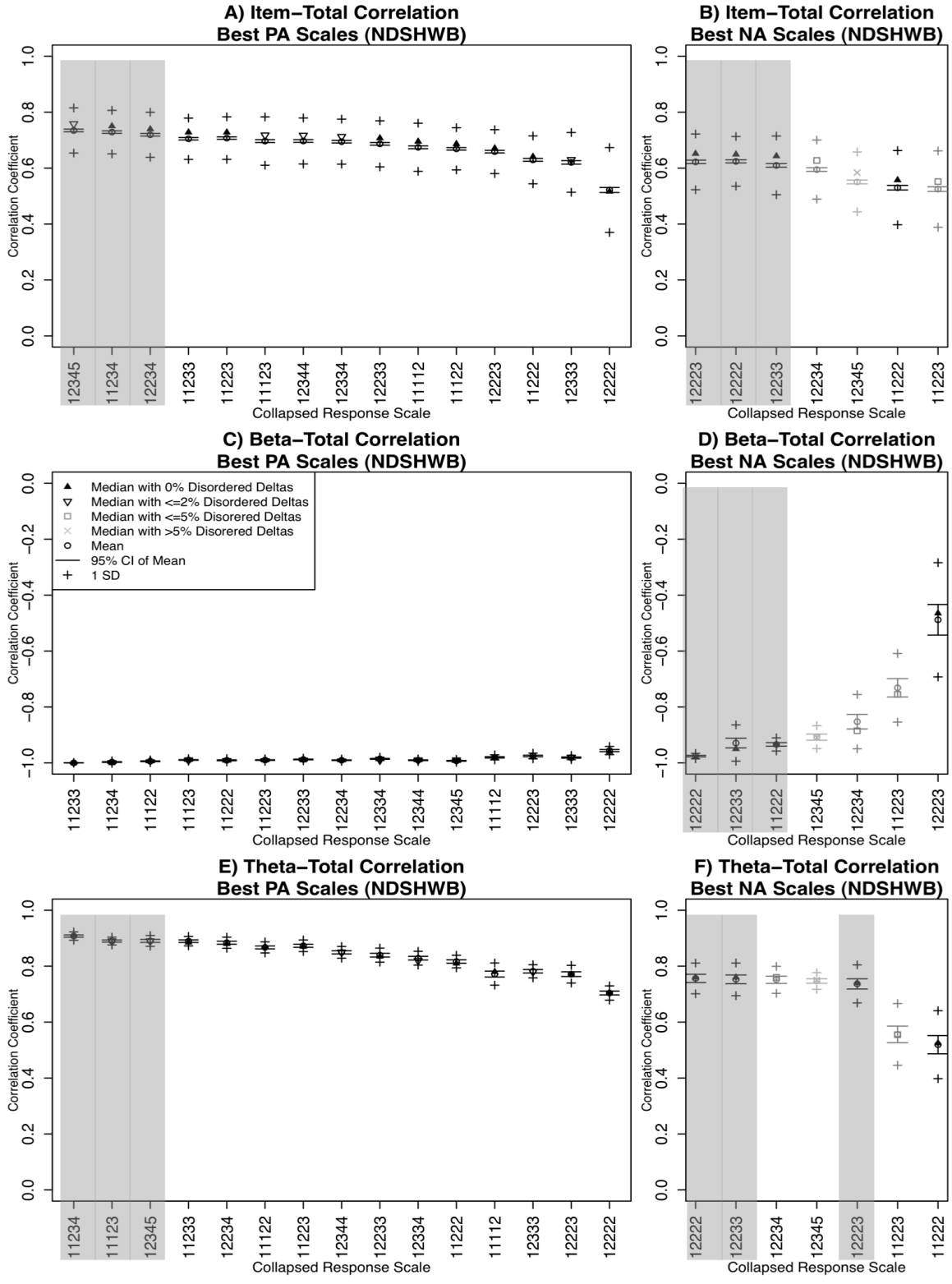

Figure 27. Summed score-parameter correlation by collapsed response scale: NDSHWB.

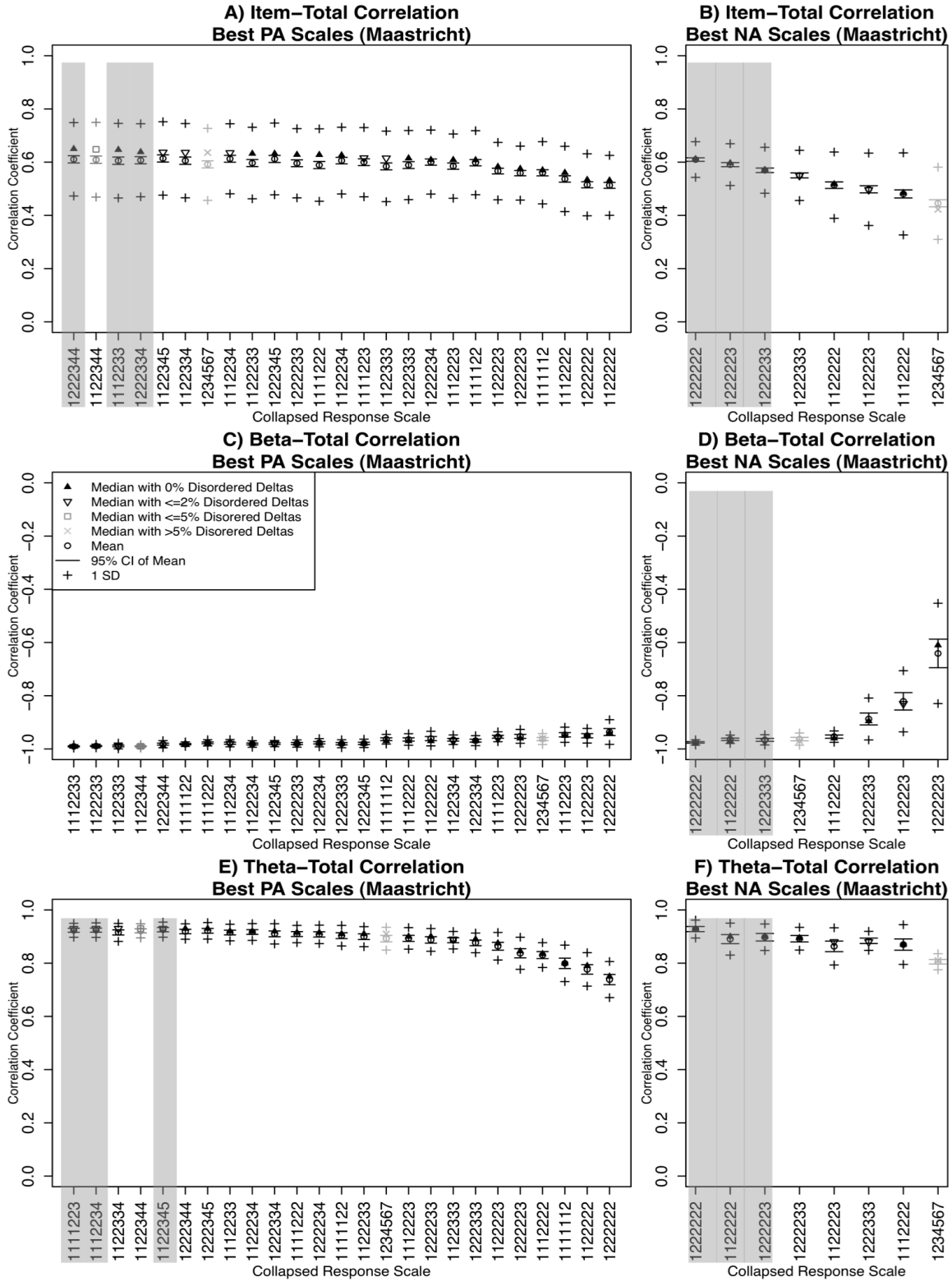

Figure 28. Summed score-parameter correlation by collapsed response scale: Maastricht.

Maastricht. The same three correlations for the Maastricht data are potted in Figure 28. Most PA response scales performed close to ideal with the exception of many of the binary response scales, which exhibited poorer performance. All three correlation coefficients favored the same binary scale for NA items, also favored by the reliability and separability statistics reviewed earlier. The collapsed response scales closest to ideal for PA responses were 4- and 5-point scales, while the response scale closest to ideal for NA responses was binary. This pattern was observed in both of the longitudinal data sets examined here.

Summary. To identify an optimal response scale for each type of affect in each data set from the scales examined here, two summed rank scores were calculated for each of the response scales in Table 1. The first score weighted response scales by performance rank. Of the three response scales with the best performance on a given statistic, the third best was given one point, the second best was given two points, and the best was given three points. The second summed rank score was unweighted, indicating how many times a response scale appeared in the table, and thus would assign the top three response scales for a given statistic one point each.

For the NDSHWB, the original 5-point scale had the highest scores on both of these ranks for PA responses, and the collapsed response scale 12222 had the highest scores on both ranks for NA. Thus, before the NDSHWB data was used to examine individual differences in affect factor structure and factor correlations, NA responses were recoded using the 12222 response scale. PA responses were not recoded, as the original response scale performed best.

In the Maastricht data, one of the 5-point collapsed response scales had the second highest rank summed scores. A binary response scale had the highest rank summed scores; however, this was mainly driven by performance on infit-related statistics and thus was not chosen as the optimal response scale. All other statistics examined favored 3-, 4-, and 5-category response scales, further supporting the choice of a polytomous scale over a binary scale. PA responses in the Maastricht data were recoded according to the response scale 1122345, before being used in individual differences analyses. For NA responses, the collapsed response scale 1222222 had the highest rank summed scores, and the data was recoded accordingly.

Conclusions From IRMs

Challenging Measurement Assumption #1

The first measurement assumption found in the affect literature challenged here is that PA and NA have similar desirable measurement characteristics. PA and NA responses are often analyzed with the same models, usually confirmatory or exploratory factor analysis e.g., Gaudreau et al., 2006; Leue & Beauducel, 2011), and differences between PA and NA, such as factor loading magnitudes, correlated errors, and large residual variances, are rarely acknowledged or discussed. Failure to acknowledge and examine these differences and continued use of the same models for PA and NA responses suggests researchers are assuming PA and NA have similar enough and desirable enough psychometric properties to warrant analyzing both types of affect with the same models, without searching for differences between the types of affect in obtained results.

Longitudinal and cross-sectional PCMs were conducted on affect responses from two longitudinal data sets to challenge this assumption. Anchored PCMs, with item parameters fixed across occasions to mean parameter values, indicated PA and NA have disparate psychometric properties. Specifically, NA items targeted participants very poorly, with barely any overlap between the item location distribution and person location distribution in both the NDSHWB and Maastricht data sets. Also, person parameters estimated from NA responses had larger standard errors than their PA counterparts. Finally, NA analyses produced lower separability coefficients, and consequently signal-to-noise ratios, than PA analyses. The signal-to-noise ratio associated with the median separability coefficient for NA PCMs in the NDSHWB data indicated there was 1.59 times the amount of signal as there was noise in the NA data, whereas the corresponding PA signal-to-noise ratio was 17.89 to 1. Separability results for the Maastricht data was even poorer, with a signal-to-noise ratio of 3.80 for PA, indicating more signal than noise, and only 0.24 for NA, indicating more noise than signal was present.

Further differences between PA and NA were found in the temporal stability of psychometric characteristics. In cross-sectional PCMs analyzing each occasion separately, the rank order of item location parameters (betas) changed order much more over the measurement period for NA items in both data sets than for PA items. The lack of overlap between item and person location distributions in NA models likely contributed to the temporal instability observed in NA beta parameters. Most of the NA items in both data sets were much too high on the latent dimension for respondents to

strongly endorse. Such poor targeting could have caused the estimates of NA item locations to be less precise than necessary. Thus, NA items likely had location parameters very close to one another on the estimated logit scale. Without enough information from well-matched respondents to have adequate precision in estimating these item betas, parameters will be more likely to change orders across occasions, resulting in the observed temporal instability of beta estimates.

A similar instability was observed in delta estimates for NA items in both data sets. Not only were deltas often in the incorrect order across occasions, the incorrect order on one occasion was often different than the incorrect order on other occasions. Delta parameters were unstable in an unsystematic way, a difficult problem to solve. Collapsing the observed response scale, a 5-point scale in the NDSHWB data and a 7-point scale in the Maastricht data, to a binary response scale greatly improved the overlap between item and person location distributions. Also, binary response scales precluded disordered deltas, as each item only had one estimated delta parameter.

In the Maastricht data, PA items exhibited temporal instability of delta parameters; however this temporal stability was more systematic than that of NA delta parameters. Often, two of the delta parameters on the lower half of the scale had disordered estimates, with the second delta parameter (between categories 2 and 3) having the lowest negative estimate and the first delta parameter (between categories 1 and 2) having a higher estimated value. It could be that administering a 7-point response scale required more precision than participants used, resulting in additional noise in the data. If participants were particularly poor at distinguishing between low levels of PA

adjectives, more noise, and consequently more disordered deltas, would be observed at the lower end of the response scale, as seen in the current Maastricht data analyses.

Alternatively, this more systematic disordering of delta estimates for PA items in the Maastricht data might have been a result of self-report bias. Perhaps social desirability bias influenced participants to be uncomfortable with reporting low levels of a PA adjective unless they had already reported high levels of many other forms of PA. For example, an individual may not have wanted to report feeling only a little bit Happy. However, if that individual has already strongly endorsed other PA items, such as Alert, Active, Attentive, and Stimulated, he or she might be less susceptible to social desirability bias, having already demonstrated a high level of PA. Then the individual might be more willing to report low levels of happiness, because a socially desirable level of happiness has already been reported. This hypothetical self-report process would result in participants with high overall PA scores, and thus high theta estimates, endorsing low categories for some PA items, such as happiness, contributing to disordered delta estimates affecting primarily the lower end of the response scale.

Regardless of the underlying mechanisms behind the observed results, one conclusion is explicit: PA and NA measures do not have similar desirable measurement properties. The NA measures in both longitudinal data sets had much less desirable psychometric characteristics than PA measures. Given that these two longitudinal data sets include two different samples, sets of affect items, response scales, and measurement times scales, it is fair to assume the observed differences between PA and NA psychometric properties would generalize to other data sets with older adult (e.g., 50- to

90- years of age) or younger female (e.g., 20- to 40- years of age) participants, measured with a wide variety of items (e.g., from the PANAS [Watson et al., 1988], the Circumplex Model of Emotion [Larsen & Diener], and other sources), using a variety of Likert-type resonse scales (e.g., with five or seven categories), over a variety of time scales (e.g., several times a day for a few days or once a day for several weeks).

Challenging Measurement Assumption #2

The second measurement assumption challenged by this work is the assumption PA and NA are adequately captured by the same Likert-type response scale. PA and NA items are rarely administered with different response scales in the affect literature (for an exception see Schmidt, 2006), and most often both PA and NA are administered with either a 5-point or 7-point Likert-type scale (e.g., Crawford & Henry 2004; Crocker, 1997; Kercher, 1992; Watson et al., 1988). This consistent use of the same scale for both types of affect implies both types of affect are captured adequately by these response scales. To challenge this assumption, all possible collapsed response scales were explored and best-performing scales for PA and NA items were chosen separately in each data set.

Results revealed some discrepancies between statistics in collapsed response scales favored. Infit mean squares and the standard deviation of infit mean squares favored collapsed response scales with fewer categories for both PA and NA. Outfit mean squares favored a mix of collapsed response scales, some with more categories, some with less. Similarly, the ratio of separation to infit mean squares standard deviation provided mixed results.

Response scales favored by the remaining five statistics, however, were consistent. Item and person reliability and separability, item response-total score correlation, item summed score-beta estimate correlation, and person summed score-theta estimate correlation all favored a 5-point response scale for measuring PA and a binary response scale for measuring NA in both data sets, regardless of the differing lengths in the original response scale administered (e.g., five categories for NDSHWB and seven categories for Maastricht). Thus, this large difference in ideal response scales for PA and NA is robust across two very different sets of items, administered with two different response scales to samples with disparate demographic characteristics, over two different time scales. This is strong evidence that commonly used response scales for affect items, specifically 5- and 7-point Likert-type scales, do not adequately capture NA. Additionally, 7-point scales may not be capturing PA adequately either. Thus, results from the current study provide strong support in opposition of the second measurement assumption commonly made in the affect literature.

It is important to note that collapsing to two or five categories after administering a response scale with more than five categories is not the same as administering a response scale with two or five categories in the first place. It is impossible to know whether administering a binary response scale for NA items will remedy any of the problematic measurement characteristics identified for these items without testing the administration empirically. Administering a binary response scale for NA items may result in participants rarely using the higher of the two response categories, just as the highest of five and seven response categories were rarely used in the present study.

The major point of this collapsed response scale investigation was that the current convention of administering 5- or 7-point Likert-type scales for PA and NA items and analyzing the resulting responses as though they have desirable measurement characteristics and are adequate representations of PA and NA is completely incorrect. Separability coefficients for NA responses in both data sets indicated much smaller signal-to-noise ratios in these responses compared to PA responses. For some persons and items, there was more noise than signal in the data. Thus, if responses on these 5- and 7-point scales were used in other analyses, results may reflect analyzed noise more than analyzed PA and NA. For this reason, responses were recoded based on the optimal collapsed response scales for each type of affect before continuing on with exploratory factor analyses. Before affect data are used in analyses, researchers must check the measurement properties of their affect measures, to detect violations of any major assumptions they may be making in other analyses, such as the two measurement assumptions challenged in this work.

It is imperative that a new framework for measuring and modeling affect be employed. This framework must follow four major assumptions. Two of these assumptions are supported by the work presented above. First, PA and NA do not have the same, desirable measurement properties, and new measurement techniques are needed for NA. Second, PA and NA are not captured adequately by the same response scale. Specifically, commonly used response scales do not adequately capture NA, and 7-point scales do not capture PA as well as 5-point scales. The last two assumptions of this new framework are reviewed in the next chapter, followed by results supporting their validity.

Chapter 7: Testing Ergodicity Assumptions

Data Analysis

In this section, analyses employed to test the individual difference assumptions often made about affect are discussed, followed by the results they produced. These two assumptions include the following: 1) PA and NA have the same factor structure across as within individuals; and 2) The relationship among PA and NA looks the same across as within individuals. First, proposed analyses and software limitations preventing their implementation are briefly discussed. Second, alternative analyses are discussed in detail. Finally, results from these alternative analyses are reviewed.

Proposed Analyses and Attempted Implementations

The proposed analyses included an application of the Idiographic Filter (IF; Nesselroade et al., 2007) to the common factor model to examine individual differences in the factor structure and correlation of PA and NA, if such differences exist. Additionally, proposed analyses involved recoding the two longitudinal data sets examined previously using the best-performing collapsed response patterns for each type of affect in each data set. PA responses in both data sets were recoded according to the 5-point collapsing patterns that produced the most desirable IRT evaluation statistics (original scale, 12345, in the NDSHWB data, and 1122345 in the Maastricht data). NA responses in both data sets were recoded according to the binary collapsed patterns that performed best in previous PCM analyses (12222, NDSHWB; 1222222, Maastricht). To account for the ordinal nature of the data and the presence of missing data, threshold models were proposed with full information maximum likelihood estimation, with

weighted least squares and ordinary least squares estimation of the same models reserved as alternative estimation procedures.

Multiple approaches of running the proposed analyses on the NDSHWB data revealed current software and optimization procedures render these analyses computationally infeasible. The smallest non-null model proposed for the NDSHWB data included four threshold parameters for each of 20 PA items across all individuals (constrained to be equal), one threshold parameter for each of 18 NA items across all individuals (constrained to be equal), 38 factor loading parameters (constrained to be equal across individuals), 38 residual variances (constrained to be equal across individuals), and one factor correlation for each individual (307 parameters). Thus, optimization would have occurred in a 421-dimensional parameter space. In the largest non-null model proposed, the factor correlation parameter was fixed to be equal across individuals, and the factor loadings and residual variances were freed, resulting in a 23,431-dimensional parameter space. Approaches employed in Mplus version 7.0 (Muthén & Muthén, 2012) with both FIML and WLS produced either irresolvable errors due to the nature of NA responses (e.g., zero variance items and empty cells), or memory errors. Memory errors also halted the use of the Information Technology Services Linux Cluster at the University of Virginia.

Scaled down null models (i.e., no individual differences) attempted in OpenMx, including a reduced set of variables from the NDSHWB data with a 71-dimensional parameter space required over 24 hours to converge for a single model. Further reduced single-group confirmatory factor analysis models indicated the amount of time required

to reach convergence increased exponentially as the number of estimated parameters in the model increased. This scaling suggests the total runtime necessary for reduced, independent-group versions of the proposed analyses would require at least one year to complete, using only one set of starting values and barring any optimization problems, such as local minima or flat likelihood spaces. Other attempts at alternative methods included confirmatory and exploratory factor analyses applied to polychoric correlation matrices. The nature of NA manifest variables often resulted in improper correlation matrices, possibly due to empty cells in the contingency tables used to obtain polychoric correlations. Thus, factor analytic models on polychoric correlation matrices were also eliminated from possible alternative analyses.

Alternative Analyses

In order to examine individual differences in affect factor structure and factor correlations to refute assumptions about the ergodicity of affect often made in the literature, methods from previous empirical investigations were adapted and employed. Returning to the rare, early empirical investigations beginning to explore individual differences in affect factor structure, methods employed by Lebo and Nesselroade (1978), Zevon and Tellegen (1982), and Feldman (1995) were combined in a series of exploratory factor analyses on correlation matrices of Pearson coefficients.

Lebo and Nesselroade (1978) factor analyzed each of five pregnant women's responses to 75 affect adjectives over 15 weeks. Their approach was person-focused, allowing for the extraction of a different number of factors for each individual. Watson and Tellegen (1982) took a slightly weaker person-focused approach, but applied the

approach to over 20 participants. Two factors were extracted from each participant's data, and compared to and classified as conventional PA and NA factors. Thus, Lebo and Nesselroade (1978) gained much information about the affect factor structure of five women, and Watson and Tellegen (1982) gained less information about more people.

In the present investigation, exploratory factor analyses were conducted on each participant's data and on the data for each measurement occasion separately. Factors were classified in a less flexible manner than in Lebo and Nesselroade (1978), but with more flexibility than the methods used by Watson and Tellegen (1982). Factors could be categorized not only as PA and NA, but also as Bipolar Affect, Affect Magnitude, and Other. These classifications will be discussed in detail shortly.

To challenge the first ergodicity assumption, that affect factor structure is the same within as across persons, variances of loadings on the same classifications of factors were contrasted between persons and occasions. Additionally, correlations of factor scores obtained from a model in which loadings and all other parameters were fixed across participants were explored to compile evidence against the second ergodicity assumption, the assumption that affect factor correlations, particularly among PA and NA, are the same across individuals as within individuals.

These procedures are reviewed in more detail below, beginning with a brief review of exploratory factor analysis, followed by a discussion of the factor classification rules implemented.

Review of Exploratory Factor Analysis (EFA)

Data for each participant and each occasion were converted into a correlation

matrix. Pearson correlation coefficients were used instead of polychoric correlation coefficients. The process of estimating polychoric correlation matrices failed to converge for most participants. Constructing a matrix from pair wise polychoric correlations failed for fewer participants, but produced nonpositive definite correlation matrices for most individuals, hindering the use of factor analytic methods. Additionally, pairwise polychoric correlations allowed thresholds for a single item to change, depending on the other item involved in the correlation. The interval-level scale assumptions made under the Pearson correlation coefficient were used to remedy this problem. Although Pearson correlation coefficients are not ideal for ordinal data, they are adequate for obtaining results that will inform more rigorous investigations of individual differences in affect factor model components.

First, a correlation matrix of Pearson correlation coefficients was created for each individual and occasion of measurement. Then, each correlation matrix was factor analyzed using exploratory factor analysis (EFA) and the common factor model. Resulting parameters were examined to challenge the assumption that the factor structure, specifically factor loading parameters, of affect does not differ between individuals.

Finally, to challenge the assumption that correlations among affect factors are the same within individuals as across individuals, an EFA constraining loadings, uniquenesses, and factor variances to be equal across participants and occasions was conducted, and individual-level correlations among factor scores were examined. Below, each step of these analyses is described in greater detail below, beginning with a brief review of the common factor model and EFA.

Review of the Common Factor. From an EFA framework, the common factor model applied to a correlation matrix can be conceptualized and expressed as the following matrix algebra (Raykov & Marcoulides, 2008):

$$\Sigma = \Lambda\Phi\Lambda^{-1} + \Psi, \quad (13)$$

where Σ is the observed symmetric p by p correlation matrix for the p variables measured, Λ is a p by m matrix of factor loadings indicating the relationship between the p manifest variables and m latent factors, Φ is the symmetric m by m factor correlation matrix, and Ψ is a diagonal p by p matrix of uniqueness parameters indicating residual variance. When factors are uncorrelated, the uniqueness parameters from an EFA of a correlation matrix are a function of the factor loadings. For each item, the uniqueness is equal to the sum of the squared loadings across all factors subtracted from one. Factor loadings, uniquenesses, and the correlations among factors are estimated to create a model-implied correlation matrix that approximates the observed matrix as closely as possible.

Estimation. All EFAs were carried out using the "fa" function in the psych package (Revelle, 2012) in R 2.15.3 (R Core Team, 2013). Four estimation methods are available through the fa function, including maximum likelihood (ML), generalized least squares (GLS,) weighted least squares (WLS), ordinary least squares (OLS; also called minimum residual estimation), and principal axis factor estimation (PAF). In test cases with individual participants, EFAs with WLS and GLS failed to run for all test cases, ML

failed for all but one test case, and OLS and PAF succeeded in obtaining reasonable estimates for all test cases. Thus, OLS estimation was used in all following analyses. In ordinary least squares estimation, the eigen values of the original correlation matrix are iteratively adjusted to minimize the squared deviations between the model-implied correlation matrix and the observed correlation matrix. Parameters are estimated from Equation 13. Results reported here are from unrotated factors.

Factor Coding. To challenge the assumption that the factor structure of affect is the same across individuals as within individuals, the variance of each of the factor loadings across participants was compared to the variance across occasions. During factor extraction, factors were extracted in order of variance explained, and thus factor order was not guaranteed to be consistent across participants or occasions. If more of the variance in one participant's data was explained by a PA factor, PA was extracted first. If a second participant's data was explained most by a NA factor, NA was extracted first. Similarly, it was possible one or more factors for each participant represented latent dimensions other than PA and NA, such as a bipolar affect construct with large positive PA loadings and large negative NA loadings or an affect magnitude factor with positive loadings on both PA and NA items. In addition, the sign of loadings on a single factor was arbitrary. The signs of all loadings on a single factor could be reversed, along with the sign of the factor correlation, without affecting any other components or parameters in the EFA, other than the factor correlation. This technique was analogous to reflecting an axis through the origin of a geometric space. To avoid inflating variance estimates by

aggregating across factors of different substantive constructs and different signs, a classification scheme was created to categorize factors into a variety of types.

First, the number of loadings more extreme than the conventional cut-off of +/- .3 (Tabachnick & Fidell, 2001) was recorded. Next, the number of PA items with loadings greater than or equal to .3 was recorded, along with the number of PA items with loadings less than or equal to -.3, the number of NA items with loadings greater than or equal to 0.3, and the number of NA items with loadings less than or equal to -.3. Each of these counts was divided by the total number of loadings as or more extreme than +/-.3, resulting in proportions summarizing the structure of each factor. For example, a proportion of .9 for PA loadings greater than the cut-off .3 on a factor indicates 90% of the loadings as or more extreme than +/- .3 on that factor are PA loadings greater than .3. This example factor would be classified as PA. A factor with a proportion of .4 for PA loadings above .3 and a proportion of .6 for NA loadings below -.3 would indicate a factor measuring bipolar affect, with high positive loadings for PA items and low negative loadings for NA items.

Rules for classifying factors were implemented by creating cut-off proportion values (see Table 2). Cut-off proportions were based on decisions made manually, by classifying several factors in the data by hand and recording the conditions used to make these classifications. For example, in the NDSHWB, a PA factor may have all 20 PA items load heavily on it, along with a small handful of NA items. If four or less of the 18 NA items also loaded on the PA factor, the most frequent manually made decision was to label the factor as PA. If five or more of the NA items loaded heavily on the factor, the

decision resulted in changing the factor classification to Bipolar Affect, or another appropriate label. In the NDSHWB data, four items made up 10.5% of the 38 total items in the data, and five corresponded to 13.2% of the total items. Thus, 12% was a reasonable cut-off percentage to separate these numbers of items. In the Maastricht data, with 16 affect items, this 12% cut-off allowed for one of the eight NA items to load on the PA factor without changing its label in the example given above.

Table 2

Factor Classification Rules.

Factor Classification	**Classification Guidelines** ***Of all loadings as or more extreme than* +/- .3:**
PA	70% or more on PA items, values ≥ .3 12% or less on NA items
Reversed PA	70% or more on PA items, values ≤ -.3 12% or less on NA items
NA	70% or more on NA items, values ≥ .3 12% or less on PA items
Reversed NA	70% or more on NA items, values ≤ -.3 12% or less on PA items
Bipolar Affect	80% or more from (PA items, values ≥ .3 with NA items, values ≤ -.3) 12% or more on PA items, values ≥ .3 alone 12% or more on NA items, values ≤ -.3 alone 12% or less on all others
Reversed Bipolar Affect	80% or more from (PA items, values ≤ -.3 with NA items, values ≥ .3) 12% or more on PA items, values ≤ -.3 alone 12% or more on NA items, values ≥ .3 alone 12% or less on all others
Affect Magnitude	80% or more from (PA items, values ≥ .3 with NA items, values ≥ .3) 12% or more on PA items, values ≥ .3 alone 12% or more on NA items, values ≥ .3 alone 12% or less on all others
Reversed Affect Magnitude	80% or more from (PA items, values ≤ -.3 with NA items, values ≤ -.3) 12% or more on PA items, values ≤ -.3 alone 12% or more on NA items, values ≤ -.3 alone 12% or less on all others
Other	Any other configuration

Factors were classified using the guidelines in Table 2. For example, a factor was classified as representing PA if, out of the total number of loadings as or more extreme than +/- .3, at least 70% of them were high loadings on PA items, and less than 12 % were high positive or low negative loadings on NA items. Guidelines were applied to the two factors extracted from each participant's correlation matrix and each occasion's correlation matrix. The two factors obtained for any given participant or occasion could receive any combination of classifications. A participant may have two PA factors, a Bipolar Affect factor and a reversed NA factor, an Affect Magnitude factor and a Reversed Affect Magnitude factor, and so on.

Variance of Factor Loadings. In each data set, one of the most common factor classifications was chosen. The variance of factor loadings over participants on this type of factor was compared to the variance of factor loadings over occasions on the same type of factor. If a factor commonly found across both people and occasions was Bipolar Affect, then factor loading variances on only factors labeled Bipolar Affect factors would be examined. This prevented variance inflation from aggregating over multiple types of factors (e.g., variance across both PA and NA factors).

Variance of Factor Correlation. The factors extracted contributed to the correlation between factor scores. If PA and NA were extracted, the factor correlation may be near zero, whereas if Bipolar Affect and Reversed Bipolar Affect were extracted, the correlation would likely be much closer to -1. Allowing individuals to have different sets of factor loadings might artificially inflate the individual differences observed in factor correlations. In the originally proposed analyses, factor loadings and uniquenesses

were fixed to avoid this inflation. To solve this problem in the EFAs examined here, a single, unrotated 2-factor EFA was conducted on the Pearson correlation matrix for the entire NDSHWB data set using OLS estimation. Then, the factor scores obtained from this whole-sample EFA were correlated for each individual and each occasion. Thus, correlations between factors were obtained while factor structure remained fixed across participants and occasions. Variation in person-specific factor correlations were compared to variation in occasion-specific factor correlations to challenge the assumption that the association between PA and NA is the same across people as within people.

Chapter 8: Ergodicity Assumption Results

EFA Results

Results from the EFAs are discussed below. First, results challenging the assumption that affect factor structure is the same across as within individuals are explored. Then, results are examined challenging the assumption that the correlation among affect factors is the same across as within persons. For each assumption, results from the NDSHWB are reviewed first, followed by those from the Maastricht data.

Variance in Factor Structure

Upon analysis completion, each participant and occasion received one unrotated 2-factor solution estimated with OLS. First, the frequencies of various combinations of factor types are summarized for participants and occasions. Second, the variances of the loading estimates are explored, and variances of loadings obtained from single-participant models are compared to those obtained from single-occasion models.

NDSHWB. Frequencies of each type of factor observed for participants and occasions are presented in Table 3. Factor order was not taken into account in these frequencies. In the row for the combination of a PA and an Affect Bipolarity factor, PA may be the first factor extracted for some observations and the second factor extracted for other observations in that row. Similarly, reversed factors were not distinguished from their counterparts in Table 3. Thus, the observation of PA and NA is included in the same frequency as the observation of Reversed PA and NA. With five different types of factors, a total of 15 different combinations of factors could be observed. Only two of these 15 were observed for occasion-specific models, while 13 different combinations

were observed for person-specific models. Factor combinations varied much more between participants than between occasions.

Table 3

Frequencies of Factor Type Combinations in the NDSHWB Data

Factor Type Combination	*N* of Occasions	*N* of Participants
Bipolar Affect, Bipolar Affect	0	16
Bipolar Affect, Affect Magnitude	19	114
Bipolar Affect, NA	37	15
Bipolar Affect, PA	0	19
Bipolar Affect, Other	0	68
Affect Magnitude, Affect Magnitude	0	2
Affect Magnitude, NA	0	0
Affect Magnitude, PA	0	6
Affect Magnitude, Other	0	7
NA, NA	0	0
NA, PA	0	4
NA, Other	0	2
PA, PA	0	5
PA, Other	0	25
Other, Other	0	9
Missing Results	0	15
Total *N*	56	292

For all of the 56 occasion-specific EFAs, Bipolar Affect was the first factor extracted. Individuals with Bipolar Affect extracted as the first factor ($n = 175$) were selected as a comparison group. Variances and standard deviations of each of the loadings across occasion- and person-specific models are shown in Table 4. For every factor loading, the associated variance between persons was approximately one order of magnitude larger than the corresponding variance between occasions. These results suggest the factor structure of affect varied much more between people than occasions, even when examining the same type of factor, Bipolar Affect. Thus, the factor structure of affect across individuals (within occasions) was not the same as within individuals.

Table 4

Factor Loading Variances Over Persons and Occasions in the NDSHWB Data.

Loading	**Variance Over *N* = 56 Occasions *SD* (*Var*)**	**Variance Over *N* = 175 Participants *SD* (*Var*)**
Active	0.048 (0.002)	0.222 (0.049)
Calm	0.050 (0.003)	0.316 (0.100)
Alert	0.047 (0.002)	0.237 (0.056)
Attentive	0.054 (0.003)	0.216 (0.047)
Elated	0.024 (0.001)	0.215 (0.046)
Determined	0.049 (0.002)	0.253 (0.064)
Stimulated	0.035 (0.001)	0.223 (0.050)
Happy	0.024 (0.001)	0.164 (0.027)
Enthusiastic	0.026 (0.001)	0.213 (0.045)
Excited	0.034 (0.001)	0.208 (0.043)
Love	0.049 (0.002)	0.240 (0.058)
Proud	0.032 (0.001)	0.284 (0.080)
Joyful	0.024 (0.001)	0.175 (0.031)
Strong	0.033 (0.001)	0.250 (0.062)
Interested	0.037 (0.001)	0.209 (0.044)
Pleased	0.028 (0.001)	0.158 (0.025)
Content	0.030 (0.001)	0.196 (0.039)
Aroused	0.046 (0.002)	0.319 (0.102)
Inspired	0.037 (0.001)	0.190 (0.036)
Euphoric	0.037 (0.001)	0.284 (0.081)
Afraid	0.051 (0.003)	0.264 (0.070)
Unhappy	0.053 (0.003)	0.207 (0.043)
Annoyed	0.060 (0.004)	0.235 (0.055)
Ashamed	0.056 (0.003)	0.269 (0.072)
Guilty	0.053 (0.003)	0.273 (0.075)
Angry	0.066 (0.004)	0.257 (0.066)
Sad	0.056 (0.003)	0.248 (0.062)
Hostile	0.052 (0.003)	0.255 (0.065)
Upset	0.052 (0.003)	0.232 (0.054)
Irritable	0.050 (0.002)	0.246 (0.061)
Depressed	0.059 (0.003)	0.227 (0.051)
Jittery	0.052 (0.003)	0.296 (0.087)
Drowsy	0.056 (0.003)	0.235 (0.055)
Sluggish	0.056 (0.003)	0.243 (0.059)
Worried	0.052 (0.003)	0.260 (0.068)
Nervous	0.052 (0.003)	0.302 (0.091)
Fatigued	0.051 (0.003)	0.241 (0.058)
Distressed	0.056 (0.003)	0.235 (0.055)

Maastricht. Results from the Maastricht data confirm those reported from the NDSHWB data. Observed frequencies of all possible factor combinations in the Maastricht data are reported for occasions and persons separately in Table 5. Only three of the 15 possible combinations were observed for occasions, whereas 14 of the combinations were observed in person-specific EFAs. Again, the factor structure of affect varied more between participants than between occasions.

Table 5

Frequencies of Factor Type Combinations in the Maastricht Data

Factor Type Combination	*N* of Occasions	*N* of Participants
Bipolar Affect, Bipolar Affect	0	43
Bipolar Affect, Affect Magnitude	27	34
Bipolar Affect, NA	22	19
Bipolar Affect, PA	0	10
Bipolar Affect, Other	0	106
Affect Magnitude, Affect Magnitude	0	1
Affect Magnitude, NA	0	1
Affect Magnitude, PA	0	2
Affect Magnitude, Other	0	9
NA, NA	0	0
NA, PA	1	4
NA, Other	0	2
PA, PA	0	4
PA, Other	0	14
Other, Other	0	14
Missing Results	0	4
Total *N*	50	263

Factor loading variances obtained from the Maastricht data also agreed with the NDSHWB findings. Almost all of the occasion-specific EFAs had first factors that could be identified as Bipolar Affect. Participants with Bipolar Affect extracted as the first factor in the EFA were selected as a comparison group (n = 150). Variances of factor

loadings between occasions were approximately an order of magnitude smaller for almost all of the items included in the analyses than the factor loading variances between persons (see Table 6). Differences in factor structure between participants were much larger than those between occasions, even when examining the same type of factor, extracted in the same order, for both persons and occasions.

Table 6

Factor Loading Variances Over Persons and Occasions in the Maastricht Data

Loading	**Variance Over *N* = 44 Occasions *SD* (*Var*)**	**Variance Over *N* = 150 Participants *SD* (*Var*)**
Cheerful	0.045 (0.002)	0.178 (0.032)
Satisfied	0.052 (0.003)	0.181 (0.033)
Energetic	0.071 (0.005)	0.253 (0.064)
Enthusiastic	0.056 (0.003)	0.205 (0.042)
Pleased	0.084 (0.007)	0.218 (0.048)
Alert	0.072 (0.005)	0.243 (0.059)
Active	0.082 (0.007)	0.257 (0.066)
Calm	0.066 (0.004)	0.306 (0.093)
Unsure	0.089 (0.008)	0.280 (0.079)
Lonely	0.078 (0.006)	0.267 (0.071)
Anxious	0.112 (0.013)	0.285 (0.081)
Gloomy	0.082 (0.007)	0.240 (0.058)
Guilty	0.100 (0.010)	0.260 (0.068)
Suspicious	0.103 (0.011)	0.269 (0.072)
Angry	0.078 (0.006)	0.227 (0.052)
Tired	0.071 (0.005)	0.242 (0.059)

Summary. In both data sets, the factor structure of affect was not the same within individuals as it was across individuals. Participants had loadings that varied much more than loadings for separate occasions, with each occasion-specific EFA offering a factor structure of affect across participants. Also, individuals had many more combinations of different types of factors than did occasions, implying person-specific factor structures

are more variable than occasion-specific (across person) affect factor structures. Thus, the factor structure of affect is not the same within individuals as it is across individuals

Variance in Factor Correlation

After fixing all factor analysis parameters to be equal across all occasions and all participants, the correlation between PA and NA factor scores was obtained for each person and each occasion. The variation in these correlation coefficients for individuals and occasions is examined below.

NDSHWB. The factor structure obtained across all participants and occasions included Bipolar Affect factor as the first factor extracted, and NA as the second. Factor scores were obtained for each response observed, with participants receiving one score on each of the two factors for each measurement occasion. The distribution of correlation coefficients representing the association between the two factors across participants is plotted in Figure 29, along with the distribution of factor correlations across occasions.

The correlation between factor scores spanned a much wider range of values in the distribution across participants. The distribution summarizing the correlation across occasions is much narrower. The mean of the correlation distribution for occasions was very close to zero (M = .02, *Median* = .03) and the standard deviation of the distribution was small (SD = .07). The correlation distribution for participants also had a mean very close to zero (M = .05, *Median* = .02); however, the standard deviation of the distribution was approximately seven times that of the distribution for occasions (SD = .51). Thus, the correlation between affect factors varied much more between participants than between occasions.

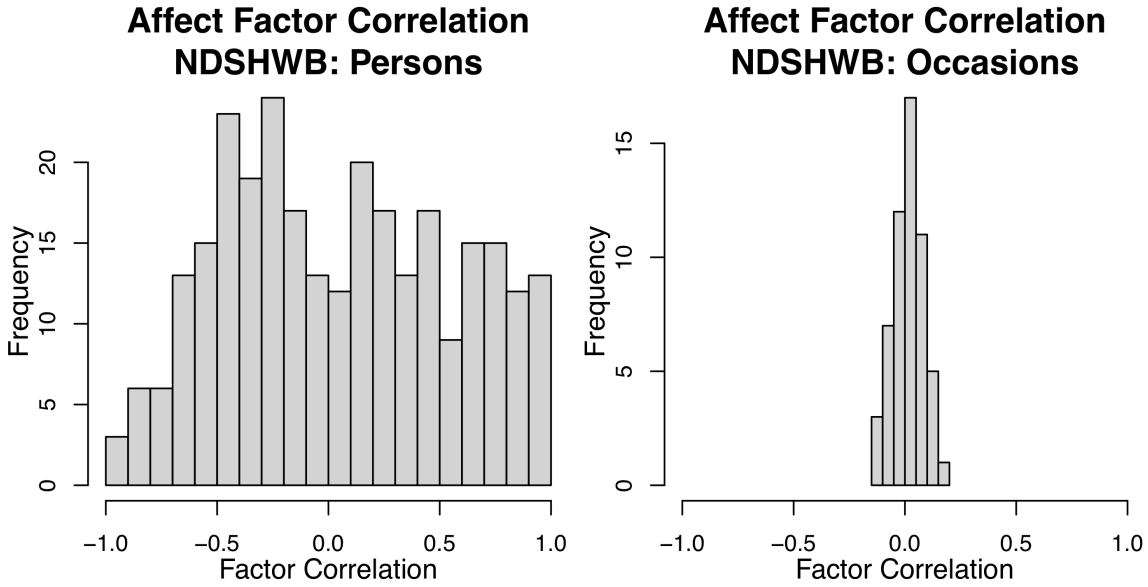

Figure 29. Distribution of factor score correlations across participants and across occasions in the NDSHWB data.

Maastricht. A similar pattern of results exists for factor correlation coefficients obtained from the Maastricht data. The factor structure fixed to be the same across all participants and occasions included Bipolar Affect as the first factor extracted and NA as the second. The distributions of factor score correlations across participants and occasions are plotted in Figure 30.

As with the NDSHWB data, the factor score correlation distribution was wider for persons than occasions. The mean of the distribution for occasions was again very close to zero (M = .01, *Median* = -.01), and the standard deviation was small (SD = 0.10). In the Maastricht data, the mean of the factor score correlation distribution for persons was farther from zero than that of occasions (M = .10, *Median* = .21), and the standard deviation was much larger than that of the occasion distribution (SD = .52). Again, the correlation among affect factors varied more by person than by occasion, indicating this

correlation was not the same within individuals as across individuals.

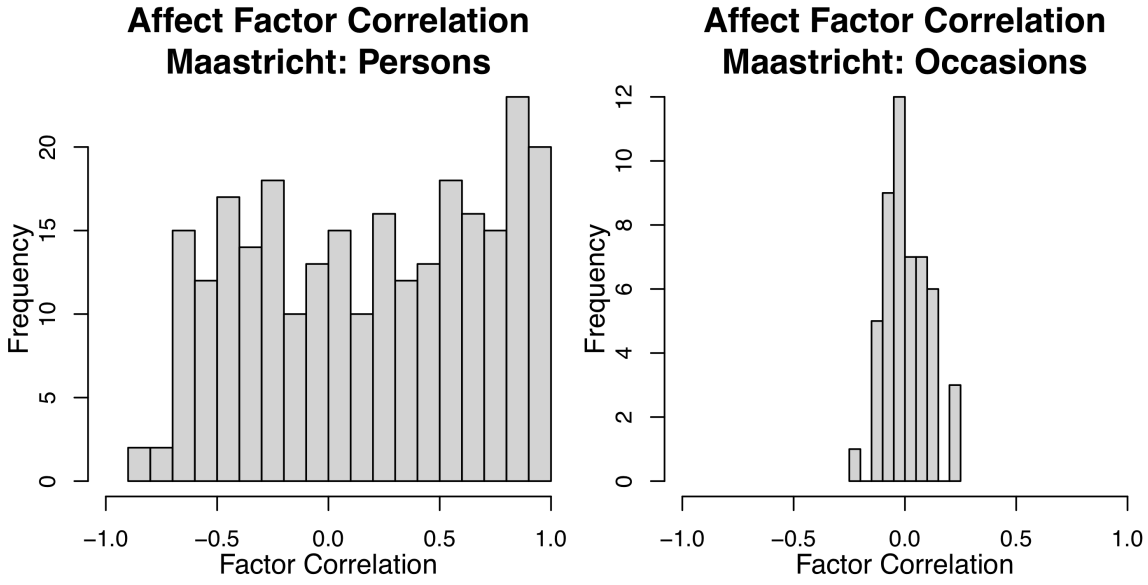

Figure 30. Distribution of factor score correlations across participants and across occasions in the Maastricht data.

Summary. Results from both data sets indicate that the correlation among affect factors varied much more between individuals than between occasions. These findings provide evidence against the assumption that the correlation of affect factors is the same within as across individuals. When occasion-specific portions of data, data across individuals, were analyzed, the differences between occasions were not at all like the differences between person-specific analyses.

Conclusions from EFAs

Challenging Ergodicity Assumption #1

Much of the investigations of affect factor structure in the literature analyze data from a cross-sectional sample of individuals and then draw inferences about the structure of a measure of affect for future respondents (e.g., Crawford & Henry, 2004; Gaudreau et

al., 2006; Watson et al., 1988). Such inferences assume the factor structure of affect is the same for all individuals. Exploratory factor analysis of data for each individual and each occasion were conducted, and variances of factor loadings over participants were compared to variances of loadings over occasions to challenge this assumption.

Factors were classified into five different types, identified as PA, NA, Bipolar Affect, Affect Magnitude, and Other. The variances of factor loadings on Bipolar Affect were examined for all individuals and occasions with Bipolar Affect extracted as the first factor in single-person and single-occasion EFAs. Variances of loadings summarizing differences in factor structure between participants were approximately an order of magnitude greater than the variances of loadings summarizing differences between occasions. Additionally, the number of different combinations of factors extracted was much larger for participants than for occasions. In the NDSHWB data, two different combinations of factors were identified from occasion-specific models, while 13 different combinations were identified in the person-specific analyses. In the Maastricht data, three combinations were identified in occasion-specific EFAs and 14 unique combinations were identified in person-specific EFAs. Participants had much larger differences in affect structure than occasions. These findings refute the assumption that the factor structure of affect is the same at the individual level as it is at the sample level. The factor structure of affect within participants is very different from the factor structure of affect across participants.

Differences between participants likely contribute to many of the conflicting findings between cross-sectional investigations of affect factor structure. Perhaps a

source of different factor solutions in different samples, particularly when the same affect measure is used, is the aggregation of individuals with very different factor structures. For example, in the NDSHWB, several participants had a Bipolar Affect factor and an Affect Magnitude factor extracted from their data. For some people, these factors may be very similar to Valence and Arousal factors. If a random sub-sample of these individuals were selected to participate in another affect study, and the majority of participants selected had a factor structure for affect similar to Valence and Arousal, a cross-sectional EFA of new affect ratings from these participants would likely produce a Valence and an Arousal factor for the sample. In contrast, if most of the selected participants had a factor structure consisting of PA and NA, the cross-sectional EFA would produce PA and NA factors. The results obtained with a cross-sectional, nomothetic approach depend on the characteristics of the sample. This sensitivity to sample characteristics greatly restricts the generalizability of results and slows scientific progress in the field of affect.

In order to prevent prolonging debates about factor structures that are actually dependent on sample characteristics, future work must refrain from assuming the factor structure of affect is the same across individuals as it is within individuals. This is not to suggest all future affect research must consist solely of multilevel investigations conducted on longitudinal data. However, researchers can and should take action to help expose the fallacy of this assumption in the literature. Specific actions for doing so are discussed in Chapter 9. Next, conclusions from results challenging the second ergodicity assumption are discussed.

Challenging Ergodicity Assumption #2

The bipolarity of PA and NA, as well as other affect factors, has been debated for decades. Several theories and models of affect posit the correlation between affect factors is the same for all individuals (e.g., Bipolar and Bivariate Models of affect, see Reich Zautra, 2002). In the present study, factor score correlations were examined for each participant and occasion to challenge this assumption. Factor scores were obtained from an EFA of data from all participants and occasions, and thus factor loadings, uniqueness parameters, and factor variances had the same estimated values for all participants and occasions.

The distribution of correlations among factor scores for participants had a much wider range and larger variance than the distribution of factor score correlations for occasions in both of the longitudinal data sets examined. Thus, differences between participants in the correlation among affect factors were much larger than differences between occasions. The correlation of affect factors within participants was not the same as the correlation of affect factors across participants. These findings provide strong support against the assumption that the correlation among PA and NA, or other affect factors, was the same within as across participants.

Research continuing to make this false assumption will hinder scientific progress by contributing to debates about affect construct correlations in which all sides are not properly informed. Future research must acknowledge individual differences in affect structure and adapt techniques to address these differences among individuals.

Chapter 9: Review of Conclusions and Discussion

Support for New Measurement Assumptions

The first goal of this dissertation project was to test the validity of two assumptions commonly made about the measurement of PA and NA in the affect literature: 1) PA and NA have similar desirable measurement characteristics; and 2) PA and NA are adequately captured by the same response scale, in two longitudinal data sets with disparate sample characteristics, item sets, response scales, and periods of measurement.

Longitudinal PCM analyses revealed large differences in the psychometric properties of PA and NA measures from both of these longitudinal data sets. Specifically, NA measures do a very poor job of targeting respondents, with most items located very high on the latent NA dimension, and most individuals located at the low extreme. While this is desirable from a psychological perspective, as individuals are reporting low levels of NA across occasions, from a measurement perspective, it is a problem. Such a large gap between item and person distributions on the latent NA dimension results in inaccurate estimates of NA scores. In addition, NA measures exhibit many instances of disordered deltas, thresholds separating adjacent categories, and unstable item locations over time. The disordered deltas observed might in part be a symptom of an inadequate response scale. When given five or seven categories in a Likert-type scale, participants responding to NA items tended to use only about the lower half of the response scale.

The PA measures in both data sets had more desirable psychometric characteristics. Both measures targeted a large proportion of the sample well and

produced more accurate estimates of person location scores. The PA measure used in the Maastricht data demonstrated many instances of disordered deltas; however, reversals appeared to be localized to the lowest two deltas, suggesting a temporally stable problem with response scale use. All of the psychometric properties examined had more desirable results for the PA measures than the NA measures, refuting the assumption that PA and NA have similar desirable measurement characteristics.

The validity of the second assumption was challenged by evaluating all possible patterns of collapsing the administered response scales. A binary, or at the very most 3-category, response scale produced much more desirable psychometric statistics than the original 5- and 7-point scales administered for both NA measures. For both PA measures, a 5-category response scale functioned well. These differences in optimal response categories, of the collapsed response scales examined, refute the second assumption that PA and NA are adequately captured by the same response scale.

Taken together, these results highlight the need for a new framework for affect measurement research. Continuing to make these false assumptions will continue to corrupt research on affect and well-being across the lifespan with inaccurate measures of NA that may contain more noise than NA signal. Future research must adopt two new assumptions that are strongly supported by the results presented in this project. First, PA and NA measures, as currently used, have very different psychometric properties, and specifically NA measures are in need of much improvement. Second, PA and NA are not adequately captured by the same response scale. Although a 5-point Likert-type scale

appears to function adequately for PA measures, investigations of alternative response schemes are sorely needed for NA measures.

None of this is suggesting researchers desist from measuring NA until perfect NA measures are constructed. Rather, this is a call for work aimed at improving measures of NA. This is also a caution to researchers who commonly use NA measures similar to the ones explored here to take other approaches in their research to prevent further slowing scientific progress in the field of affect. Operating under a new framework in which measurement properties are a major focus and measurement problems are acknowledged will only help us learn more about affect, particularly about NA. As better measures are developed, more will be learned about how individuals evaluate NA and why we often see such low levels of NA in self-report data in research in the United States. The more we discover about these topics, the more accurately we will be able to uncover affect's role in health and well-being across the lifespan.

Discussion

In this discussion, recommendations for affect researchers and anyone who measures affect are provided. Next, limitations and strengths of the work presented here are discussed, followed by directions for future research.

Recommendations for Researchers. First and foremost, it is highly recommended that researchers check the measurement properties of any assessments used in empirical investigations. IRMs are an excellent method to use in making this check when data consist of dichotomous or polytomous ordinal item responses. Of course, using multiple methods, such as IRMs and factor analysis will yield more information than

limiting measurement evaluations to only one method. If the psychometric properties of any measures used are poor, researchers can use a variety of techniques to attempt to improve measurement properties, such as removing items or collapsing response categories.

With affect specifically, researchers are urged to examine the measurement properties of any affect items used, particularly those used to measure NA. Recall that one of the NA measures explored in this work had a signal-to-noise ratio below 1, indicating more noise than signal, under the original response scale. Collapsing response categories down to a binary recoding of NA responses improved the signal-to-noise ratio three-fold. Without checking measurement characteristics, there is no way to know whether the measures employed are measuring constructs with any reasonable level of accuracy or precision. Checking measurement properties must become a critical component of empirical investigations of affect.

If it is not possible to examine the measurement characteristics of affect assessments used in a study, researchers are urged to administer items measuring PA with response scales of no more than five categories. Similarly, it is strongly recommended that researchers administer NA items with response scales with a smaller number of categories, or that they administer a response scale with no more than five categories and plan to examine results of recoding NA responses into a binary or 3-level ordinal variable with methods other than IRMs. Exploring various recoding schemes for NA data may lead to better detection of effects of interest if collapsing categories reduces imprecision induced by overly complex response scales. Regardless of whether this imprecision

exists, examining results of a study with and without recoding the data will provide more information about the nature of NA responses than only examining the original data, particularly when measurement characteristics are not evaluated.

Limitations. The investigation of collapsed response scales may be too impractical for many researchers employing affect assessments to use. Ideally, affect assessment users will employ measurement models to examine the psychometric properties of any measures they use in an empirical investigation and address problematic properties of these measures with further analyses or modifications of the data. However, depending on the goal behind the measurements being collected, affect assessment users may not deem it worthwhile to carry out such investigations. For example, if a researcher is exploring relationships between affect factors and Schizophrenic symptoms, the sample available to the researcher may be too small to make intensive psychometric analyses feasible or worthwhile. This infeasibility is why alternative suggestions are provided in the previous section for measuring affect and evaluating affect measurement with methods that do not succumb to the two false measurement assumptions explored here.

Strengths. Perhaps the most important strength of this work is its generalizability to individuals, items, measurement periods, and response scales similar to those used in either of the longitudinal data sets examined. The NDSHWB data included older adults as participants, aged 53 to 91 years, administered 38 affect items from a variety of affect measures with a 5-point response scale once a day for 56 consecutive days. The Maastricht study included younger female adults aged 18 to 46 years, administered 16

affect items, 9 of which matched items in the NDSHWB data, ten times a day for five consecutive days with a 7-point response scale. Despite the different participants, items, measurement time scales, and response scales used, similar results refuting the measurement assumptions commonly made about affect measures were obtained from both data sets. Thus, results can be generalized to other samples of older adult participants or younger women, other measures of affect with items similar to item set used in either of the NDSHWB and Maastricht studies, administered daily or several times a day, with either a 5- or 7-point response scale. It is likely these results will hold in a variety of other studies.

Future Research. Much work remains to be done in future studies on affect measurement. The measurement of NA poses a particularly large problem. Although collapsed response scale evaluation indicated NA was best captured by a binary collapsed response scale, there is no guarantee administering NA items with a binary scale would not result in the same low category frequency problems observed with the 5- and 7-point scales used in the NDSHWB and Maastricht studies, respectively. This is not to suggest such administration should not be attempted. A study testing a variety of response scales by administering various scales to participants rather than exploring post-hoc adjustments to a single response scale would greatly inform the literature on affect measurement. Such a study is sorely needed.

However, exploring alternative methods of parsing apart self-reported NA into self-reported experience and social desirability bias would also be incredibly beneficial to the field. For example, a social desirability measure could be administered along with

affect items, and the relationship between these constructs could be partialled out of affect responses before including the responses in other analyses. Another possible technique could include altering an individual's frame of reference for evaluating NA by asking the individual how irritable or sad he or she is compared to the best he or she has ever felt. Finally, it may be the case that NA is so strongly linked to external events that measuring NA without context is not worthwhile. To explore this possibility, future research could attempt to induce facets of NA and compare responses from questions asking participants to report how much of those NA facets they are experiencing that reference the event to questions that do not reference the event.

These suggestions are a small sample of the work necessary for informing major improvements in affect measures. It is critical to research on affect and well-being across the lifespan that new techniques for measuring affect are explored. Continuing to assume that PA and NA have similar desirable measurement properties and are captured adequately by the same response scale will grievously slow progress in field of affect research by producing inaccurate, noisy measurements of NA that should not be used in further analyses.

It is imperative that a new framework for measuring and modeling affect be employed. This framework must follow four major assumptions. Two of these assumptions are supported by the work reviewed above. First, PA and NA do not have the same desirable measurement properties, and new measurement techniques are needed specifically for NA. Second, PA and NA are not captured adequately by the same

response scale. The work presented here supporting the last two assumptions of this new framework are discussed in the following section.

Support for New Individual Differences Assumptions

The second goal of this dissertation project was to evaluate the validity of two ergodicity assumptions commonly made in the affect literature: 1) The factor structure of PA and NA is the same across individuals as it is within individuals; and 2) The correlation among PA and NA is the same across as within individuals, using data recoded according to the best-performing collapsed response scales for PA and NA in each of the two longitudinal data sets examined.

Results from person- and occasion-specific EFAs revealed substantial differences in factor structures between individuals but not between occasions, when a 2-factor solution was forced. Person-specific models contained a larger number of combinations of identified factor types (e.g., PA, NA, Bipolar Affect) in both data sets, whereas occasion-specific models conducted on both data sets showed much more consistent factor combinations. Additionally, factor loadings varied more across person-specific analyses than across occasion-specific analyses, often by an order of magnitude.

Finally, Results from whole-data EFAs showed similar individual variation in factor correlation coefficients that was not found across occasions. In sum, these findings refute the ergodicity assumptions examined and prompt the incorporation of two new assumptions into the new framework for affect research discussed here: 1) The factor structure of PA and NA differs by individual; and 2) The correlation among affect factors

differs by individual. Notably, these differences are not observed, at least not at the same magnitude, between occasions (across individuals).

Future affect research must acknowledge the presence of these individual differences. Failure to do so will only prolong debates in the affect literature that are likely caused by the aggregation over individual differences and thus severely hinder scientific progress. Acknowledging these individual differences will allow for the development of new, more accurate theories of the structure and bipolarity of affect, as well as of the influence of affect components on other constructs. Recommendations on how to adhere to these new individual differences assumptions are discussed below, along with limitations and strengths of the presented work, and suggestions for future research.

Discussion

Recommendations for Researchers. Changes in current affect measurement and modeling techniques are necessary in order to continue conducting empirical investigations of affect with scientific rigor. This is not to suggest every future study of affect must be longitudinal and account for individual differences in affect structure and bipolarity; however, steps must be taken to prevent the false ergodicity assumptions examined here to continue being made in the literature. Of course, if collecting longitudinal data is feasible, repeated measures and multilevel methods can be employed to explore individual differences and/or to filter out individual differences from other analyses, depending on whether the goal of the study is to examine individual differences or control for them.

Without longitudinal data, techniques for grouping individuals with similar affect structures, such as cluster analysis or latent class analysis, can be used to more accurately approximate the structure of affect and the correlations among affect factors for individual participants. For example, consider a researcher attempting to link PA and NA to a measure of physical functioning with a sample of chronic pain sufferers and one measurement occasion. Running a confirmatory factor analysis on all PA items, NA items, and physical functioning items over the entire sample will yield one correlation coefficient representing the association between PA and physical functioning for the entire sample and one representing the association between NA and physical functioning. However, affect factor structure is not the same for each individual. Thus, the estimated factor correlations may be poor representations of the actual correlations between PA, NA, and physical functioning for many of the participants.

To better approximate the correlations among the three constructs for every participant, individuals could be partitioned into groups of people with very similar affect structures with a grouping technique like cluster analysis. Once the optimal number of groups or clusters is chosen, the same confirmatory factor analysis examining relationships between affect and physical functioning can be conducted on data from each group separately. Perhaps when the whole sample is analyzed together, the factor correlations estimated are very weak. However, when participants are grouped by affect responses and reanalyzed by group, the researcher may find these factor correlations are strong for one or more groups, but are attenuated by another group with factor correlations close to zero. Now the researcher has better explained each individual's data

and has learned more about relationships between affect and other variables in the study. After an analysis similar to the one above, other differences between groups or clusters can be explored that may aid in explaining the different effects found in each group.

Even without using conventional grouping techniques, individual differences can still be explored by analyzing grouping covariates already included in the data set, such as demographic variables. Finally, regardless of whether any methods suggested here are used, researchers must be careful about how strongly they generalize results. Given that affect structure and factor correlations vary greatly by individual, it is likely analyses of affect and related constructs will produce results that vary by the sample analyzed. It is imperative not to generalize results beyond the samples, items, and time scales to which they are reasonably similar.

Limitations. The exploratory factor analyses used to challenge assumptions of ergodicity have two main limitations. First, due to the nature of NA responses, Pearson correlation coefficients had to be used in place of polychoric correlation coefficients, as polychoric correlations too often resulted in nonpositive definite correlation matrices. Second, ordinary least squares (OLS) estimation was used in place of full information maximum likelihood (FIML) estimation due to issues with improper correlation matrices. When missing data are included in the data set, FIML is the preferred estimation method, as it uses all of the data available (e.g., see Jöreskog & Moustaki, 2012). Due to the nature of the data, particularly the NA responses (i.e., zero variance items, empty cells), maximum likelihood methods of estimation and weighted least squares estimation failed for most participants and was unfortunately infeasible in the work presented here.

Strengths. As with the measurement results presented, one of the most important strengths of this factor analytic work is the generalizability of the results, allowed by the two longitudinal data sets examined. Despite differences in sample characteristics (e.g., older adults compared to younger women), item sets (e.g., 38 items, most from the PANAS and Circumplex Model of Emotion compared to 16 items not taken from any measure in particular), and time scales (e.g., daily measurements for eight weeks compared to ten randomly timed measurements a day for five days), differences in affect factor structure and factor correlations between participants were much larger than differences between occasions in both data sets. Thus, it is very likely these findings would be replicated by future studies with a variety of samples of participants, items, and measurement periods similar to the ones studied here.

With these individual differences results, the replication across different sets of items is particularly impressive. A large portion of the affect literature involves constructing measures intended to capture various quadrants of a circumplex of affect adjectives (e.g., Russell, 1980; Watson et al., 1988). The individual differences found in the present work, in two different item sets, indicates there are likely substantial individual differences in affect factor structure and factor correlations in many of the item sets commonly borrowed from the affect circumplex. Future research must refrain from assuming affect factor structure and factor correlations are the same across individuals as within individuals.

Future Research. Much research remains to be done on individual differences in affect factor structure and factor correlations. First, more stringent confirmatory factor

analysis models could be used to more rigorously test for individual differences in various components of factor models by employing multilevel methods, such as the Idiographic Filter (Nesselroade et al., 2007). For example, freeing loadings to vary by person and then additionally freeing factor correlations to vary by person will provide insight into whether individual differences in both of these factor model components exist simultaneously.

Unfortunately, applying these methods to ordinal data is computationally taxing and for many models may be infeasible with current statistical software. This raises another area in need of further research: the implementation of multilevel factor models with ordinal data containing missing values. In IRMs, estimation of parameters from ordinal data is extremely fast due to the sparse block-diagonal Hessian matrix used for optimization (Baker & Kim, 2004). At each iteration, each block of the Hessian can be separately updated, saving massive amounts of computational time. If it is possible to make similar constraints on the Hessian in factor models, research exploring methods of doing so would greatly contribute to future individual differences work with ordinal data.

Finally, much more work is needed on explaining why and how the factor structure of affect and correlations of affect factors differ between individuals. Finding groups of individuals with very similar factor structures and factor correlations and determining what other characteristics these individuals have in common would be a good beginning to this explanation.

Chapter 10: Final Statements

A New Framework for Affect Research

In this dissertation work, the validity of four assumptions commonly made in affect research was examined in two different longitudinal data sets. Results from item response model analyses refuted the two measurement assumptions tested, revealing that PA and NA measures do not have the same desirable measurement properties and PA and NA constructs are not captured adequately by the same Likert-type response scales. Specifically, items and response scales commonly used in measuring NA do not target participants well, do not have stable psychometric properties over time, and do not produce accurate estimates of NA levels. Thus, it is critical that future affect research employs different, better methods for measuring affect, particularly NA.

Similarly, results from exploratory factor analyses refuted the two individual differences assumptions tests, indicating the factor structure of affect and affect factor correlations differ substantially between individuals, but not between occasions. These individual differences were found in both longitudinal data sets. The false assumption that these individual differences do not exist has contributed to debates in the affect literature spanning decades, such as the debate regarding the bipolarity of PA and NA. Continuing these debates based on results obtained under assumptions that are clearly invalid substantially hinders scientific progress in affect research. To prevent further slowing of affective science advances, future research must acknowledge the presence of these individual differences and account for them as much as practically possible.

In sum, this work highlights the need for a new framework under which to conduct affect research. Under this framework, researchers must explore new methods of more adequately measuring and modeling affect and its influences on health and well-being.

First, new methods of measuring affect, particularly NA, are needed. Investigations of a variety of response scales would inform these new measurement methods. Similarly, exploring new types of affect items, ones that take context into account or alter the frame of reference from which individuals are evaluating affect, would also greatly contribute to developing affect measures that produce sufficiently accurate quantifications of PA and NA. Additionally, studies employing a wider variety of measurement evaluation methods, using both factor analytic techniques and IRT techniques, rather than continuing to report factor analytic results alone, would be informative.

Second, investigations of affect factor structure, correlations among affect factors, and associations between affect and other constructs must acknowledge the presence of individual differences in affect factor structure, and in the relationship among dimensions of affect. While multilevel modeling studies will go a long way in informing future investigations, less sophisticated methods will also greatly contribute to the quality of future research. In cross-sectional studies, researchers can investigate whether group differences exist in any findings involving affect. For example, in a study exploring PA as a mediator of the influence of NA on health behaviors, researchers could check for group differences in the direct and indirect effects found and in the relationship between

PA and NA. Groups could be created from demographic variables, personality assessments that were included in the study, or any other variables measured.

Additionally, when generalizing results of a study, it is important to remember the evidence supporting individual differences in affect factor structure and factor correlations. The relationships between PA, NA, and other constructs may also differ between individuals, limiting the generalizability of empirical findings.

The work completed in this dissertation project supports the need for a new framework for conducting affect research, a framework in which measurement and individual differences are major foci. Developing and testing new methods of measuring and modeling affect will go a long way in advancing the study of affect and improving the rigor with which affect research is conducted. This new framework will allow researchers to better uncover the role of affect in health and well-being across the lifespan.

References

Antoni, M. H., LaPerriere, A., Schneiderman, N., & Fletcher, M. A. (1991). Stress and immunity in individuals at risk for AIDS. *Stress Medicine, 7,* 35-44.

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any "threshold disorder controversy". *Educational and Psychological Measurement*, *73*(1), 78-124.

Baker, F. B., & Kim, S-H. (2004). *Item response theory parameter estimation techniques* (2nd ed.). Boca Raton, FL: CRC Press.

Beck, J. G., Novy, D. M., Diefenbach, G. J., Stanley, M. A., Averill, P. M., & Swann, A. C. (2003). Differentiating anxiety and depression in older adults with Generalized Anxiety Disorder. *Psychological Assessment, 15*(2), 184-192.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison–Wesley.

Bond, T. G. and Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences. Psychology Press.

Carstensen, L. L., Isaacowitz, D. M., & Charles, S. T. (1999). Taking time seriously: A theory of socioemotional selectivity. *American Psychologist, 54,* 165-181.

Charles, S.T., Mather, M., & Carstensen, L.L. (2003). Focusing on the positive: Age differences in memory for positive, negative, and neutral stimuli. *Journal of Experimental Psychology, 85,* 163–178.

Coifman, K. G., Bonanno, G. A., & Rafaeli, E. (2007). Affect dynamics, bereavement and resilience to loss. *Journal of Happiness Studies, 8*, 371-392.

Crawford, J. R., & Henry, J. D. (2004). The positive and negative affect schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology, 43,* 245-265.

Crocker, P. R. E. (1997). A confirmatory factor analysis of the Positive and Negative Affect Schedule (PANAS) with a youth sport sample. *Journal of Sport & Exercise Psychology, 19,* 91-97.

Davis, M. C., Zautra, A. J., & Smith, B. W. (2004). Chronic pain, stress, and the dynamics of affective differentiation. *Journal of Personality, 72*(6), 1133-1160.

Dua, J. K. (1993). The role of negative affect and positive affect in stress, depression, self-esteem, assertiveness, type A behaviors, psychological health, and physical health. *Genetic, Social & General Psychology Monographs, 119*(4).

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495-515.

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

Erbacher, M. K., Schmidt, K. M., & Bergeman, C. S., (Under review). An idiographic approach using derivatives for studying relationships among longitudinal measurements of positive and negative affect in later life.

Erbacher, M. K., Schmidt, K. M., Boker, S. M., & Bergeman, C. S., (2012). Measuring positive and negative affect in older adults over 56 days: Comparing trait level scoring methods using the Partial Credit Model. *Journal of Applied Measurement*.

Erbacher, M. K., Schmidt, K. M., & Schroeder, J. R. (In prep). The role of positive affect in the experience of chronic pain.

Ersner-Hershfield, H., Mikels, J. A., Sullivan, S. J., & Carstensen, L. L. (2008). Poignancy: Mixed emotional experience in the face of meaningful endings. *Journal of Personality and Social Psychology, 94*(1), 158-167.

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology, 69*(1), 153-166.

Feldman-Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of affect. *Journal of Personality and Social Psychology, 74,* 967–984.

Feldman-Barrett, L. F., & Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science, 8,* 10–14.

Ferrucci, L., Harris, T. B., Guralnik, J. M., Tracy, R. P., Corti, M. C., & Cohen, H. J., … Havlik, R. J. (1999). Serum IL-6 level and the development of disability in older persons. *Journal of the American Geriatrics Society, 47*(6), 639-646.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491.

Gable, S. L., Reis, H.T, & Elliot, A. J. (2000). Behavioral activation and inhibition in everyday life. *Journal of Personality and Social Psychology, 78,* 1135–1149.

Gaudreau, P., Sanchez, X., and Blondin, J. (2006). Positive and negative affect states in a performance-related setting: Testing the factorial structure of the PANAS across two samples of French-Canadian participants. *European Journal of Psychological Assessment, 22*(4), 240-249.

Goldstein, M. D., & Strube, M. J. (1994). Independence revisited: The relation between positive and negative affect in a naturalistic setting. *Personality and Social Psychology Bulletin, 20,* 57–64.

Green, D. P., Salovey, P., & Truax, K. M. (1999). Static, dynamic, and causative bipolarity of affect. *Journal of Personality and Social Psychology, 76,* 856–867.

Green, K. E. & Frantom, C. G. (2002). *Survey development and validation with the Rasch model.* Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC, November.

Horn, J. L. (1965). A rationale and a test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hu, J., & Gruber, K. J. (2008). Positive and negative affect and health functioning indicators among older adults with chronic illnesses. *Issues in Mental Health and Nursing, 29*, 895-911.

Jackson, B. R., & Bergeman, C. S. (2011). How does religiosity enhance well-being? The role of perceived control. *Psychology of Religion and Spirituality, 3*(2), 139-161.

Jannarone, R. J. (2010). Models for locally dependent responses: Conjunctive item response theory. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern Item Response Theory* (465-479). New York: Springer.

Joiner, T. E., Sandin, B., Chorot, P., Lostao, L., & Marquina, G. (1997). Development and factor analytic validation of the SPANAS among women in Spain: (More) cross-cultural convergence in the structure of mood. *Journal of Personality Assessment, 68*(3), 600-615.

Jöreskog, K. G., & Moustaki, I. Factor analysis of ordinal variables with full information maximum likelihood. Unpublished manuscript downloaded from http://www.ssicentral.com. Accessed August, 2012.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136-153.

Kawata, A. K. (2006). Measurement invariance of the Positive and Negative Affect Schedule (PANAS) in a community sample of older people. *Dissertation Abstracts International B: the Sciences and Engineering, 66*(9), 5136.

Kercher, K. (1992). Assessing well-being in the old-old: The PANAS as a measure of orthogonal dimensions of positive and negative affect. *Research on Aging, 14*(2), 131-168.

Kiecolt-Glaser, J. K., McGuire, L., Robles, T. F., & Glaser, R. (2002). Psychoneuroimmunology: Psychological influences on immune function and health. *Journal of Consulting and Clinical Psychology, 70*(3), 537-547.

Kratz, A. L., Davis, M. C., & Zautra, A. J. (2007). Pain acceptance moderates the relation between pain and negative affect in female osteoarthritis and fibromyalgia patients. *Annals of Behavioral Medicine, 33*(3), 291-301.

Labouvie-Vief, G., & Medler, S. M. (2002). Affect optimization and affect complexity: Modes and styles of regulation in adulthood. *Psychology and Aging, 17,* 571–587.

Labouvie-Vief, G., Diehl, M., Jain, E., Zhang, F. (2007). Six-year change in affect optimization and affect complexity across the adult lifespan: A further examination. *Psychology and Aging, 22*(4), 738-751.

Lamoureux, E. L., Pesudova, K., Pallant, J. F., Rees, G., Hassell, J. B., Caudle, L. E., & Keeffe, J. E. (2008). An evaluation of the 10-item Vision Core Measure 1 (VCM1) Scale (the core module of the Vision-Relation Quality of Life Scale) using Rasch analysis. *Ophthalmic Epidemiology, 15*, 224-233.

Larsen, R. J., & Diener, E. (1992). Promises and problems with the Circumplex Model of Emotion. In M. S. Clark (Ed.), *Emotion*. Thousand Oakes, CA: Sage Publications.

Lebo, M. A. & Nesselroade, J. R. (1978). Intra-individual differences dimensions of mood change during pregnancy identified in five *p*-technique factor analyses. *Journal of Research in Personality,* 205-224.

Leue, A. & Beauducel, A. (2011). The PANAS structure revisited: On the validity of a bifactor model in community and forensic samples. *Psychological Assessment, 23*(1), 215-225.

Likert, R. (1931). A technique for the measurement of attitudes. *Archives of Psychology*. New York: Columbia University Press.

Linacre, J. M. (2005). Dichotomous and polytomous category information. *Rasch Measurement Transactions*, *19*(1), 1005-1006.

Linacre, J. M. (2012). Winsteps Manual. www.winsteps.com.

Lundgren-Nilsson, A., Grimby, G., Ring, H., Tesio, L., Lawton, G., Slade, A., ... Tennant, A. (2005). Cross-cultural validity of functional independence measure items in stroke: A study using Rasch analysis. *Journal of Rehabilitation Medicine, 37*, 23-31.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174.

Mather, M., Canli, T., English, T., Whitfield, S., Wais, P., Ochsner, K., Gabrieli, J.D.E., & Carstensen, L.L. (2004). Amygdala responses to emotionally valenced stimuli in older and younger adults. *Psychological Science, 15,* 259–263.

Mather, M., & Carstensen, L.L. (2003). Aging and attentional biases for emotional faces. *Psychological Science, 14,* 409–415.

Matsunaga, M., Isowa, T., Kimura, K., Miyakoshi, M., Kanayama, N., Murakami, H., … Ohira, H. (2008). Associations among central nervous, endocrine, and immune activities when positive emotions are elicited by looking at a favorite person. *Brain, Behavior, and Immunity, 22,* 408-417.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*(4), 201-218.

Molloy, G. N., Pallant, J. F., & Kantas, A. (2001). A psychometric comparison of the Positive and Negative Affect Schedule across age and sex. *Psychological Reports, 88*, 861-862.

Montpetit, M. A., Bergeman, C. S., Deboeck, P. R., Tiberio, S. S., & Boker, S. M. (2010). Resilience-as-process: Negative affect, stress, and coupled dynamical systems. *Psychology and Aging, 25*(3), 631-640.

Muthén, L. K.., & Muthén, B. O. (20012). Mplus User's Guide (Seventh Edition). Los Angeles, CA: Muthén & Muthén.

Nesselroade, J. R., Gerstorf, D., Hardy, S. A., & Ram, N. (2007). Idiographic filters for psychological constructs. *Measurement, 5*(4), 217-235.

Nesselroade, J. R., & Schmidt-McCollam, K. M. (2000). Putting the process in developmental processes. *International Journal of Behavioral Development, 24*(3), 295-300.

Ong., A. D., Bergeman, C. S., & Bisconti, T. L. (2004). The role of daily positive emotions during conjugal bereavement. *Journal of Gerontology: PSYCHOLOGICAL SCIENCES, 59B*(4), P168-P176.

Ong., A. D., Bergeman, C. S., Bisconti, T. L., & Wallace, K. A. (2006). Psychological resilience, positive emotions, and successful adaptation to stress in later life. *Journal of Personality and Social Psychology, 91*(4), 730-749.

Ong, A. D., Zautra, A. J., & Carrington-Reid, M. (2010). Psychological resilience predicts decreases in pain catastrophizing through positive emotions. *Psychology and Aging, 25*(3), 516-523.

Potter, P. T., Zautra, A. J., & Reich, J. W. (2000). Stressful events and information processing dispositions moderate the relationship between positive and negative affect: Implications for pain patients. *Annals of Behavioral Medicine, 22,* 1002–1012.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rafeali, E., Rogers, G. M., & Revelle, W. (2007). Affective synchrony: Individual differences in mixed emotions. *Personality and Social Psychology Bulletin, 33*(7), 915-932.

Raykov, T., & Marcoulides, G. A. (2008). *An Introduction to Applied Multivariate Analysis.* New York: Taylor & Francis.

Ready, R. E., Carvalho, J. O., & Weinberger, M. I. (2008). Emotion complexity in younger, midlife, and older adults. *Psychology and Aging, 23*(4), 928-933.

Ready, R. E., Robinson, M. D., & Weinberger, M. (2006). Age differences in the organization of emotion knowledge: Effects involving valence and time frame. *Psychology and Aging, 21*(4), 726-736.

Reich, J. W. & Zautra, A. J. (2002). Arousal and the relationship between positive and negative affect: An analysis of the data of Ito, Cacioppo, and Lang (1998). *Motivation and Emotion, 26*(3), 209-222.

Reich, J. W., Zautra, A. J., & Potter, P. T. (2001). Cognitive structure and the independence of positive and negative affect. *Journal of Social and Clinical Psychology, 20,* 99–115.

Revelle, W. (2012). psych: Procedures for Personality and Psychological Research. Northwestern University, Evanston. R package version 1.2.8.

Robazza, C., Bortoli, L., Nocini, F., Moser, G., Arslan, C. (2000). Normative and idiosyncratic measures of positive and negative affect in sport. *Psychology of Sport and Exercise, 1*, 103-116.

Roth, R. S., Geisser, M. E., Theisen-Goodvich, M., & Dixon, P. J. (2005). Cognitive complaints are associated with depression, fatigue, female sex, and pain catastrophizing in patients with chronic pain. *Archives of Physical Medicine and Rehabilitation, 86,* 1147-1154.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125,* 3–30.

Russel, A., Bergeman, C. S., Deboeck, P., Baird, B., Monpetit, M., & Ong, A. (2011). Emotional control during later life: The relationship between global perceptions and daily experience. *Personality and Individual Differences, 50*, 1084-1088.

Schmidt, K. M. (2006). *Calibrating a Multidimensional CAT for Chronic Pain Assessment.* Technical Report for Barron Associates, Inc.

Smith, R. M., Schumacker, R. E., and Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*(1), 66-78.

Staud, R., Price, D. D., Robinson, M. E., & Vierck, C. J. (2004). Body pain area and pain-related negative affect predict clinical pain intensity in patients with fibromyalgia. *The Journal of Pain, 5*(6), 338-343.

Stone, M. H. (1998). Rating scale categories: Dichotomy, double dichotomy, and the number two. *Popular Measurement, 1*(1), 61-65.

Strand, E. B., Zautra, A. J., Thoresen, M., Odegard, S., Uhlig, T., & Finset, A. (2006). Positive affect as a factor of resilience in the pain-negative affect relationship in patients with rheumatoid arthritis. *Journal of Psychosomatic Research, 60*, 477-484.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics*. Boston: Allyn and Bacon.

Tellegen, A., Watson, D., & Clark, L. A. (1999). On the dimensional and hierarchical structure of affect. *Psychological Science, 10*(4), 297-303.

Terracciano, A., McCrae, R. R., Hagemann, D., & Costa, P. T. Jr. (2003). Individual difference variables, affective differentiation, and the structures of affect. *Journal of Personality, 71*(5), 669-704.

Thayer, R.E. (1989). *The biopsychology of mood and activation.* New York: Oxford University Press.

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Zautra, A. J., Johnson, L. M., & Davis, M. C. (2005). Positive affect as a source of resilience for women in chronic pain. *Journal of Consulting and Clinical Psychology, 73*(2), 212-220.

Zautra, A. J., Potter, P. T., & Reich, J. W. (1997). The independence of affects is context dependent: An integrative model of the relationship between positive and negative affect. In K. W. Schaie & M. Powell (Eds.), *Annual Review of Gerontology* (Vol. 17, pp. 75–103). New York: Springer.

Zautra, A. J., Reich, J. W., Davis, M. C., Nicolson, N. A., & Potter, P. T. (2000). The role of stressful events in the relationship between positive and negative affects: Evidence from field and experimental studies. *Journal of Personality, 68,* 927–951.

Zautra, A., Smith, B., Affleck, G., & Tennen, H. (2001). Examinations of chronic pain and affect relationships: Applications of a dynamic model of affect. *Journal of Consulting and Clinical Psychology, 69,* 786–795.

Zevon, M. A. & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology, 43*(1), 111-122.

Zhang, X., Ersner-Hershfield, H., & Fung, H. H. (2010). Age differences in poignancy: Cognitive reappraisal as a moderator. *Psychology & Aging, 25*(2) 310-320.

Zhang, Z., Browne, M. W., & Nesselroade, J. R. (2011). Higher order factor invariance and idiographic mapping of constructs to observables. *Applied Developmental Sciences, 15*(4), 186-200.

Appendix A: Collapsed Response Scales

```
# R Code for Obtaining All Possible Collapsed Response Scales
# Note: Function Originated from Kathy Gerber, MS (University of
# Virginia)

f <- function (n=10)
{
     ret <- matrix(0, nrow=2^n, ncol=1+n)
    x <- 0:(2^n-1)
    for(j in (n+1):2) {
       ret[,j] <- x%%2
       x <- x %/% 2
    }
    if (n>=2) for(j in 2:(n+1)) {
       ret[,j] <- ret[,j-1] + ret[,j]
    }
    ret
}

# FOR NDSHWB SCALES:
# Give the function one less than the number of categories in the
# original scale

coll.scales5 <- f(n=4)

# By default, categories begin at 0. Add 1 to each element of
# coll.scales5

coll.scales5 <- coll.scales5 + 1

# Remove the first row (all 1's)
# NDSHWB SCALES: 15 Collapsed Patterns

coll.scales5[-1,]

[,1] [,2] [,3] [,4] [,5]
 [1,]    1    1    1    1    2
 [2,]    1    1    1    2    2
 [3,]    1    1    1    2    3
 [4,]    1    1    2    2    2
 [5,]    1    1    2    2    3
 [6,]    1    1    2    3    3
 [7,]    1    1    2    3    4
 [8,]    1    2    2    2    2
 [9,]    1    2    2    2    3
[10,]    1    2    2    3    3
[11,]    1    2    2    3    4
[12,]    1    2    3    3    3
[13,]    1    2    3    3    4
[14,]    1    2    3    4    4
[15,]    1    2    3    4    5
```

```
# FOR MAASTRICHT SCALES:
# Give the function one less than the number of categories in the
# original scale

coll.scales7 <- f(n=6)

# By default, categories begin at 0. Add 1 to each element of
# coll.scales7

coll.scales7 <- coll.scales + 1

# Remove the first row (all 1's)
# MAASTRICHT SCALES: 63 Collapsed Patterns

coll.scales7[-1,]

[,1] [,2] [,3] [,4] [,5] [,6] [,7]
 [1,]    1    1    1    1    1    1    2
 [2,]    1    1    1    1    1    2    2
 [3,]    1    1    1    1    1    2    3
 [4,]    1    1    1    1    2    2    2
 [5,]    1    1    1    1    2    2    3
 [6,]    1    1    1    1    2    3    3
 [7,]    1    1    1    1    2    3    4
 [8,]    1    1    1    2    2    2    2
 [9,]    1    1    1    2    2    2    3
[10,]    1    1    1    2    2    3    3
[11,]    1    1    1    2    2    3    4
[12,]    1    1    1    2    3    3    3
[13,]    1    1    1    2    3    3    4
[14,]    1    1    1    2    3    4    4
[15,]    1    1    1    2    3    4    5
[16,]    1    1    2    2    2    2    2
[17,]    1    1    2    2    2    2    3
[18,]    1    1    2    2    2    3    3
[19,]    1    1    2    2    2    3    4
[20,]    1    1    2    2    3    3    3
[21,]    1    1    2    2    3    3    4
[22,]    1    1    2    2    3    4    4
[23,]    1    1    2    2    3    4    5
[24,]    1    1    2    3    3    3    3
[25,]    1    1    2    3    3    3    4
[26,]    1    1    2    3    3    4    4
[27,]    1    1    2    3    3    4    5
[28,]    1    1    2    3    4    4    4
[29,]    1    1    2    3    4    4    5
[30,]    1    1    2    3    4    5    5
[31,]    1    1    2    3    4    5    6
[32,]    1    2    2    2    2    2    2
[33,]    1    2    2    2    2    2    3
[34,]    1    2    2    2    2    3    3
[35,]    1    2    2    2    2    3    4
[36,]    1    2    2    2    3    3    3
[37,]    1    2    2    2    3    3    4
```

```
[38,]    1    2    2    2    3    4    4
[39,]    1    2    2    2    3    4    5
[40,]    1    2    2    3    3    3    3
[41,]    1    2    2    3    3    3    4
[42,]    1    2    2    3    3    4    4
[43,]    1    2    2    3    3    4    5
[44,]    1    2    2    3    4    4    4
[45,]    1    2    2    3    4    4    5
[46,]    1    2    2    3    4    5    5
[47,]    1    2    2    3    4    5    6
[48,]    1    2    3    3    3    3    3
[49,]    1    2    3    3    3    3    4
[50,]    1    2    3    3    3    4    4
[51,]    1    2    3    3    3    4    5
[52,]    1    2    3    3    4    4    4
[53,]    1    2    3    3    4    4    5
[54,]    1    2    3    3    4    5    5
[55,]    1    2    3    3    4    5    6
[56,]    1    2    3    4    4    4    4
[57,]    1    2    3    4    4    4    5
[58,]    1    2    3    4    4    5    5
[59,]    1    2    3    4    4    5    6
[60,]    1    2    3    4    5    5    5
[61,]    1    2    3    4    5    5    6
[62,]    1    2    3    4    5    6    6
[63,]    1    2    3    4    5    6    7
```