Modeling Extreme Response Style Using Item Response Trees

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Arts in the Graduate School of The Ohio State University

By

Adam Tapal, Mgr.

Graduate Program in Psychology

The Ohio State University

2016

Master's Examination Committee:

Paulus De Boeck, Advisor

Michael C. Edwards

Duane Wegener
Abstract

Responses to Likert scales are potentially influenced by response styles and not only by the latent trait level or attitude strength to be measured. This thesis focuses on the extreme response style (ERS) in responses to a Big Five personality test. An attempt is presented to jointly model the latent personality trait and ERS using an extension of Item Response Theory – the Item Response Trees, such that the response style effect can be removed. In order to do so, it is assumed that the ERS effect is independent of the latent trait. This attempt consists of two separate studies that are identical in the used methods but focus on two different Big Five datasets.
Acknowledgments

I thank my advisor, Paul De Boeck, for his valuable expertise, advice and support.
Vita

June 2012 .......................................................Bc. Psychology, Masaryk University
June 2015 .......................................................Mgr. Psychology, Masaryk University
August 2015 to August 2016 .........................Graduate Fellow, Department of Psychology, The Ohio State University
October 2015 to May 2016 ............................Graduate Research Assistant, Office of Institutional Research and Planning, The Ohio State University
August 2016 to present .................................Graduate Teaching Associate, Department of Psychology, The Ohio State University

Publications


Fields of Study

Major Field: Psychology
Table of Contents

Abstract ............................................................................................................................... ii

Acknowledgments .......................................................................................................... iii

Vita ................................................................................................................................... iv

List of Tables .................................................................................................................. viii

List of Figures ................................................................................................................. xi

Introduction ....................................................................................................................... 1

Psychological and Sociodemographic Correlates of ERS ........................................... 6

Theoretical Explanations of ERS ................................................................................... 8

Measuring Extreme Response style .............................................................................. 10

  Traditional approaches ............................................................................................... 10

  Model-based approaches ......................................................................................... 13

Overview of Item Response Theory ............................................................................ 22

  IRTree models ......................................................................................................... 28

Study 1 ............................................................................................................................. 37

  Sample and data ...................................................................................................... 37

  Method ...................................................................................................................... 38
List of Tables

Table 1. The mapping matrix representing the original responses as sub-responses ...... 29
Table 2. The mapping matrix for the example of a nested model. ......................... 31
Table 3. The mapping matrix for Study 1 ................................................................. 41
Table 4. The discrimination matrices of sub-items for Study 1 (all IRTree models) ...... 46
Table 5. Model fit of the proposed and alternative models (Agreeableness). .......... 48
Table 6. EAP reliabilities of latent trait estimates (Agreeableness). ....................... 48
Table 7. Summary of discrimination parameter estimates at each node (Agreeableness).49
Table 8. Model fit of the proposed and alternative models (Extraversion). .......... 50
Table 9. EAP reliabilities of latent trait estimates (Extraversion). ......................... 50
Table 10. Summary of discrimination parameter estimates at each node (Extraversion).51
Table 11. Model fit of the proposed and alternative models (Conscientiousness). ... 52
Table 12. EAP reliabilities of latent trait estimates (Conscientiousness). ............... 52
Table 13. Summary of discrimination parameter estimates at each node (Conscientiousness).53
Table 14. Model fit of the proposed and alternative models (Neuroticism). .......... 54
Table 15. EAP reliabilities of latent trait estimates (Neuroticism). ....................... 54
Table 16. Summary of discrimination parameter estimates at each node (Neuroticism).55
Table 17. Model fit of the proposed and alternative models (Openness). ............. 56
Table 18. EAP reliabilities of latent trait estimates (Openness). ......................... 56
Table 19. Summary of discrimination parameter estimates at each node (Openness). .... 57
Table 20. Covariances of the exogenous variables in the structural model, 2D model. .... 61
Table 21. Covariances of the personality trait scores in the structural model, 3D model. 63
Table 22. Covariances of the response style score residuals in the structural model, 3D model. 63
Table 23. The mapping matrix for Study 2. ................................................................. 72
Table 24. The discrimination matrices of sub-items for Study 2 (all IRTree models). .... 75
Table 25. Model fit of the proposed and alternative models (Agreeableness). ........ 77
Table 26. EAP reliabilities of latent trait estimates (Agreeableness). .................... 77
Table 27. Summary of discrimination parameter estimates at each node (Agreeableness). ........................................................ 78
Table 28. Model fit of the proposed and alternative models (Extraversion). ........ 79
Table 29. EAP reliabilities of latent trait estimates (Extraversion). ..................... 79
Table 30. Summary of discrimination parameter estimates at each node (Extraversion). 80
Table 31. Model fit of the proposed and alternative models (Conscientiousness). .... 80
Table 32. EAP reliabilities of latent trait estimates (Conscientiousness). ............ 81
Table 33. Summary of discrimination parameter estimates at each node (Conscientiousness). .......................................................... 81
Table 34. Model fit of the proposed and alternative models (Neuroticism). ........ 82
Table 35. EAP reliabilities of latent trait estimates (Neuroticism). .................. 82
Table 36. Summary of discrimination parameter estimates at each node (Neuroticism). 83
Table 37. Model fit of the proposed and alternative models (Openness). .............. 83
Table 38. EAP reliabilities of latent trait estimates (Openness) .................................. 84
Table 39. Summary of discrimination parameter estimates at each node (Openness) .... 84
Table 40. Covariances of the exogenous variables in the structural model, 2D model.... 86
Table 41. Covariances of the personality trait scores in the structural model, 3D model. 88
Table 42. Covariances of the response style score residuals in the structural model,

3D model.................................................................................................................. 90
List of Figures

Figure 1. Rasch model item response functions for different item difficulty values........ 26
Figure 2. 2PL item response functions for different discrimination parameter values..... 26
Figure 3. Example of an IRTree. ...................................................................................... 28
Figure 4. Example of a nested IRTree model.. ................................................................. 31
Figure 5. The depiction of the tree structure for Study 1................................................ 40
Figure 6. Structural model of the EAP estimates from Study 1, 2D model....................... 60
Figure 7. Structural model of the EAP estimates from Study 1, 3D model....................... 62
Figure 8. The depiction of the tree structure for Study 2................................................ 71
Figure 9. Structural model of the EAP estimates from Study 2, 2D model....................... 87
Figure 10. Structural model of the EAP estimates from Study 2, 3D model..................... 89
Introduction

In psychology as in other social science disciplines, rating scales such as Likert scales are widely popular for measurement of attitudes, opinions, personality characteristics or personal preferences. Such scales usually have a simple format where the person being questioned indicates the degree of their agreement with a given statement or to what degree it describes them. This degree of agreement or identification with a statement is expressed by endorsing one of a number of ordered response categories, the number of categories typically being the same for all items within a scale and sometimes containing a neutral or ambivalent response (e.g., 1 = “Disagree”, 2 = “Somewhat disagree”, 3 = “Neither agree nor disagree”, 4 = “Somewhat agree”, 5 = “Agree”).

Although popular and widely used, rating scales are not devoid of problems associated with their use. It is generally assumed that when answering on a scale, the choice is driven solely by the respondent’s trait level or attitude strength. However, the responses to scale items can also be influenced by systematic factors called “response styles”. A response style can be defined as a content-irrelevant, consistent pattern of responding or the tendency to prefer or avoid certain response categories (Hamilton, 1968; Wetzel & Carstensen, 2015).

A number of response styles have been investigated in the past years. Among the ones more elaborately studied are acquiescence (ARS, tendency to agree or identify with
a statement, regardless of content), disacquiescence (DRS, the opposite of acquiescence),
midpoint responding (MRS, tendency to use the middle response category, regardless of
content) and extreme response style (ERS, tendency to endorse the most extreme
response categories, regardless of content), as noted by multiple authors (Baumgartner &
Steenkamp, 2001; Bachman & O’Malley, 1984; Khorramdel & Davier, 2014; Bolt &
Johnson, 2009). Extreme response style is – aside from acquiescence – probably the one
that has gained the most attention in the past as well as more recently, and it will be the
primary subject of interest in this thesis.

As noted previously, extreme response style can be described as the tendency of
respondents to favor or avoid choosing the extreme end-points of a rating scale, relatively
independently of content (De Jong, Steenkamp, Fox & Baumgartner, 2008). This
response style can be responsible for biasing effects on both the correlations between
constructs and on the level of scores and thus on comparisons between different persons
(Greenleaf, 1992). Consider, for example, the situation when two persons respond to a
scale designed to measure trait aggressiveness. If, as is quite usual, it is assumed that
purely the person’s trait aggressiveness is underlying the responses to the scale and thus a
sum score is obtained on the basis of a unidimensional model, the two persons can be
compared simply with respect to the difference in their sum scores. However, if one also
assumes that extreme response style is present and that it influences the responses
alongside trait aggressiveness, this procedure might no longer be valid. If the two
respondents are equally low on the trait aggressiveness, but one of them exhibits strong
ERS, his or her sum score might be lower due to persistently choosing the more extreme
low category on the response scale. The other person might, on the other hand, avoid extreme responses and end up with a sum score that is considerably higher simply because moderately worded responses were chosen instead. It could, thus, be concluded that the extreme responder’s trait level is lower than that of the moderate responder, even though their true latent trait level would be the same.

Furthermore, as noted previously, extreme response style can possibly be a biasing factor in estimating the true correlations between two constructs. The variance of raw scores obtained from responses to a scale can be decomposed into true score variance and error variance, which can itself consist of random and/or systematic variance. Assuming that ERS is stable and generalizes over multiple instruments (a notion which will be elaborated on later), it can be a source of systematic error variance which – unless ERS is controlled for – is inseparable from true score variance and might inflate, deflate or otherwise confound the covariance between two true scores.

Therefore, extreme response style can pose a threat to the validity of psychological measurement and, consequently, to the validity of psychological findings. Nevertheless, as Baumgartner and Steenkamp (2001) note, it is not frequently considered a major threat in the literature (if the method sections of empirical and scale development articles are to serve as a reference), although its potentially confounding effects on statistical inference have been noted by a number of researchers (Bachman & O’Malley, 1984; Eid & Rauber, 2000; Naemi, Beal & Payne, 2008; among others). According to Naemi et al. (2008), extreme responders account to 25% - 30% of total respondents in most surveys and questionnaires. Bachman and O’Malley (1984) found that 48% of items
in one survey exhibited the pattern of having a greater percentage of extreme responses 
(on both ends) made by one ethnic group compared to the other – an observation 
attributed to the difference in extreme response style between the two groups (the topic of 
cross-cultural research on ERS will be briefly discussed later on). These are only a few 
selected examples of the breadth of challenges extreme responding can pose.

On the other hand, not all research up to date has viewed ERS as a mere statistical 
nuisance that has to be controlled for and of which the scores of the construct of interest 
have to be purified. Instead, it has been approached as a phenomenon worth studying by 
itself. Extreme responding has been found by some researchers to be quite stable over 
time and to generalize across different instruments. Bachman and O’Malley (1984) 
reported relatively high stability of ERS over a period of up to four years and Weijters, 
Geuens and Schillewaert (2010) have found it to be stable over at least a one-year period. 
Similarly, Berg and Collier (1953) found test-retest reliability of the ERS measure 
employed by them approximately .80 over a period of 15 days and a similar result was 
reported by Merrens (1970), although over a shorter time span. A significant number of 
authors have noted the consistency of ERS across instruments designed to measure 
different constructs (for example, Hamilton, 1968; Merrens, 1970; Greenleaf, 1992; 
Wiley & Wiley, 1970; Austin, Deary & Egan, 2006), supporting the notion that ERS 
generalizes over instruments and is not scale-specific. It is fair to point out that other 
authors, however, did not find ERS to be stable and doubted the validity of viewing it as 
a systematic response style (Innes, 1977; Hui & Triandis, 1985). Differing, contradictory 
findings and lack of consensus regarding extreme response style might have been the
result of years of research during which no uniform way of measuring it was broadly established in the scientific community and various researchers and research teams used various methods (not unusually their very own) of assessing ERS (Greenleaf, 1992). Measurement problems such as this one will be more broadly addressed in a stand-alone section to follow.

Naturally, once the stability and generality of ERS seemed demonstrated (at least to some researchers), it became plausible to consider extreme responding as a manifestation of certain personality characteristics, of stable individual differences between persons. However, extreme response style (and other response styles, by that matter) is most likely best understood as not being of a purely dispositional nature, but rather as the product of an interaction between dispositional and situational determinants. Responders differ in their tendency to engage in extreme responding, and this tendency might be bolstered or weakened by situational factors such as the format of the response scale, time pressure, perceived importance of the scale or questionnaire or the content of items.

That being said, an extensive amount of findings have been accumulated in the past decades concerning personality correlates of extreme response style, as literature was more concerned with the dispositional understanding of ERS. While theoretical explanation of extreme responding is not the main purpose of this thesis, such findings are valuable and helpful for a contextual understanding of this response style as a meaningful psychological construct. In the next section, the conclusions of some of these inquiries into ERS are briefly presented, as well as attempts to understand the
psychological processes behind extreme responding.

Psychological and Sociodemographic Correlates of ERS

As already noted, the tendency to provide extreme responses can be viewed as a manifestation of some aspect of personality. With stability and generality supported, a number of researchers have tried to explore the relationships of ERS and some well-known or relevant personality and sociodemographic variables. However, McGee (1962) points out the insufficient representation of behavioral data in the research on psychological correlates of ERS – a limitation which persists to this day. The research up to date can be classified into three broad categories based on the main methodology: *correlational, criterion-based* and *factor-analytic* (Hamilton, 1968). The following part summarizes the main findings from each of these categories of research.

Correlational studies

Correlational evidence mostly covers the relationship of ERS and variables measured by pre-existing and well-established scales and tests. Of considerable interest were epistemic and cognitive dispositions – for example, ERS was reported to correlate positively with intolerance of ambiguity (or rigidity; Brim & Hoff, 1957; Brengelmann, 1959, 1960; Naemi, Beal & Payne, 2009) and negatively with conceptual complexity (White & Harvey, 1965), while mixed evidence was reported for a relationship with authoritarianism or intelligence (Hamilton, 1968; Naemi, Beal & Payne, 2009). Among emotional variables, extreme responding was mainly reported to correlate positively with anxiety (Lewis & Taylor, 1955). Other personality traits of interest included Big Five
traits, and the findings of such studies are of high relevance to this thesis. ERS was found to be positively correlated with extraversion (Austin, Deary & Egan, 2006; Meiser & Machunsky, 2008) and conscientiousness (Austin, Deary & Egan, 2006), but no significant relationships with other personality traits in the Big Five framework have been reported. A plethora of other correlates, mainly those concerned with mental health and adjustment, was summarized by O’Donovan (1965).

Criterion-based studies

Studies in this category were generally focused on comparing two or more groups with respect to the differences in their tendency for extreme responding. More often than not, females were found to exhibit a stronger ERS; as did children and elderly (Hamilton, 1968; Naemi, Beal & Payne, 2009; Austin, Deary & Egan, 2006; Eid & Rauber, 2000). Groups with lower achieved education were also reported to be more prone towards extreme responding (Marin, Gamba & Marin, 1992) and differences between groups of varying manifest anxiety levels seem to support the relationship reported above (Berg & Collier, 1953).

A specific category of studies that investigated group membership and extreme response style is constituted by those that focused on religious and ethnic minorities or cross-cultural differences in general. It is difficult to summarize such research effectively, mainly because of the sheer variety of different cultural groups and minorities that were investigated with respect to differences in ERS in the past decades. Nonetheless, cross-cultural differences have been repeatedly demonstrated in different contexts by multiple

Factor-analytic studies

Some researchers have tried to explore the structure of correlations between ERS and other personality variables by the means of factor analysis of summed scores obtained from a diverse set of scales, and scores or indices of extreme responding. Two studies reported conflicting findings regarding the relationship between ERS and authoritarianism – Zuckerman and Norton (1961) found them loading in an opposing way on a single factor, whereas Schutz and Foster (1963) reported otherwise. The latter authors also found extreme response style to positively load on a factor dubbed “inflexibility”, supporting the above mentioned link between ERS and rigidity. Hamilton (1968) also mentions a factor-analytic study which linked extreme responding to a motivation akin to the need for cognitive closure (Webster & Kruglanski, 1994).

Theoretical Explanations of ERS

Although no consensual explanation exists of why individuals differ in their tendency to respond extremely, there have been attempts at making sense of the rather fragmented findings reported above. However, none of these attempts represent the result of systematic effort to uncover the underpinnings of extreme responding – rather, they should be understood as hypotheses or ideas produced as a result of a review of findings. The first proposition by Brim and Hoff (1957) is centered on the concepts of intolerance for ambiguity and rigidity. Since not only measures of intolerance for
ambiguity, but also both authoritarianism and dogmatism scales have been linked to ERS (although not always conclusively), and both are considered related to intolerance of ambiguity (Hamilton, 1968; Leary & Hoyle, 2009), this idea seems to have support in empirical findings. The main explanation lies in the fact that responding in the extremes might serve a function of reducing ambiguity and epistemic tension, introducing structure and clarity.

Another hypothesis, mentioned by Hamilton (1968) as well as Baumgartner and Steenkamp (2001), reflects on the association between ERS and anxiety. From this point of view, extreme responses are due to a greater intensity of responses to stimuli (in this case, items) by persons with higher anxiety level, accompanied by decreased ability to inhibit such responses.

A final explanation presented here is focused on the development of cognitive structures (Shulman, 1973). Based on this explanation, persons with less differentiated cognitive structures are more likely to respond in an extreme way simply due to a decreased ability to discriminate between, for example, various nuances of agreeing (or disagreeing) with a statement. Such an idea was prompted by some of the findings listed above, such as greater extremity of responses in children and elderly, as well as lower conceptual complexity among extreme responders. Somewhat hinting towards this hypothesis is also the reasoning of Hui and Triandis (1989) who speculate that the number of subjective categories of judgment might simply be different for some persons from the number of response categories provided by a response scale. If this assumption is accepted, one can hypothesize that persons with less nuanced cognitive schemas might
tend towards using only the extreme end-points on a scale due to a lack of more fine-grained subjective judgment categories.

Measuring Extreme Response style

Over the course of the past decades, different approaches to measuring ERS have emerged in the scientific community. According to some (e.g., Greenleaf, 1992), the almost overwhelming variability in how ERS is operationalized and measured is one of the reasons why many contradictory reports about its stability, generality and about the relationship to other psychological and sociodemographic variables exist. This state of affairs continues up to the present time. Approaches to study extreme responding can be categorized into two distinct groups – the so-called “traditional” approaches, prevalent since the 1950s and still widely popular because of their simplicity, and rather recent, “model-based” approaches that have begun to sprout early in the 21st century. In the following section, a brief review of both approaches is presented.

Traditional approaches

The most classical and popular approach to quantifying ERS is measuring it as the proportion of times a respondent answers a given number of rating scale items in an extreme interval (Greenleaf, 1992). An estimate of ERS is, under this approach, often simply constructed by summing up the number of extreme responses a respondent endorsed (for example, the number of times a respondent “strongly agreed” or “strongly disagreed”) over a set of items and dividing this number by the total number of items the
respondent answered. This technique is usually employed when five- or seven-point scales are used. In cases when scales with a greater number of response categories are used, often the two most extreme categories on each end are collapsed and both count as an extreme response (Hamilton, 1968).

Although this frequency-based method was (and still is) the most prevalent one in the domain of the “traditional” methods, it is by no means the only one. Other proposed ways of quantifying extreme responding under the traditional approach include computing the standard deviation of respondent’s item scores (the more dispersed person’s scores are, the more extreme of a responder a person is; Bolt & Newton, 2011) or calculating a “bimodality index” (Peak, Muney & Clay, 1960 in Hamilton, 1968) which is the average number of extreme responses minus the average number of middle-category responses.

Two different ways of employing the traditional approach can be further distinguished based on which items are included in the set used to compute said proportion or index – researchers either choose the relevant items ad-hoc from substantive scales used in their research, or use dedicated ERS instruments (De Jong, Steenkamp, Fox & Baumgartner, 2008). Both ways, however, suffer from inherent limitations. The use of ad-hoc ERS scales, first of all, means that more often than not, items used for measuring ERS have rarely been used for the same purpose before, which poses a threat to the validity of this method as well as making between-studies comparisons difficult. Furthermore, it is common that the items chosen for measuring ERS all originate from a single substantive scale – among the examples from literature
are the Dogmatism scale by Rokeach (1956), the F-Scale (Adorno, Fenkel-Brunswik, Levinson & Sanford, 1950) or parental attitudes scale; as noted by Hamilton (1968) or Greenleaf (1992). This is highly problematic, because it is impossible to determine whether an extreme response was provided for trait-related reasons (respondent might have chosen the extreme response simply because he or she indeed identifies or agrees strongly with a given statement), stylistic reasons (the extreme response was endorsed due to respondent’s tendency to endorse extreme responses) or a synergy of both. Difficulty to separate stylistic and substantive variance is of course at the very heart of measuring ERS, but this problem becomes exacerbated, since items from such scales are commonly highly correlated.

The use of dedicated ERS measures was originally conceived as an attempt to ameliorate this issue by assessing ERS across items heterogeneous in content and “psychologically diffuse” (Greenleaf, 1992; Baumgartner & Steenkamp, 2001). The logic behind this approach is that while respondents might extremely agree or disagree with some items for substantive reasons, it is unlikely that they will do so consistently across items of different content, which should only be a strategy of extreme responders. In addition, measures thought to be “content-free” have been used to measure ERS for the same reason under the assumption that extreme response style will be the most pronounced when item content is the most ambiguous or absent. Among these content-free measures are, for example, Berg’s (1953) Perceptual Reaction Test or the Personal Friends Questionnaire by Soueif (1958).
However, not even dedicated ERS measures have eluded criticism (Baumgartner & Steenkamp, 2001; Jin & Wang, 2014; Khorramdel & Davier, 2014). Their addition into a survey or test battery makes it longer and is thus costly in terms of respondent fatigue. Using even dedicated ERS instruments under the traditional approach does not reflect the interactionist view of response styles, that is, it doesn’t reflect the idea that not only persons differ in their tendency for extreme responses, but items, too, differ in how they prompt extreme responding. Persons in general respond to some items more extremely than to others, and an extreme response to some items might be more informative of a person’s ERS than an extreme response to some other items. Furthermore, the usefulness of an item for measuring ERS might vary across different groups – a limitation which the traditional approach cannot cope with. On top of all this, a major limitation of the traditional methods is that they mostly offer ways of quantifying ERS while not correcting for its influence (Bolt & Johnson, 2009).

Recently, attempts to overcome abovementioned problems with ERS measurement and correction for extreme responding have resulted in incorporating ERS into psychometric models based on item response theory (IRT). A general overview and notable examples of the model-based approaches will be presented in the following section.

Model-based approaches

Current work on IRT-based models for extreme response style can be grouped into two distinct approaches (Wetzel & Carstensen, 2015). Some researchers have
considered ERS as a categorical variable so that persons differ qualitatively in their extreme response style (that is, a person either is or is not an extreme responder, although this approach can be extended to accommodate more than two distinctions) – this represents the so-called *categorical (or discrete) approach*. Other researchers view ERS as a continuous variable with persons differing quantitatively in the degree or strength of extreme responding – this represents the so-called *dimensional approach* and is conceptually closer to the traditional approaches presented earlier.

The *categorical approach* has been dominated by the use of mixture item response theory models (Bolt & Johnson, 2009), most commonly the mixed partial credit model. In these models, different latent classes are identified for which different item parameters apply – typically two classes that are distinguished by the supposed presence or absence of extreme response style. In other words, such models allow for the existence of sub-groups within the population of interest which possess different item and person parameter estimates (for further details on IRT nomenclature, see the Method section below). It is important to note that this approach is different from the so-called differential item functioning (DIF) analysis in that DIF analyses are concerned with differences in item parameters between manifest groups (i.e., groups that differ with respect to a manifest categorical variable such as ethnicity, gender or level of education), whereas mixture models are concerned with identifying groups that do not systematically differ as such.

The categorical approach has mostly been fundamentally exploratory in nature, since determining whether the difference between two (or more) latent classes lies in the
presence or absence of response style is done by carefully examining and interpreting the
differences in threshold parameter estimates. As Jin and Wang (2014) note, this approach
should be used with caution, since it may not always reveal true latent classes and might
lead to identification of spurious latent classes instead. Nevertheless, studies using
mixture models seem to be quite successful in identifying latent classes that differ in
extreme response style (Eid & Rauber, 2000; Gollwitzer, Eid & Jürgensen, 2005; Austin,
Deary & Egan, 2006) and class membership has been reported to be consistent across
different scales (Wetzel, Carstensen & Böhnke, 2013). An important discussion remains,
however, on whether it is reasonable to simplify the problem of extreme response style
by reducing its nature to that of a dichotomous variable.

Unlike the categorical approach, the *dimensional approach* is more conceptually
similar to the traditional approaches described earlier in that response style is viewed and
modeled as a continuous variable. In this sense, each individual is considered to possess a
unique location on a latent continuum that represents intensity or strength of tendency for
extreme responding. Modeling and measurement efforts under the umbrella of
dimensional approach are also considerably more diverse in terms of concepts and
techniques, ranging from a use of pre-existing or classical IRT models for purposes of
ERS measurement to formulation of specialized models with response styles specifically
in mind.

For instance, De Jong, Steenkamp, Fox and Baumgartner (2008) have used a two-
parameter IRT model to measure extreme response style. They recoded responses to
items as either extreme (1) or not extreme (0) and the model was fit on this recoded data.
Although this approach might be useful for measuring ERS, it does not allow for measuring the target latent trait and only focuses on the response style. Bolt and Johnson (2009) have used a multidimensional nominal response model (MNRM, Takane & de Leeuw, 1987, in Falk & Cai, 2016) to explore potential presence of ERS dimension in categorical responses to items from a tobacco dependence scale. In their study, a model with two independent latent dimensions exhibited the best fit among a number of models explored. Working subsequently with the two-dimensional model, the category slope estimates were rotated to maximize their variance for the first latent dimension. After the rotation, the category slope estimates were inspected and while the slope estimates were strictly ordered (the lowest slope for response category “1” and the highest slope for response category “7”) for the first latent dimension, suggesting that it corresponds to tobacco dependence, they were not ordered so for the second dimension. In fact, in the case of the second dimension, the only positive slope estimates were those of the most extreme response categories, while the slope estimates of the non-extreme categories were all negative and about equal in size. Based on this pattern, the authors concluded the second dimension corresponds to a trait distinguishing extreme and non-extreme responses. This approach is in some respects superior to that of De Jong et al. (2008) because the substantive dimension of interest is being modeled alongside the response style dimensions which allows for not only measuring ERS, but also correcting the substantive scores for response style bias. It is important to note, however, that this approach is rather exploratory in nature, since the model was not explicitly formulated to measure ERS.
Their method was further extended by Bolt and Newton (2011) for multi-scale measurement of ERS using MNRM, however, with significant adjustments. Unlike in the case of Bolt and Johnson (2009), the number of latent dimensions was chosen a priori and the category slope parameters for each latent dimension were constrained. In case of substantive dimensions, slope parameters were constrained to -3, -1, 1 and 3 for the categories of 1, 2, 3 and 4, respectively, however for the extreme response style dimension, slope parameters were constrained to 1, -1, -1 and 1 instead. These additional constraints represent a shift from the exploratory approach taken by Bolt and Johnson (2009) to an approach that is inherently confirmatory. Wetzel and Carstensen (2015) have confronted the issue in a similar manner, but used a multidimensional partial credit model (Kelderman, 1996) rather than MNRM and instead of constraining slope parameters, different scoring functions were applied to the response categories. Because the aim of their study was to demonstrate modeling of different response styles, the scoring functions were adjusted accordingly. For example, for acquiescence, the two categories indicating the highest agreement were scored as 1 while other categories were scored as 0. Analogously, in case of extreme response style, the two most extreme categories were scored as 1 while other categories were scored as 0.

Finally, Falk and Cai (2016) have introduced flexible, full-information modeling of response styles using MNRM. Under this approach, it is possible to model multiple response styles of choice (not only extreme response style). The model proposed by the authors does not require constraining the category slope parameters as in the case of Bolt & Johnson (2009) in order to identify which latent dimension is a substantive dimension.
and which is a response style dimension (and of what type) – instead, an item slope parameter is freely estimated for each item and latent dimension modeled. This is enabled by defining each latent dimension through an a priori specified category scoring function, akin to the approach chosen by Wetzel and Carstensen (2015). Each response category can be assigned a different score which represents the order of the categories for the item and how the categories relate to any given modeled latent dimension. For example, a scoring function of 0, 1, 2, 3 and 4 can be used to represent the order of five categories for the measurement of a substantive trait, a scoring function of 1, 0, 0, 0 and 1 can be used to represent the order of the categories for the measurement of ERS, and so forth. This allows for specifying the order of categories while allowing for item slope parameters to be estimated for each item and latent dimension – enabling the researcher to evaluate the loading of each item on a particular latent variable.

The approaches just described share two main features. First of all, they represent instances of using IRT models in resourceful new ways to measure extreme response style. Secondly, they can both incorporate ERS into the measurement model as a unique latent variable. By contrast, different paths were taken by Jin and Wang (2014) and Johnson (2003) in which ERS is no longer modeled as a latent variable. Johnson (2003) proposes a heterogeneous threshold ordinal regression model in which item category thresholds are treated as random across persons. The logic behind this approach is straightforward – the distances between category thresholds for extreme responders should be significantly smaller than those for non-extreme responders. In other words, it should take only a relatively smaller increase in agreement (or disagreement) for an
An extreme responder to endorse a response category indicating stronger agreement (or disagreement) over a category indicating weaker agreement (or disagreement). Jin and Wang’s (2014) IRT model, on the other hand, keeps item category thresholds homogeneous across persons, however, an additional person parameter – the weight parameter $\omega$ – serves as a multiplier of the item category threshold parameters. Large weight parameters indicate larger distances between thresholds so that extreme responses are less likely to be endorsed – a logic similar to Johnson’s. The value of this weight parameter varies across persons and quantifies a person’s tendency for extreme responding.

A third category of dimensional approaches that will be presented here could be considered highly similar to the first category in that it represents a use of pre-existing IRT models and incorporates ERS into the measurement model as an additional latent dimension. However, the models falling into this category are highly relevant for this thesis because they represent a like-minded approach. Recently, Khorramdel and von Davier (2014) and Plieninger and Meiser (2014) made use of item response trees (De Boeck & Partchev, 2012; Böckenholt, 2012) to model the response processes behind respondents’ choosing of one response category over another. This approach aims to effectively decompose the observed responses into multiple binary sub-responses (which represent said response processes) on the basis of a hypothesized processing tree schema. The resulting sub-responses (responses to pseudo-items which are the products of decomposition of original items) can subsequently be fitted with common IRT models.
Because the research reported in this thesis uses the item response trees approach, it is worth describing the studies conducted by the authors in greater detail. Khorramdel and von Davier (2014) decomposed the original ordinal response items in a Big Five questionnaire into binary pseudo-items based on a processing tree which assumes three sequential processes. The first process represents the respondent’s decision whether a non-neutral or neutral (midpoint category) response will be chosen (Process I). If a non-neutral response was chosen, the second process (Process II) represents the choice between a positive or negative response. Subsequently, the third process (Process III) embodies the choice between an extreme or a non-extreme response. Using this schema, each original item was decomposed into three pseudo-items corresponding to the postulated processes, and a multidimensional two-parameter logistic model was fit on the decomposed data. Process I was assumed to model neutral responding, Process II was assumed to model the personality trait measured by the original item, and Process III was assumed to model extreme responding. Utilizing all items from the Big Five measure they were using, the authors aimed at simultaneously modeling extreme response style, neutral responding and all the Big Five personality traits based on the pseudo-item responses. The authors demonstrated usefulness of this approach in modeling extreme and midpoint responding and concluded that the used approach is suitable for measuring and correcting for the effect of response styles.

Similarly, Plieninger and Meiser (2014) have also employed the tree approach in an almost identical way, although with a number of important differences. The original ordinal response items were decomposed on the basis of a tree model which assumed four
sequential processes and not three as was the case in the study by Khorramdel and von Davier (2014). The fourth additional process modeled the choice between categories that were neither neutral nor extreme (i.e., a choice between “1” and “2” or “4” and “5” on the “0” to “6” scale used by the authors). Additionally, the sub-responses were fitted with a one-parameter logistic model. The authors’ conclusions were similar to the ones made by Khorramdel and von Davier (2014) and, furthermore, an effort was made to address concerns about the validity of this approach in modeling response styles. The authors have used extraneous criteria to assess the construct and criterion validity of their model and reported satisfactory validity evidence.

Because of the promising results of research on modeling extreme response style using a tree-based approach, and due to the relative flexibility it allows for, item response trees will be the modeling framework of choice in this thesis. I aim to expand on the work done by Khorramdel and von Davier (2014) and Plieninger and Meiser (2014) in a number of ways. First of all, the approach used in both cited studies assumes that responses to pseudo-items are strictly unidimensional. This constraint means that the personality trait supposedly measured by the original item is only modeled through responses to pseudo-items corresponding to Process II (the process modeling either disagreement or agreement with the item), making the actual intensity of agreement or disagreement as measured with the original response scale irrelevant for measuring anything else aside from ERS. In other words, the respondent’s choice of a response category of “5” over “4” on a 5-point Likert scale is assumed to be only due to extreme response style and not due to the trait in question. This is a fairly strong assumption that
does not take into account the possibility that both the personality trait and ERS can influence the intensity of a response. Secondly, extreme responses (represented by Process III) are grouped together irrespective of the direction of a response which assumes that a respondent – not taking into account their personality trait level – is just as likely to respond with an extreme agreement as they are with extreme disagreement. Finally, in the case of Plieninger and Meiser (2014), only the most extreme responses (“0” and “6”) are considered to reflect extreme response style, although one could argue that for 7-point response scales, the two most extreme responses on either side might be indicative of ERS – with the most extreme responses being simply stronger indicators (Falk & Cai, 2016).

In this thesis, I aim to carry out research which does not impose these assumptions. Following this section is a concise, basic overview of item response theory in general, which then makes way for introducing item response trees as formulated by De Boeck and Partchev (2012) and Jeon and De Boeck (2015), followed by specification of the chosen approach.

Overview of Item Response Theory

The term Item Response Theory (IRT) denotes a class of psychometric models centered on the notion that the probability of a person’s response to an item (usually dichotomous, however, multiple extensions to polychotomous items exist) is a function of his or her location on some latent trait continuum. Such “latent trait” is a construct that is not directly measurable, but it is assumed to cause or influence a person’s response or
behavior (such as answering a test item correctly or endorsing a statement). Examples of such constructs are personality traits (such as optimism, extraversion or need for cognition), mental abilities (analytical reasoning, verbal ability or abstract thinking, among others), but also behavioral tendencies or attitudes. From an IRT perspective, persons differ in how strong their traits or how good their abilities are in such way that they can be ordered with respect to their latent trait levels – that is, persons differ in a quantitative rather than qualitative manner.

As mentioned before, a person’s latent variable level is not directly measurable and thus has to be estimated from information that can be obtained and measured directly – from responses to a set of items that are believed to be related to the latent variable of interest. Items, like persons, are positioned on the same latent continuum. The item’s position on the latent continuum is usually interpreted as its difficulty or severity and, just as with persons, items can be ordered with respect to this property. The concept of item difficulty can be understood in a simple manner within the IRT context – the greater a person’s trait level is compared to the item’s difficulty, the more likely a person is to answer that item correctly or to endorse it. Conversely, as a person’s trait level decreases with respect to the item’s location, the probability of answering correctly or endorsing the item decreases as well.

In order to more specifically introduce Item Response Theory, two concise sections below are dedicated to an overview of two frequently used IRT models that are the most relevant to the content of this thesis – the classic Rasch model (Rasch, 1960)
and the two-parameter logistic model (Birnbaum, 1968). A brief introduction of a multidimensional variant of the two-parameter logistic model is also provided.

**Rasch model**

The simplest and most straightforward of Item Response models is the Rasch model. It is used to estimate latent variable levels when available data come from responses to items that are dichotomous. The Rasch model can be expressed with the following formula:

\[
P(Y_{ij} = 1 \mid \theta_i) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}
\]

where \(Y_{ij}\) is the response of person \(i\) to item \(j\) (for simplicity, let us assume that “1” denotes a correct response and “0” denotes an incorrect response), \(\theta_i\) is the latent trait level of person \(i\) and \(\beta_j\) is the difficulty (or severity) of item \(j\). Simply put, the equation expresses the probability of a person with trait level \(\theta\) answering an item of difficulty \(\beta\) correctly.

In Figure 1 below, a plot of this function (a so-called “item response function” or “item characteristic curve”) for different values of \(\beta\) is provided for illustration purposes. As can be seen, the probability of a correct response monotonically increases with the level of latent trait (indicated as “Ability” in the plot), and the probability is equal to 0.5 whenever \(\beta = \theta\).

**Two-parameter logistic model (2PL)**
The two-parameter logistic model is, just as in the case of the Rasch model, used in cases when data are obtained from dichotomously scored items. We can express the model using the formula below:

\[
P(Y_{ij} = 1 \mid \theta_i) = \frac{\exp(\alpha_j(\theta_i - \beta_j))}{1 + \exp(\alpha_j(\theta_i - \beta_j))}
\]  

(2.2)

where \(Y_{ij}\) is the response of person \(i\) to item \(j\), \(\theta_i\) is the latent trait level of person \(i\) and \(\beta_j\) is the difficulty (or severity) of item \(j\). The difference between the Rasch and the 2PL models lies in the presence of the so-called discrimination parameter, \(\alpha_j\). This parameter can vary over items and represents how well a given item differentiates between persons of different latent trait levels. As can be seen, higher values of the discrimination parameter cause the difference between \(\theta\) and \(\beta\) to more strongly affect the response probabilities.

Another way of looking at the discrimination parameter is understanding it in the same way as item loadings are understood in the context of factor analysis. Both ways of thinking about the discrimination parameter imply an important principle – the larger the discrimination parameter, the more information a given item yields about the latent trait in question, in other words, the more relevant the item is for measuring the latent trait. In Figure 2 below, item response functions for different values of the discrimination parameter (but of the same difficulty value, \(\beta = 0\)) are shown. As can be seen, the larger discrimination parameters the greater the slope of the function, which is especially pronounced around the area where \(\theta = \beta\).
Figure 1. Rasch model item response functions for different item difficulty values. The figure shows three item response functions (indicated by numbers 1, 2 and 3) for different item difficulty values ($\beta_1 = 0, \beta_2 = -1, \beta_3 = 1$).

Figure 2. 2PL item response functions for different discrimination parameter values. The figure shows three item response functions (indicated by numbers 1, 2 and 3) for different discrimination parameter values ($\alpha_1 = 0.5, \alpha_2 = 1, \alpha_3 = 2$). Note that the item difficulty parameter is the same for all functions, $\beta = 0$. 
Naturally, it is obvious that the Rasch model is a special case of the two-parameter model, one for which all items’ discrimination parameters are equal. Other than this difference, the parameters of both models should be interpreted in the same manner.

**Multidimensional two-parameter logistic model**

The models presented thus far were both modeling a person’s response as a function of a single latent trait parameter. Indeed, unidimensionality is one of the assumptions of these models. Item Response models that relax this assumption do however exist and are frequently used. Such models extend the idea of a latent variable continuum into the concept of a multidimensional latent space. The multidimensional variety of the two-parameter logistic model that was previously covered can be expressed with the following equation:

\[
P(Y_{ij} = 1 \mid \theta_i) = \frac{\exp(\alpha_j^\prime \theta_i - d_j)}{1 + \exp(\alpha_j^\prime \theta_i - d_j)}
\]

where \( \theta_i \) is no longer a single parameter, but a vector of length \( m \), corresponding to \( m \) latent dimensions. The same is true for \( \alpha_j \), which means that each item has a dimension-specific discrimination parameter. The discrimination parameters can, in this case, be understood in the same way as discrimination parameters in the unidimensional two-parameter model – indicating the item’s loading on a given latent trait (or the amount of information the item can provide about the latent trait). The \( d_j \) parameter is an intercept which plays a similar role as the \( \beta_j \) parameter in unidimensional models.
IRTree models

In the studies that constitute this thesis, I have utilized IRT models with a tree structure (so-called IRTrees; de Boeck & Partchev, 2012; Jeon & de Boeck, 2015). An IRTree model represents the observed categorical responses of person i to item j as outcomes of a series of dichotomous responses $Y_{ijn}^*$ (responses to so-called “sub-items” or “nodes” with stemming branches representing the two responses, forming a tree structure). The tree structure is based on a number of postulated internal processes (represented by the sub-items) that are hypothesized to underlie the polytomous response, such that the final response is the result of progressing through the tree to its bottom. An example of such a tree is provided in Figure 3.

Figure 3. Example of an IRTree.

*Originally polytomous categorical responses (0, 1 or 2) are represented as outcomes of a series of dichotomous sub-responses $Y_{ij1}^*$ and $Y_{ij2}^*$. 

28
As can be seen from the figure above, each of the three original responses can now be formulated in terms of the dichotomous sub-responses. The way in which the original responses can be recoded is shown in Table 1 in what is termed the model’s *mapping matrix*. The purpose of the mapping matrix is to clearly delineate how the final answer is arrived at in terms of a series of sub-responses and can be thought of as the matrix representation of the model schematic as shown in Figure 3.

<table>
<thead>
<tr>
<th>( Y_{ij} )</th>
<th>( Y_{ij1}^* )</th>
<th>( Y_{ij2}^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_{ij} = 0 )</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>( Y_{ij} = 1 )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( Y_{ij} = 2 )</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. The mapping matrix representing the original responses as sub-responses. The original response of \( Y_{ij} = 1 \) is recoded as responses \( Y_{ij1}^* = 1 \) (sub-response 1 or Node 1 of item \( j \)) and \( Y_{ij2}^* = 0 \) (sub-response 2 or Node 2 of item \( j \)). Since the response \( Y_{ij} = 0 \) is already reached through the response \( Y_{ij1}^* = 0 \), the response \( Y_{ij2}^* \) on sub-item 2 is missing (NA).

An interesting interpretation for the sub-items (or “nodes”) is that they represent the processes of arriving at the final response on the original response scale of 0, 1 or 2.

In order to illustrate this logic, we can consider the item \( j \) to be “*Would you buy a Volkswagen car?*” and the original responses to stand for “*No*”, “*Maybe*” and “*Yes*”, respectively. At first, the respondent would decide whether it is at all likely that he or she would be willing to buy a Volkswagen – this decision is represented by the first sub-item. If the purchase of a Volkswagen would be ruled out straight away, they would arrive at...
the answer “No”. If the purchase would not be ruled out straight away, they would decide on how likely the purchase would be – whether they are somewhat (“Maybe”) or absolutely (“Yes”) certain. This process is represented by the second sub-item. Therefore, the ultimate answer is conceived as consisting of first granting any likelihood whatsoever to the purchase of a Volkswagen, and then considering how likely the purchase would really be.

The tree model shown in Figure 3 represents one variety of IRTree models, a so-called linear tree. For linear trees, at least one branch from any node leads directly to a final response. In case of the example presented, the response of “0” to the first sub-item directly leads to a final answer of “no” ($Y_{ij} = 0$) while the response of “1” leads to another node. Some IRTree models, however, can contain nodes that have both branches leading to another node, and thus any response to such nodes would lead to an eventual response only indirectly, through intermediate nodes. An example of such a tree is provided in Figure 4.
Figure 4. Example of a nested IRTree model. Either response at the top node leads to one of the lower-level nodes, and only responses at those lower-level sub-items ($Y_{ij2}^*$ or $Y_{ij3}^*$) lead to the final response $Y_{ij}$.

Just as with the linear tree model presented in Figure 3, the nested model has a mapping matrix (see Table 2).

<table>
<thead>
<tr>
<th></th>
<th>$Y_{ij1}^*$</th>
<th>$Y_{ij2}^*$</th>
<th>$Y_{ij3}^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_{ij} = 0$</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>$Y_{ij} = 1$</td>
<td>0</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>$Y_{ij} = 2$</td>
<td>1</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>$Y_{ij} = 3$</td>
<td>1</td>
<td>NA</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. The mapping matrix for the example of a nested model. As before, the original responses (0 through 3) are recoded as a series of responses to the respective sub-items. The responses to sub-items which are not reached on the way to the final response through the tree are missing (NA).
The nested tree model can be illustrated in a similar way as the linear tree model. Consider again that the item $j$ would be “Would you buy a Volkswagen car?” and the response options (0 to 3 in Figure 4) would stand for “Absolutely no”, “Probably no”, “Probably yes” and “Absolutely yes”. At first, the respondent would (perhaps heuristically) decide whether, say, the likelihood of them buying a Volkswagen is below or above 50%. The response of “below 50%” would be represented as the response of “0” at the first node and would lead to the second node (denoted as sub-item 2 in Figure 4), while the response of “above 50%” would be represented as the response of “1” and would lead to the third node (designated as sub-item 3 in Figure 4). At those lower nodes, the decision would be dealing with the further determining the likelihood of the purchase. At Node 2, for example, the respondent would decide whether the likelihood is 0% or close to 0% (response of “0”, arriving at the final response of “Absolutely no”) or whether it is higher than that (but still less than 50%), arriving at the final response of “Probably no”.

Put in other words, the data are expanded such that each item $j$ is “broken down” into $n$ dichotomous sub-items corresponding to each node on the postulated “tree” and every response category of the original item $j$ is represented as a unique sequence of sub-responses to those $n$ sub-items ($Y_{ij} = 0$ as [0,NA], $Y_{ij} = 1$ as [1,0] and $Y_{ij} = 2$ as [1,1]). Unless some response categories are merged or considered identical, the number $n$ of nodes is always $k-1$ where $k$ is the number of response categories (this rule does not hold for cases where the sub-items are not binary, see Jeon & De Boeck, 2015). As can be seen, due to the way the tree model is constructed, some sub-responses are coded as not
observed (NA) since the corresponding pseudo-items were not reached by the respondent along the way to the final response. This can be classified as data missing at random (MAR, Rubin & Little, 2002) since the missingness depends on previous observations and thus it does not pose a problem for model estimation using maximum likelihood. The probabilities of responding “0” or “1” at each node (i.e., the probabilities of following the left or right branch) can be modeled using an IRT model. For simplicity, let’s consider modeling the sub-responses with a Rasch model. At each node, the probability of a response of “1” vs. a response of “0” is modeled with the following function:

$$P(Y_{ijn}^* = 1 \mid \theta_{in}, v_{ij}) = \frac{\exp(\theta_{in} - \beta_{jn})}{1 + \exp(\theta_{in} - \beta_{jn})}$$

where $Y_{ijn}$ is the response of person $I$ to sub-item $n$ belonging to item $j$, $\theta_{in}$ is the propensity of person $I$ to answer “1” at node $n$ (i.e., the propensity to follow the right branch at node $n$), $\beta_{jn}$ can be interpreted as the intensity of left-branch induction of node $n$ belonging to item $j$ (i.e., the strength with which node $n$ of item $j$ induces a response of “0”) and $v_{ij}$ is the vector of responses of person $I$ to sub-items belonging to item $j$ that have had to be necessarily provided in order for sub-item $n$ to be reached (i.e., the probability of a response of $Y_{ij3}^*$ from the example given in Figure 4 is conditional on the preceding response of $Y_{ij1}^*$ being “1”) – for the first sub-item, $v_{ij}$ does not contain any elements.

Based on the model as shown in Figure 3, the probabilities of all the final responses $Y_{ij}$ can be expressed as follows:

$$P(Y_{ij} = 0 \mid \theta_{i}) = P(Y_{ij1}^* = 0 \mid \theta_{i1})$$

33
\[ P(Y_{ij} = 1 \mid \theta_i) = P(Y_{ij1}^* = 1 \mid \theta_{i1}) P(Y_{ij2}^* = 0 \mid Y_{ij1}^* = 1, \theta_{i2}) \]  
\[ P(Y_{ij} = 2 \mid \theta_i) = P(Y_{ij1}^* = 1 \mid \theta_{i1}) P(Y_{ij2}^* = 1 \mid Y_{ij1}^* = 1, \theta_{i2}) \]

where \( \theta_i = (\theta_{i1}, \theta_{i2}) \), \( \theta_i \sim MVN(0, \Sigma_\theta) \) and \( \beta_j = (\beta_{j1}, \beta_{j2}) \). The probability of an end-node response can be simply calculated as the product of (conditional) probabilities of sub-responses that lead to it (as can be seen, the model can be easily formulated for any number of nodes). The multiplicative nature of the IRTree models as just described is rooted in the fact that the probabilities of all but the first sub-responses are conditioned on preceding sub-responses, which satisfies the assumption of local independence. This important aspect refers to so-called discrete survival models (see De Boeck and Partchev, 2012). These models are thus fundamentally different from the so-called \textit{divide-by-total} models of choice behavior (Bradley & Terry, 1952; Luce, 1959) in line with the nominal response model (NRM), often used for modeling response styles (see the \textit{Model-based approaches} section of the theoretical introduction above). The main difference is that while the probability of the final response is expressed in the IRTree framework as the product of (conditional) probabilities of sub-responses leading to that final response, the same probability is expressed in \textit{divide-by-total} models as the value of the final response divided by the sum of the values of all potential final responses.

It is important to note that the above formulae and descriptions of the IRTree model present both the propensity and the induction intensity as dependent on the node. This quality of IRTree models means that they allow for different latent variables for each sub-item and also for different item parameters for each sub-item created from the same original item. Such formulation, however, can be subject to model constraints on
either of those properties – a latent variable can be assumed to be measured on multiple nodes, for instance. Conversely, this enables answering substantive research questions relating to the meaning of the response categories and/or the underlying response processes – a quality that is shared with the multidimensional variants of divide-by-total models such as the multidimensional nominal response model.

**Extension of the IRTree models**

The type of models presented above and discussed by De Boeck and Partchev (2012) have an important limitation – they rely on a one-parameter logistic (or probit) model for modeling the responses to sub-items. However, with a certain set of assumptions, a more complex modeling function can be incorporated into the general tree model (see Jeon & De Boeck, 2015).

Although we have previously considered modeling the sub-responses with a Rasch model, there is no reason we couldn’t model them using a two-parameter logistic model instead. In such a case, the probability at each node of a response of “1” vs. a response of “0” could be modeled using the following function:

\[
P(Y_{ijn}^* = 1 \mid \theta_{in}, \nu_{ij}) = \frac{\exp(\alpha_{jn}(\theta_{in} - \beta_{jn}))}{1 + \exp(\alpha_{jn}(\theta_{in} - \beta_{jn}))}
\]  

where \(\alpha_{jn}\) is the discrimination parameter for node \(n\) belonging to item \(j\) and all other parameters are as described previously. As introduced in the 2PL section above, the discrimination parameter can be thought of as the strength with which the latent trait (or response propensity) \(\theta\) influences the probability of responding “1” (or “0”) to the given
sub-item. In other words – how relevant the latent trait is in determining the particular sub-response.

Similarly, the tree models need not be limited to one-dimensional IRT models for formulating the probability of a response to a sub-item. Just as described in the section on multidimensional IRT models above, the probabilities of sub-responses can be modeled as functions of, for example, two person parameters:

\[
P(Y_{jn}^* = 1 \mid \theta_{in1}, \theta_{in2}, \nu_{ij}) = \frac{\exp(\alpha_{jn1}\theta_{in1} + \alpha_{jn2}\theta_{in2} + d_{jn})}{1 + \exp(\alpha_{jn1}\theta_{in1} + \alpha_{jn2}\theta_{in2} + d_{jn})}
\]  

(2.9)

where \(\alpha_{jn1}\) is the discrimination parameter for the latent trait \(\theta_{in1}\), \(\alpha_{jn2}\) is the discrimination parameter for the latent trait \(\theta_{in2}\) and \(d_{jn}\) is the intercept, akin to the difficulty (or induction intensity) \(\beta_{jn}\) in the previously shown models. The item parameters and person parameters are again presented as node-specific, however, this need not be the case if additional constraints are imposed on the model. In order for the model to be identified, it is necessary to impose some constraints on the latent distribution. Thus, the distribution of the latent traits is constrained to be multivariate normal with means of 0. Additionally, the variance-covariance matrix of the latent traits is constrained, with diagonal elements set to 1 and off-diagonal elements corresponding to latent traits measured on the same sub-items set to 0.

For the purposes of this thesis, the probabilities of sub-responses were modeled with a multidimensional two-parameter logistic model with the aim to model them as functions of sub-item difficulty parameters and multiple latent trait parameters with node-by-trait-specific item discrimination parameters. A set of latent dimensions is formulated
and a specific subset of these dimensions is declared for each sub-item beforehand (details on this formulation are provided in the respective studies below). As mentioned earlier, unlike the research previously carried out (Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014), the specification used here does not assume that extreme responses are only due to ERS and not at all due to the substantive personality trait being measured – both latent dimensions are assumed to influence the extreme responses, as will be explained later on. Instead, the relative contributions of the two latent dimensions can be evaluated through comparing their respective sub-item-specific discrimination parameters. For each of the following studies, the depiction of the proposed model tree structure and the rationale behind it is provided, as well as the mapping matrix which represents each original response category as a series of sub-responses. Moreover, the matrix indicating which latent dimensions are included in the probability function for each sub-response is provided as well. This structure should facilitate comprehension, with the models and modeling decisions being easier to understand given the general description of the modeling approach described thus far.

Study 1

Sample and data

Study 1 used data from a study of Smits, Dolan, Vorst, Wicherts and Timmerman (2011) downloaded via the Journal of Open Psychology Data (Smits, Dolan, Vorst, Wichers & Timmerman, 2013). The original dataset contained responses to a Dutch localization of a Big Five personality test (Vijf PersoonlijkheidsFactoren Test; Elshout &
Akkerman, 1975) from 9070 psychology freshmen (the published dataset contained responses from 8954 participants due to data pre-processing) at University of Amsterdam that had participated in the data collection in exchange for course credit in the years 1982 – 2007.

The questionnaire contained 70 items aiming to measure the Big Five personality traits of Agreeableness, Extraversion, Conscientiousness, Neuroticism and Openness to Experience. Each trait was measured by 14 items scored on a 7-point Likert scale with responses ranging from 1 (The description from the item *does not apply at all*) to 7 (The description *applies very well*). Reverse-scored items have been recoded in the data pre-processing phase by the authors and, due to copyright issues, the exact wording of items was unavailable.

For the purposes of this study, participants with at least one missing response were dropped from the dataset, reducing the sample to N = 7623. The reduced sample consisted of 69.6 % females, with the mean age of participants being 20.15 years, SD = 1.86.

*Method*

As stated before, the questionnaire was devised to measure five independent personality traits. Given the fact that the purpose of this study was modeling of extreme response style, the proposed model presented below should be understood as a general model formulation that was subsequently specified to fit all the items measuring one specific trait, and this process was repeated five times (once for each trait). Such separate
fitting of the model on different parts of the dataset (as opposed to jointly modeling all
the Big Five personality traits and possibly a single, general ERS dimension) allowed
focusing primarily on modeling ERS while keeping the five separate models relatively
simple. If a single joint model would have been fit on the entire dataset, it would have
introduced a significant level of complexity and, furthermore, modeling ERS as one
latent variable independent of the latent Big Five personality traits would have been too
strict an assumption. Taking these considerations into account, the specific-models
approach was found as more preferable, given that such an approach still allows for
further investigation of the degree of ERS generality, as will be explained further.

Thus, five instances of the proposed IRTree model were fit on the data (one for
each of the Big Five traits; for a general overview of IRTree models, see the introductory
Method section), as well as a number of slightly different alternative models which will
be briefly described. Below is the depiction of the general tree structure (Figure 5) which
was subsequently used for each model instance, and the mapping matrix (Table 3) that
represents the original response categories in terms of sub-responses.
The original responses 1 (Does not apply at all) through 7 (Applies very well) are modeled as outcomes of a series of sub-responses to sub-items (nodes) N1 through N6. For each node, following the right branch is indicated as “1” and following the left branch is indicated as “0”.

Rationale for the model formulation

The rationale for the chosen models is as follows: the first node (N1) represents the split between a neutral response category (4 – “Neutral”) and other response categories that have either positive or negative direction. The propensity for following the left / right branch at this node is best understood as a latent variable that is a mixture of different tendencies and attitudes, such as unwillingness to reveal information about oneself, uncooperativeness, lack of opinion or insight into oneself, lack of motivation to provide any other answer, not understanding the items, not being satisfied with the response categories provided, etc. (Baumgartner & Steenkamp, 2001; Kulas & Stachowski, 2013; Khorramdel & von Davier, 2014).
Table 3. The mapping matrix for Study 1.

<table>
<thead>
<tr>
<th>(Y_{ij} = 1)</th>
<th>(Y_{ij} = 2)</th>
<th>(Y_{ij} = 3)</th>
<th>(Y_{ij} = 4)</th>
<th>(Y_{ij} = 5)</th>
<th>(Y_{ij} = 6)</th>
<th>(Y_{ij} = 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>(NA)</td>
<td>0</td>
<td>(NA)</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>(NA)</td>
<td>1</td>
<td>(NA)</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>(NA)</td>
<td>(NA)</td>
<td>(NA)</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>(NA)</td>
<td>(NA)</td>
<td>(NA)</td>
<td>(NA)</td>
<td>(NA)</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>(NA)</td>
<td>0</td>
<td>(NA)</td>
<td>(NA)</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>(NA)</td>
<td>1</td>
<td>(NA)</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

This mapping matrix represents the original response categories as sets of sub-responses. For each sub-item, following the right branch is indicated as “1” and following the left branch is indicated as “0”, with \(NA\) for sub-responses that are not reached on the way to the particular response category.

Worthy (1969) considers neutral responding to be indicative of extreme response style, although some empirical research indicates otherwise (Plieninger & Meiser, 2014). Given that our goal was not to explain neutral or non-neutral answering, the first node was be considered as measuring a single latent tendency – non-neutral responding, \(\Theta_R\) (although the term non-neutral responding is much less common than neutral responding, it is used here purely for achieving consistency with the idea that “endorsing” a sub-item, that is, following the right branch, is indicative of the latent trait measured at that sub-item).
The second node (N2) represents the split between disagreement (1, 2 and 3) and agreement (5, 6 and 7) categories. Answering “0” to this sub-item means ultimately choosing some form of disagreement, whereas answering “1” will lead towards some form of agreement. As such, the sub-responses given at this node were considered to be influenced by the personality trait in question (i.e., one of the Big Five traits, $\Theta_{BF}$). Moreover, the sub-responses at this node should be in fact the most informative about a person’s level of the latent trait – indeed, one of the possible sub-responses ultimately indicates some form of disagreement with the given statement purportedly reflecting the latent trait (although the disagreement can be of varying severity), while the other possible sub-response ultimately indicates some form of agreement with such statement. If we were to offer a dichotomous response format instead, then all forms of agreement (or disagreement) would fall under one category and responses in such a format would be analogous to the sub-responses at Node 2 as described above. As such, it was hypothesized that the values of discrimination parameters (or loadings) for the latent personality trait should generally be the largest for this sub-item, and comparatively smaller for Nodes 3 and 4, and Nodes 5 and 6, because these lower-level nodes merely represent finer nuances of agreement or disagreement.

Node 2 does not yield any information about how extreme a response category the person will ultimately choose. For the subsequent lower-level nodes (Nodes 3, 4, 5 and 6), however, this is no longer true. These sub-items represent the split between a more extreme response category (or categories) and a less extreme response category. An example can be drawn from Node 3. Answering “1” at Node 3 (i.e., choosing the right
branch) results in arriving at a final response of “3”. The response of “3” is among the negative responses (that is, among the ones indicating disagreement with the given statement) while it is also the least extreme of the negative responses. Thus, simply put, such a response would indicate some level of disagreement, but not extreme responding. Answering “0” at Node 3 (i.e., choosing the left branch) results in arriving at Node 5 which then leads to either a final response of “1” or a final response of “2”, based on the corresponding sub-response. Both final responses indicate more severe disagreement with the given statement and also a more extreme response than a final response of “3”.

What can be drawn from this example is that following either the left or right branch of the lower-level sub-items is hypothetically still informative of a person’s personality trait level (in other words, persons with higher trait level should tend to choose the right branch at either node more often compared with persons with lower trait level), but it was also hypothesized that the sub-responses were influenced by a person’s extreme response style, $\Theta_{ERS}$. Following the left branch at Node 3 is indicative of more extreme response style than following the right branch (because it means ultimately choosing one of the more extreme responses) and this likewise applies to following the right branch at Node 4 (note the difference - due to the fact that choosing the right branch is indicative of ERS for Node 4 and Node 6, but by contrast, the same is true for choosing the left branch for Node 3 and Node 5, the ERS discrimination parameters should be of opposite signs for these two sets of nodes). In addition, the responses provided at the lowest-level nodes (Node 5 and Node 6) should be more informative of a person’s level of extreme response style than Nodes 3 and 4, because specific sub-responses at these
nodes can lead to the most extreme response categories (“1” or “7”). The absolute values of discrimination parameters of the sub-items N5 and N6 were, therefore, hypothesized to be comparatively larger than those of the sub-items N3 and N4, respectively. Thus, while responses to sub-items N1 and N2 were considered to be reflective of a single latent trait (Non-neutral responding and one of the Big Five personality traits, respectively), the other four sub-responses were hypothesized to be influenced by both extreme response style and the Big Five personality trait in question. This structure can be expressed with a discrimination matrix provided in Table 4 below. As stated before, in order for the model to be identified, the variance of latent variables was fixed to 1 and covariance between $\Theta_{BF}$ and $\Theta_{ERS}$ were constrained to 0. For interpretational simplicity, the covariances between $\Theta_R$ and both $\Theta_{BF}$ and $\Theta_{ERS}$ were also constrained to 0, although a model without the restricted covariance between $\Theta_R$ and $\Theta_{ERS}$ was fit and will be briefly discussed under the Unrestricted 3D model section of the Summary and discussion of the results below.

In addition to the three-dimensional model (3D) described above, the fit of alternative models that share identical tree structure with the 3D model but differ in the number of latent dimensions was also investigated. One of the alternative models was a one-dimensional IRTree model (1D) with no response style dimensions – that is, only the substantive trait was measured on Nodes 2-6 and no trait was measured on Node 1. The same tree structure was chosen for this model to allow for simple fit comparisons, but it is important to note that such structure is not optimal for a one-dimensional model since the responses to one sixth of the sub-items are modeled using only an intercept (and no
latent variable) – the choice between a neutral category and a non-neutral category cannot be meaningfully modeled with the substantive trait. For completeness, a one-dimensional model with a linear tree structure where this choice can be meaningfully modeled with a substantive trait was also fit (for more information on linear tree models, see the relevant chapter above or De Boeck & Partchev, 2012), but it is not included in the results since its fit was comparatively worse than that of the other tree models, as was the case for the 1D nested tree model reported here. A two-dimensional IRTree model (2D) was also fit, with the non-neutral responding and extreme response style latent dimensions joined into a single ERS dimension – the substantive trait was measured on Nodes 2-6 and the ERS dimension was measured on Nodes 3-6 and Node 1. Such an approach is for example suggested by Worthy (1969). The loading structure of the 1D and 2D models is also shown in Table 4. Furthermore, a unidimensional graded response model (GRM, Samejima, 1969) with no response style dimensions was fit on the data as a representative of an IRT model that would typically be fit on this kind of data.
Table 4. The discrimination matrices of sub-items for Study 1 (all IRTree models). Table 4 indicates which latent dimensions are hypothesized to influence which sub-responses in a given IRTree model. The number “0” indicates that the discrimination parameter of the respective sub-item for a given latent dimension was fixed to 0; and an $\alpha_{j1\Theta}$ parameter indicates that the discrimination parameter for latent dimension $\Theta$ at node n of item j was freely estimated. $\Theta_R$ represents “Non-neutral responding”, $\Theta_BF$ represents a Big Five personality trait and $\Theta_{ERS}$ represents “Extreme response style”.

<table>
<thead>
<tr>
<th></th>
<th>$\Theta_R$</th>
<th>$\Theta_{BF}$</th>
<th>$\Theta_{ERS}$</th>
<th>$\Theta_{BF}$</th>
<th>$\Theta_{ERS}$</th>
<th>$\Theta_{BF}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>$\alpha_{j1R}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\alpha_{j1ERS}$</td>
<td>0</td>
</tr>
<tr>
<td>N2</td>
<td>0</td>
<td>$\alpha_{j2BF}$</td>
<td>0</td>
<td>$\alpha_{j2BF}$</td>
<td>0</td>
<td>$\alpha_{j2BF}$</td>
</tr>
<tr>
<td>N3</td>
<td>0</td>
<td>$\alpha_{j3BF}$</td>
<td>$\alpha_{j3ERS}$</td>
<td>$\alpha_{j3BF}$</td>
<td>$\alpha_{j3ERS}$</td>
<td>$\alpha_{j3BF}$</td>
</tr>
<tr>
<td>N4</td>
<td>0</td>
<td>$\alpha_{j4BF}$</td>
<td>$\alpha_{j4ERS}$</td>
<td>$\alpha_{j4BF}$</td>
<td>$\alpha_{j4ERS}$</td>
<td>$\alpha_{j4BF}$</td>
</tr>
<tr>
<td>N5</td>
<td>0</td>
<td>$\alpha_{j5BF}$</td>
<td>$\alpha_{j5ERS}$</td>
<td>$\alpha_{j5BF}$</td>
<td>$\alpha_{j5ERS}$</td>
<td>$\alpha_{j5BF}$</td>
</tr>
<tr>
<td>N6</td>
<td>0</td>
<td>$\alpha_{j6BF}$</td>
<td>$\alpha_{j6ERS}$</td>
<td>$\alpha_{j6BF}$</td>
<td>$\alpha_{j6ERS}$</td>
<td>$\alpha_{j6BF}$</td>
</tr>
</tbody>
</table>

Results

All IRTree models were fit using the function `tam.mml.2pl` in the TAM package for IRT (Kiefer, Robitzsch, Wu, 2016) for R (R Core Team, 2015). The function utilizes full-information maximum likelihood estimation with EM optimizer, multidimensional integrals were approximated using Quasi-Monte Carlo integration with 3000 nodes. The GRM model was fit using the `mirt` package for R (Chalmers, 2012) utilizing full-information maximum likelihood estimation with EM optimizer and 61 Gaussian nodes.
The results below are organized as follows: separate sections are dedicated to each of the Big Five traits (Agreeableness, Extraversion, Conscientiousness, Neuroticism and Openness to Experience) and in each section the results of fitting the models are presented alongside model fit indices and short summaries of the parameter estimates for some of the models. After all models are presented, the overall results are summarized and briefly discussed.

**Agreeableness**

Data for Agreeableness contained responses of 7623 subjects to 14 items. The results of fitting all models are summarized in Table 5. The number of parameters for each model was the following: 224 parameters for the 3D and 2D models (84 sub-item $\beta$ parameters, because the data contained 14 items and the tree structure consisted of 6 nodes, and 140 $\alpha$ parameters – two for each sub-item except for the sub-items representing the first and second nodes where only a single latent dimension was measured. The number of parameters was equal because the only difference between the two models was that the two response style dimensions in the 3D model were joined into one in the 2D model), 154 parameters for the 1D model (84 sub-item $\beta$ parameters and 70 $\alpha$ parameters – the sub-items representing the first node did not measure any latent dimension) and 98 parameters for the GRM model (six threshold parameters per item – 84 in total, and one $\alpha$ parameter per item – 14 in total). The number of parameters for the models does not depend on the Big Five trait that is considered.
Table 5. Model fit of the proposed and alternative models (Agreeableness).
The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>nPar</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>306668</td>
<td>224</td>
<td>307116</td>
<td>308670</td>
</tr>
<tr>
<td>2D</td>
<td>306600</td>
<td>224</td>
<td>307047</td>
<td>308601</td>
</tr>
<tr>
<td>1D</td>
<td>314785</td>
<td>154</td>
<td>315093</td>
<td>316161</td>
</tr>
<tr>
<td>GRM</td>
<td>311000</td>
<td>98</td>
<td>311196</td>
<td>311876</td>
</tr>
</tbody>
</table>

The reliabilities of EAP estimates (computed as $\nu / (s + \nu)$ where $\nu$ is the variance of EAP estimates and $s$ is the mean squared error variance of measurement; Adams, 2005) of the latent traits from each model are listed in Table 6:

<table>
<thead>
<tr>
<th>Model</th>
<th>$\theta_R$</th>
<th>$\theta_A$</th>
<th>$\theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>.49</td>
<td>.72</td>
<td>.71</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>.70</td>
<td>.80</td>
</tr>
<tr>
<td>1D</td>
<td>-</td>
<td>.82</td>
<td>-</td>
</tr>
<tr>
<td>GRM</td>
<td>-</td>
<td>.85</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 6. EAP reliabilities of latent trait estimates (Agreeableness).

Table 7 below shows the mean values and the standard deviations of the discrimination parameter estimates at each node and latent trait for the two best fitting
models (3D and 2D). The mean values are shown because they neatly illustrate the overall trend in the values of parameter estimates as the tree progresses to its bottom.

Note that the SDs represent the dispersion of the discrimination parameter estimates, not the standard error of the mean discrimination parameter.

Table 7. Summary of discrimination parameter estimates at each node (Agreeableness). Means are written in bold, while standard deviations are in parentheses and italicized.  

<table>
<thead>
<tr>
<th>Node</th>
<th>$\Theta_R$</th>
<th>$\Theta_A$</th>
<th>$\Theta_{ERS}$</th>
<th>$\Theta_A$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td><strong>0.81</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td><strong>0.53</strong></td>
</tr>
<tr>
<td>N2</td>
<td>0</td>
<td><strong>1.52</strong></td>
<td>0</td>
<td>1.53</td>
<td>0</td>
</tr>
<tr>
<td>N3</td>
<td>0</td>
<td><strong>0.58</strong></td>
<td><strong>-0.81</strong></td>
<td>0.76</td>
<td><strong>-0.76</strong></td>
</tr>
<tr>
<td>N4</td>
<td>0</td>
<td><strong>0.56</strong></td>
<td><strong>0.89</strong></td>
<td><strong>0.28</strong></td>
<td><strong>0.99</strong></td>
</tr>
<tr>
<td>N5</td>
<td>0</td>
<td><strong>0.60</strong></td>
<td><strong>-1.14</strong></td>
<td><strong>0.81</strong></td>
<td><strong>-1.06</strong></td>
</tr>
<tr>
<td>N6</td>
<td>0</td>
<td><strong>0.23</strong></td>
<td><strong>1.46</strong></td>
<td><strong>-0.20</strong></td>
<td><strong>1.44</strong></td>
</tr>
</tbody>
</table>

Extraversion

Data for Extraversion contained responses of 7623 subjects to 14 items. The results of fitting all models are summarized in Table 8, with EAP reliabilities of the latent traits in Table 9, and the mean values and standard deviations of the discrimination parameter estimates for the two best fitting models are shown in Table 10:
<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>nPar</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>331314</td>
<td>224</td>
<td>331762</td>
<td>333316</td>
</tr>
<tr>
<td>2D</td>
<td>331304</td>
<td>224</td>
<td>331752</td>
<td>333306</td>
</tr>
<tr>
<td>1D</td>
<td>338524</td>
<td>154</td>
<td>338832</td>
<td>339900</td>
</tr>
<tr>
<td>GRM</td>
<td>334348</td>
<td>98</td>
<td>334544</td>
<td>335224</td>
</tr>
</tbody>
</table>

Table 8. Model fit of the proposed and alternative models (Extraversion). *Means are written in bold, while standard deviations are in parentheses and italicized.*

<table>
<thead>
<tr>
<th>Model</th>
<th>$\Theta_R$</th>
<th>$\Theta_E$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>.48</td>
<td>.83</td>
<td>.65</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>.82</td>
<td>.72</td>
</tr>
<tr>
<td>1D</td>
<td>-</td>
<td>.87</td>
<td>-</td>
</tr>
<tr>
<td>GRM</td>
<td>-</td>
<td>.89</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 9. EAP reliabilities of latent trait estimates (Extraversion).
Table 10. Summary of discrimination parameter estimates at each node (Extraversion). *Means are written in bold, while standard deviations are in parentheses and italicized.*

<table>
<thead>
<tr>
<th>Node</th>
<th>$\Theta_R$</th>
<th>$\Theta_E$</th>
<th>$\Theta_{ERS}$</th>
<th>$\Theta_R$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>0.67 (.20)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.49 (0.12)</td>
</tr>
<tr>
<td>N2</td>
<td>0</td>
<td>1.74 (0.82)</td>
<td>0</td>
<td>1.75 (0.83)</td>
<td>0</td>
</tr>
<tr>
<td>N3</td>
<td>0</td>
<td>0.67 (0.34)</td>
<td>-0.79 (0.13)</td>
<td>0.67 (0.33)</td>
<td>-0.79 (0.14)</td>
</tr>
<tr>
<td>N4</td>
<td>0</td>
<td>0.79 (0.39)</td>
<td>0.78 (0.15)</td>
<td>0.64 (0.37)</td>
<td>0.84 (0.18)</td>
</tr>
<tr>
<td>N5</td>
<td>0</td>
<td>0.47 (0.34)</td>
<td>-1.24 (0.41)</td>
<td>0.43 (0.33)</td>
<td>-1.19 (0.39)</td>
</tr>
<tr>
<td>N6</td>
<td>0</td>
<td>0.58 (0.35)</td>
<td>1.31 (0.27)</td>
<td>0.34 (0.29)</td>
<td>1.29 (0.31)</td>
</tr>
</tbody>
</table>

*Conscientiousness*

Data for Conscientiousness contained responses of 7623 subjects to 14 items. The results of fitting all models are summarized in Table 11, with EAP reliabilities of the latent traits in Table 12, and the mean values and standard deviations of the discrimination parameter estimates for the two best fitting models are shown in Table 13:
<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>nPar</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>340648</td>
<td>224</td>
<td>341096</td>
<td>342651</td>
</tr>
<tr>
<td>2D</td>
<td>340396</td>
<td>224</td>
<td>340844</td>
<td>342398</td>
</tr>
<tr>
<td>1D</td>
<td>347936</td>
<td>154</td>
<td>348244</td>
<td>349313</td>
</tr>
<tr>
<td>GRM</td>
<td>345352</td>
<td>98</td>
<td>345548</td>
<td>346228</td>
</tr>
</tbody>
</table>

Table 11. Model fit of the proposed and alternative models (Conscientiousness). The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\theta_R$</th>
<th>$\theta_C$</th>
<th>$\theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>.39</td>
<td>.81</td>
<td>.67</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>.80</td>
<td>.71</td>
</tr>
<tr>
<td>1D</td>
<td>-</td>
<td>.82</td>
<td>-</td>
</tr>
<tr>
<td>GRM</td>
<td>-</td>
<td>.85</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 12. EAP reliabilities of latent trait estimates (Conscientiousness).
### Table 13. Summary of discrimination parameter estimates at each node (Conscientiousness).

*Means are written in bold, while standard deviations are in parentheses and italicized.*

**Neuroticism**

Data for Neuroticism contained responses of 7623 subjects to 14 items. The results of fitting all models are summarized in Table 14, with EAP reliabilities of the latent traits in Table 15, and the mean values and standard deviations of the discrimination parameter estimates for the two best fitting models are shown in Table 16:

<table>
<thead>
<tr>
<th></th>
<th>3D Model</th>
<th></th>
<th>2D model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Theta_R$</td>
<td>$\Theta_C$</td>
<td>$\Theta_{ERS}$</td>
<td>$\Theta_C$</td>
<td>$\Theta_{ERS}$</td>
</tr>
<tr>
<td>N1</td>
<td>0.57 (.07)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.39 (0.08)</td>
</tr>
<tr>
<td>N2</td>
<td>0</td>
<td>1.35 (0.69)</td>
<td>0</td>
<td>1.36 (0.69)</td>
<td>0</td>
</tr>
<tr>
<td>N3</td>
<td>0</td>
<td>0.59 (0.46)</td>
<td>-0.78 (0.16)</td>
<td>0.59 (0.46)</td>
<td>-0.80 (0.17)</td>
</tr>
<tr>
<td>N4</td>
<td>0</td>
<td>0.59 (0.32)</td>
<td>0.78 (0.14)</td>
<td>0.54 (0.31)</td>
<td>0.81 (0.16)</td>
</tr>
<tr>
<td>N5</td>
<td>0</td>
<td>0.45 (0.36)</td>
<td>-1.24 (0.21)</td>
<td>0.44 (0.36)</td>
<td>-1.23 (0.21)</td>
</tr>
<tr>
<td>N6</td>
<td>0</td>
<td>0.45 (0.25)</td>
<td>1.28 (0.21)</td>
<td>0.36 (0.23)</td>
<td>1.27 (0.22)</td>
</tr>
</tbody>
</table>
Table 14. Model fit of the proposed and alternative models (Neuroticism).
The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>nPar</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>329883</td>
<td>224</td>
<td>330331</td>
<td>331885</td>
</tr>
<tr>
<td>2D</td>
<td>329808</td>
<td>224</td>
<td>330256</td>
<td>331811</td>
</tr>
<tr>
<td>1D</td>
<td>338421</td>
<td>154</td>
<td>338729</td>
<td>339798</td>
</tr>
<tr>
<td>GRM</td>
<td>333738</td>
<td>98</td>
<td>333935</td>
<td>334615</td>
</tr>
</tbody>
</table>

Table 15. EAP reliabilities of latent trait estimates (Neuroticism).

<table>
<thead>
<tr>
<th>Model</th>
<th>$\Theta_R$</th>
<th>$\Theta_N$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>.48</td>
<td>.83</td>
<td>.68</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>.82</td>
<td>.76</td>
</tr>
<tr>
<td>1D</td>
<td>-</td>
<td>.88</td>
<td>-</td>
</tr>
<tr>
<td>GRM</td>
<td>-</td>
<td>.91</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 14. Model fit of the proposed and alternative models (Neuroticism).
The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.

Table 15. EAP reliabilities of latent trait estimates (Neuroticism).
### Table 16. Summary of discrimination parameter estimates at each node (Neuroticism).

Means are written in bold, while standard deviations are in parentheses and italicized.

<table>
<thead>
<tr>
<th></th>
<th>3D Model</th>
<th></th>
<th>2D model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Theta_R$</td>
<td>$\Theta_N$</td>
<td>$\Theta_{ERS}$</td>
<td>$\Theta_N$</td>
</tr>
<tr>
<td>N1</td>
<td>0.71 (0.21)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N2</td>
<td>0</td>
<td>1.86 (0.81)</td>
<td>0</td>
<td>1.87 (0.81)</td>
</tr>
<tr>
<td>N3</td>
<td>0</td>
<td>0.83 (0.41)</td>
<td>-0.80 (0.12)</td>
<td>0.63 (0.37)</td>
</tr>
<tr>
<td>N4</td>
<td>0</td>
<td>0.91 (0.32)</td>
<td>0.85 (0.13)</td>
<td>0.93 (0.29)</td>
</tr>
<tr>
<td>N5</td>
<td>0</td>
<td>0.58 (0.32)</td>
<td>-1.53 (0.30)</td>
<td>0.22 (0.24)</td>
</tr>
<tr>
<td>N6</td>
<td>0</td>
<td>0.76 (0.32)</td>
<td>1.25 (0.25)</td>
<td>0.73 (0.27)</td>
</tr>
</tbody>
</table>

**Openness to Experience**

Data for Openness contained responses of 7623 subjects to 14 items. The results of fitting all models are summarized in Table 17, with EAP reliabilities of the latent traits in Table 18, and the mean values and standard deviations of the discrimination parameter estimates for the two best fitting models are shown in Table 19:
### Table 17. Model fit of the proposed and alternative models (Openness).

The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>nPar</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>324187</td>
<td>224</td>
<td>324635</td>
<td>326189</td>
</tr>
<tr>
<td>2D</td>
<td>323926</td>
<td>224</td>
<td>324374</td>
<td>325928</td>
</tr>
<tr>
<td>1D</td>
<td>332315</td>
<td>154</td>
<td>332623</td>
<td>333691</td>
</tr>
<tr>
<td>GRM</td>
<td>327856</td>
<td>98</td>
<td>328053</td>
<td>328733</td>
</tr>
</tbody>
</table>

### Table 18. EAP reliabilities of latent trait estimates (Openness).

<table>
<thead>
<tr>
<th>Model</th>
<th>$\Theta_R$</th>
<th>$\Theta_O$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>.53</td>
<td>.77</td>
<td>.66</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>.76</td>
<td>.77</td>
</tr>
<tr>
<td>1D</td>
<td>-</td>
<td>.81</td>
<td>-</td>
</tr>
<tr>
<td>GRM</td>
<td>-</td>
<td>.85</td>
<td>-</td>
</tr>
</tbody>
</table>

The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.
### Table 19. Summary of discrimination parameter estimates at each node (Openness).

*Means are written in bold, while standard deviations are in parentheses and italicized.*

<table>
<thead>
<tr>
<th>Node</th>
<th>$\Theta_R$</th>
<th>$\Theta_O$</th>
<th>$\Theta_{ERS}$</th>
<th>$\Theta_O$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td><strong>0.73 (0.20)</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td><strong>0.59 (0.14)</strong></td>
</tr>
<tr>
<td>N2</td>
<td>0</td>
<td><strong>1.45 (0.52)</strong></td>
<td>0</td>
<td><strong>1.46 (0.54)</strong></td>
<td>0</td>
</tr>
<tr>
<td>N3</td>
<td>0</td>
<td><strong>0.51 (0.33)</strong></td>
<td><strong>-0.85 (0.19)</strong></td>
<td><strong>0.67 (0.33)</strong></td>
<td><strong>-0.85 (0.20)</strong></td>
</tr>
<tr>
<td>N4</td>
<td>0</td>
<td><strong>0.65 (0.31)</strong></td>
<td><strong>0.85 (0.18)</strong></td>
<td><strong>0.38 (0.31)</strong></td>
<td><strong>0.94 (0.21)</strong></td>
</tr>
<tr>
<td>N5</td>
<td>0</td>
<td><strong>0.15 (0.35)</strong></td>
<td><strong>-1.48 (0.52)</strong></td>
<td><strong>0.38 (0.37)</strong></td>
<td><strong>-1.34 (0.38)</strong></td>
</tr>
<tr>
<td>N6</td>
<td>0</td>
<td><strong>0.43 (0.22)</strong></td>
<td><strong>1.31 (0.21)</strong></td>
<td><strong>0.03 (0.25)</strong></td>
<td><strong>1.31 (0.25)</strong></td>
</tr>
</tbody>
</table>

The results of the above part of Study 1, together with the results of the following part, are summarized and discussed in the *Summary and discussion of the results* section on page 64 and then further with the results of Study 2 in the *General discussion* section at the end of the thesis on page 93.

**Relating the constructs across models**

Although an overview of the two best fitting models (2D and 3D) reveals that the discrimination parameter estimates for extreme response style followed the expected pattern (the results of fitting the models and the parameter estimates will be discussed in a while), that is not itself sufficient in establishing the validity of modeling ERS in this manner. First of all, modeling the responses for each of the Big Five traits separately does
not allow to draw conclusions regarding the similarity of essence of whatever was
modeled as extreme response style across the five models. Do the five different estimates
for each person reflect the same response tendency? Does ERS as modeled in the
Agreeableness model have anything in common at all with ERS as modeled in, say, the
Neuroticism model? Answering these questions is crucial in making sense of the modeled
response style.

Furthermore, the way the models have been formulated did not allow for
examining the relationship between ERS (and non-neutral responding, for that matter)
and the personality traits of the Big Five. Such an examination is, however, important for
putting the measured variables in context and assessing the construct validity of extreme
response style, although it is by no means sufficient.

For these purposes, two structural models were fit on the EAP score estimates of
each modeled latent variable for each person obtained from either the 2D or the 3D
model. Thus, the data for each person contained an EAP score of all the Big Five
personality traits (used as random predictors in this model), five extreme response style
EAP scores (each from a different Big Five model) and, in case of estimates from the 3D
model, five non-neutral responding EAP scores. Both the extreme response style scores
and the non-neutral responding scores were used in a measurement model as variables
loading on a (higher-order) latent variable. These higher-order latent variables,
representing general response style factors, were allowed to correlate (although this only
applies to the model of the 3D estimates, since in the 2D case there is only one general
response style factor present) and were regressed on the Big Five personality trait
estimates. Moreover, each Big Five personality trait estimate was allowed to correlate with the residuals of response style estimates coming from the same IRTree model (so, for example, the Extroversion EAP estimate was allowed to correlate with the residuals of non-neutral responding and ERS estimates from the Extroversion 3D IRTree model). Finally, in case of the SEM model using the 3D IRTree estimates, the residuals of non-neutral responding and ERS estimates from the same Big Five IRTree model were allowed to correlate as well. It should be pointed out that using EAP estimates from IRT models as variables in the structural models means not taking advantage of the disattenuation for imperfect reliabilities SEM is normally able to provide. The models were fit using normal maximum likelihood estimation (with conventional standard errors computed using the observed information matrix) in the package lavaan (Rosseel, 2012) for R.

2D model

For the structural model based on latent score estimates from the 2D IRTree model, the (non-adjusted) minimum function test statistic was $\chi^2(20) = 924, p < 0.01$; RMSEA = 0.077 (90% CI = 0.073; 0.081), sRMR = 0.038, TLI = 0.91, CFI = 0.96. The path diagram of the model is shown in Figure 6. Note that only the regression parameters, factor loadings, residual variances and the latent variable variance (all standardized) are shown in Figure 6. In order to reduce clutter, the covariance parameters of the exogenous variables are shown in Table 20 below.
Figure 6. Structural model of the EAP estimates from Study 1, 2D model. 

The variables in this model are labeled using the first letter of the respective Big Five personality trait from the model of which they were extracted. The single letter labels refer to estimates of the actual personality trait, while labels ending with “_E” refer to extreme response style estimates (e.g., “A” denotes Agreeableness, and “A_E” denotes extreme response style, both as modeled in the Agreeableness model). The latent variable “ERS” refers to extreme response style. For the italicized parameter estimates, p > 0.05.
Table 20. Covariances of the exogenous variables in the structural model, 2D model. Each of the 5 ERS score residuals was allowed to covary only with the Big Five personality trait score from the same IRTree model – thus, the BF_E row contains the covariances of the Big Five personality trait score in the given column with the respective ERS score residual.

\* = p > .05

3D model

For the structural model based on latent score estimates from the 3D IRTree model, the (non-adjusted) minimum function test statistic was $\chi^2(59) = 990.5$, $p < 0.01$; RMSEA = 0.046 (90% CI = 0.043; 0.048), sRMR = 0.030, TLI = 0.95, CFI = 0.97. The path diagram of the model is shown in Figure 7. Note that only the regression parameters, factor loadings, residual variances and the latent variable variances (all standardized) are shown in Figure 7. In order to reduce clutter, the covariance parameters of the exogenous variables are shown in Tables 21 and 22 below.
Figure 7. Structural model of the EAP estimates from Study 1, 3D model. The variables are labeled in an identical way to Figure 6. Additionally, labels ending with “_R” refer to non-neutral responding (e.g., “A_R” denotes non-neutral responding as modeled in the Agreeableness model). The latent variable “R” refers to non-neutral responding. For the italicized parameter estimates, $p > 0.05$. 
<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
<th>E</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>.09</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.25</td>
<td>.03</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-.26</td>
<td>-.09</td>
<td>-.27</td>
<td>1</td>
</tr>
<tr>
<td>O</td>
<td>-.04</td>
<td>.04</td>
<td>.11</td>
<td>-.06</td>
</tr>
<tr>
<td>BF_E</td>
<td>.21</td>
<td>-.08</td>
<td>-.07</td>
<td>-.06</td>
</tr>
<tr>
<td>BF_R</td>
<td>.38</td>
<td>.02*</td>
<td>.13</td>
<td>-.23</td>
</tr>
</tbody>
</table>

Table 21. Covariances of the personality trait scores in the structural model, 3D model. Each of the 5 pairs of ERS and non-neutral responding score residuals was allowed to covary only with the Big Five personality trait score from the same IRTree model – thus, the BF_E row contains the covariances of the Big Five personality trait score in the given column with the respective ERS score residual, and the same holds for the BF_R row of non-neutral responding score residuals covariances.

* = p > .05

<table>
<thead>
<tr>
<th>A_R</th>
<th>C_R</th>
<th>E_R</th>
<th>N_R</th>
<th>O_R</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_E</td>
<td>.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_E</td>
<td></td>
<td>.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E_E</td>
<td></td>
<td></td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td>N_E</td>
<td></td>
<td></td>
<td></td>
<td>.16</td>
</tr>
<tr>
<td>O_E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 22. Covariances of the response style score residuals in the structural model, 3D model.
Summary and discussion of the results

Considering fit of the IRT models, it can be concluded that the multidimensional models (2D and 3D) are superior in fit to the unidimensional models (1D and GRM), while the 2D model consistently fits slightly better than the proposed 3D model, although the differences are rather minimal. There might be two plausible explanations for this observation. One possible explanation is that differentiating the non-neutral responding dimension from extreme response style is redundant since the two dimensions are actually parts of the same construct (as proposed by Worthy, 1969). An alternative explanation is that they might not be the parts of the same construct but are in fact related (see the structural model for the 3D IRTree model EAP estimates and the Unrestricted 3D model section below) and forcing them to be strictly independent under the 3D model hindered model fit. As mentioned earlier, non-neutral responding most likely represents a mixture of different tendencies which might include (but not be limited to) both ERS and the Big Five trait in question or be related to them – in such case, modeling it as a part of either one of these traits might improve the model fit as compared to completely separating it as an independent dimension. In fact, it can be seen that across all the 3D IRTree models, non-neutral answering reliabilities were consistently quite low, mostly below .5 – such low values might reflect that this latent variable represents a mixture of different tendencies, although also the fact that it was measured only at single node (and thus little information was available compared with the other latent variables) is very likely to hinder reliability.
The reliabilities of the Big Five trait estimates were higher for the simpler models (GRM and 1D), however, even for the more complex and better fitting models (2D and 3D) the values seem sufficiently large. The lower reliabilities might suggest that if the IRTree models are indeed a more precise reflection of the underlying processes, then there is in fact less trait-related information in the data than unidimensional models assume. This would mean that it is in fact not the multidimensional models underestimating true reliability, but unidimensional models overestimating it. As for the ERS estimates, the reliabilities mostly range between .65 and .80, with the two-dimensional IRTree model providing overall more reliable estimates. This can be assumed to be the consequence of ERS being measured over a higher number of sub-responses in the 2D model as opposed to the 3D model. Naturally, the reliability of ERS estimates might increase significantly if modeled jointly with all Big Five latent traits as a unidimensional trait, however the generality of extreme response style should not be assumed automatically.

Another point of interest are the discrimination parameters of the two best fitting models. Listed in the tables above are the means of discrimination parameter estimates across all sub-responses for each latent variable modeled. We hypothesized that for the Big Five personality dimensions, the average value of the discrimination parameter estimates would decrease as the tree progresses to its bottom, reflecting that lower-order sub-items are less and less relevant in (i.e., yield less information about) measuring the trait in question. It can be clearly seen that this trend generally holds for both the 2D and the 3D models. The discrimination parameters for the second node were, on average,
always the largest. The third and fourth node had smaller discrimination parameters, and those were smaller still for the fifth and sixth nodes (only once, in the case of the Agreeableness model, the fifth node actually had a slightly larger mean of the discrimination parameter estimates than the third node). In two cases, the mean discrimination parameter for the sixth node was actually negative (Agreeableness, see Table 7) or near-zero (Openness, see Table 19), suggesting that the choice between responses of “5” or “6” was not at all indicative or even inversely indicative of the Big Five trait in question. Moreover, there is no consistent trend in the standard deviation values of the discrimination parameter estimates, suggesting that the dispersion of estimates does not change in any systematic fashion. For ERS, the opposite was predicted – the mean absolute value of the discrimination parameter estimates would be larger for lower-level nodes compared to nodes immediately one level above. The tables above show that this pattern was indeed present across both the 2D and 3D models, suggesting that lower-level sub-responses were more intensely influenced by a person’s level of ERS.

The structural equation models of EAP estimates from both the 2D and 3D models are postulating the influence of a general latent response style factor on the response style estimates obtained from all five trait-specific IRTree models fitted on items measuring each of the Big Five personality traits. Moreover, direct effects of the Big Five personality trait estimates on the general response style factors are modeled. Both structural models fit well and the trait-specific ERS estimates had fairly high loadings on the general ERS factor in both cases. In the case of the 3D model, trait-
specific non-neutral responding estimates loaded less highly on their respective general factor and a sizeable proportion of their variance remains unexplained. In some cases, however, their unique variance was correlated with the Big Five personality trait estimate from the same IRTree model (for example, Agreeableness, Neuroticism or Openness to Experience), which was not the case for ERS estimates in the 3D model (with the exception of Agreeableness). The unique variances of ERS estimates did correlate more with the Big Five personality trait estimates in case of the 2D model, but as can be noted from the comparison of the two SEM models, it was due to the fact that the trait-specific ERS and trait-specific non-neutral responding were not differentiated and were modeled as a single latent variable in the 2D model. It is clear that the non-neutral responding estimates did not share much common variance, which is attributable to the potentially mixed character of the response style as modeled in the IRTree models, as mentioned above, and to the way the response style was being measured (through only a single node). In a number of cases (such as for Agreeableness and Openness in the 2D model or Agreeableness, Extroversion or Neuroticism in the 3D model, for example), the relationship between a particular personality trait and a general response style factor (either ERS or non-neutral responding) was markedly different in size or even in sign from the trait’s relationship with the residual of the corresponding trait-specific response style estimate. This is an interesting observation – although it could be argued that this might be partially due to artifactual covariance (see the General discussion section at the end of the thesis), it could also be reasoned that this reflects a different relationship between the personality traits and general versus trait-specific response styles.
The two general response style factors correlate in the 3D structural model, though the trait-specific response style latent variables were modeled as independent in all instances of the 3D IRTree model. The correlation is not surprising, since imposing the constraint of independence on two latent variables within a model does not mean that their estimates will be uncorrelated, and furthermore, it was the general response style factors that were correlated in the SEM model.

As can be seen from both structural models, Agreeableness had a consistently mild relationship with the general ERS factor, and Extraversion was related to both general response styles (a link between Extraversion and ERS has previously been reported in the literature, see Austin, Deary & Egan, 2006 or Meiser & Machunsky, 2008).

**Unrestricted 3D model**

As mentioned above, two covariances in the 3D IRTree models were constrained to zero in order to improve interpretational simplicity – the covariance between non-neutral responding and ERS and the covariance between non-neutral responding and the corresponding Big Five personality trait. The three main priorities in this thesis were assessing relative model fit of the IRTree models, studying the pattern of discrimination parameters at all nodes for each IRTree model and evaluating the relationships between the latent variable estimates in the SEM models. The following short summary focuses on what are the implications regarding these three priorities when these two covariances are unconstrained.
In general, the unrestricted 3D IRTree models fitted better than the restricted 3D model or the 2D model presented above, which is not surprising as a correlation between non-neutral responding and ERS is to be expected. IRT reliabilities of both response style dimensions were slightly improved and the pattern of discrimination parameters reported above remained unchanged. With respect to the SEM models, if non-neutral responding would be allowed to correlate with the substantive trait within the IRTree model but ERS would not, this would mean, in effect, that the two response styles would be modeled under unequal conditions and it would in turn prove more difficult to interpret and compare their relationship with the substantive traits. Not restricting the correlation of non-neutral responding and extreme response style within the IRTree model would pose a difficulty in disentangling these two response styles coming from the same trait-specific IRTree model and capturing general response style variance in the SEM model. The SEM model fit on the EAP estimates from the unrestricted model fitted slightly worse than its restricted 3D counterpart, the loadings of ERS estimates on the general ERS factor were decreased and the reverse was true for the loadings of non-neutral responding estimates on their corresponding general factor. The resulting covariance of the two general response style factors in the SEM model was very large (almost .90) and the covariances of the Big Five trait estimates with the residuals of trait-specific ERS estimates were in some cases as high as those of the Big Five trait estimates with the residuals of trait-specific non-neutral responding estimates. These results suggest that not restricting the covariance between the two response style dimensions within the IRTree model blurred the boundary between them to a substantial degree.
Study 2

Sample and data

This study used data obtained with a Big Five Factor Markers from the International Personality Item Pool (Goldberg, 1992) administered online on a public website (http://personality-testing.info). The data were downloaded from a repository on the same website (http://personality-testing.info/_rawdata) on March 2016, and represent a sample of self-selected participants that filled out the questionnaire and agreed with their data being stored and made publicly available. For the purposes of this study, only data from participants identifying as American whose first language is English were used (N = 7885, 66% female, mean age 27.1, SD = 13). The questionnaire contained 50 items (10 corresponding to each trait) with a response scale ranging from 1 (Strongly Disagree) through 3 (Neutral) to 5 (Strongly Agree).

Method

As in Study 1, the questionnaire was designed to measure the Big Five personality traits. The purpose of Study 2 is, again, modeling of extreme response style, and so the proposed model presented below should be understood as a general model formulation afterwards specified and fit on items measuring each specific trait.

Five instances of the proposed IRTree model were fit on the data (one model per one Big Five trait; for a general overview of IRTree models, see the introductory Method section), as well as a number of slightly different alternative models which will be briefly
described. Below is the depiction of the general tree structure (Figure 8) which was subsequently used for each model instance and the mapping matrix (Table 23) that represents the original response categories in terms of sub-responses.

Figure 8. The depiction of the tree structure for Study 2.

The original responses 1 (Strongly Disagree) through 5 (Strongly agree) are modeled as outcomes of a series of sub-responses to sub-items (nodes) N1 through N4. For each node, following the right branch is indicated as “1” and following the left branch is indicated as “0”.

71
Table 23. The mapping matrix for Study 2.

This mapping matrix represents the original response categories as sets of sub-responses. For each sub-item, following the right branch is indicated as “1” and following the left branch is indicated as “0”, with NA for sub-responses that are not reached on the way to the particular response category.

Rationale for the model formulation

The rationale for the chosen models is similar as in Study 1: the first node (N1) represents the split between a neutral response category (3 – “Neutral”) and other response categories that indicate either agreement or disagreement. Again, the first node was considered as measuring a single latent tendency – non-neutral answering ($\Theta_R$).

The second node (N2) represents the split between response categories indicating disagreement (1 and 2) and those indicating agreement (5 and 6). Answering “0” to this sub-item means ultimately choosing some form of disagreement, whereas answering “1” will lead towards some form of agreement. As such, the sub-responses given at this node was considered to be influenced by the personality trait in question (i.e., one of the Big Five traits, $\Theta_{BF}$). As in the previous study, the sub-responses at this node should be in fact
the most informative about a person’s level of the latent trait and so the discrimination parameters (or loadings) for the latent personality trait should generally be the largest for this sub-item, while being smaller for Nodes 3 and 4.

Nodes 3 and 4 represent the split between an extreme response category and a non-extreme response category. Just as in Study 1, following either the left or right branch of the lower-level sub-items should still be informative of a person’s personality trait level - the higher the trait level of a person, the more likely they are to choose the right branch at either node. The sub-responses are also influenced by a person’s extreme response style ($\Theta_{ERS}$). Following the left branch at Node 3 is indicative of more extreme response style than following the right branch (because it means ultimately choosing one of the more extreme responses) and this likewise applies to following the right branch at Node 4. What has been hypothesized in Study 1 – i.e. that the discrimination parameters of the lowest-level sub-responses for extreme response style will be comparatively higher than the discrimination parameters of the sub-responses immediately one level above – cannot be hypothesized in this case. The number of original response categories does not allow for the same six-node structure as was the case in Study 1.

Again, while responses to sub-items N1 and N2 are considered to be reflective of a single latent trait (non-neutral responding and one of the Big Five personality traits, respectively), the other two sub-responses are hypothesized to be influenced by both ERS and the Big Five personality trait in question. This structure can be expressed with a discrimination matrix provided in Table 24 below. Again, in order for the model to be identified, the variance of latent variables was fixed to 1 and covariance between $\Theta_{BF}$ and
Θ_{ERS} was constrained to 0. Just as in Study 1, the covariances between Θ_{R} and both Θ_{BF} and Θ_{ERS} were also constrained to 0 and a model without the restricted covariance between Θ_{R} and Θ_{ERS} was again fit on the data. The reasons for this decision as well as conclusions about the model fit are identical to Study 1. For more information, refer to the Unrestricted 3D model section in the results of Study 1.

Identically to Study 1, the fit of alternative models that share an identical tree structure with the three-dimensional (3D) model but differ in the number of latent dimensions was also investigated. A one-dimensional IRTree model (1D) with no response style dimensions, a two-dimensional IRTree model (2D) with the non-neutral responding and extreme response style latent dimensions joined into a single ERS dimension and a unidimensional graded response model (GRM) with no response style dimensions were fit on the data. The loading structure for the 2D and 1D IRTree models is also shown in Table 24 below. For a more in-depth description of the alternative models see the corresponding section in Study 1 – the only difference between the alternative models in Study 1 and Study 2 lies in a different tree structure due to the smaller number of original response categories. Note that in both studies (as described earlier) the fit of a linear tree variant of the 1D model was also investigated, but it is not included in the results since its fit was comparatively worse than that of the other tree models, as was the case for the nested tree 1D model presented here.
Table 24. The discrimination matrices of sub-items for Study 2 (all IRTree models).

|       | 3D Model | | | 2D Model | | | 1D Model | |
|-------|----------|---------|---------|----------|---------|---------|----------| |
|       | $\theta_R$ | $\theta_{BF}$ | $\theta_{ERS}$ | $\theta_{BF}$ | $\theta_{ERS}$ | $\theta_{BF}$ | |
| N1    | $\alpha_{j1R}$ | 0 | 0 | 0 | $\alpha_{j1ERS}$ | 0 | |
| N2    | 0 | $\alpha_{j2BF}$ | 0 | $\alpha_{j2BF}$ | 0 | $\alpha_{j2BF}$ | |
| N3    | 0 | $\alpha_{j3BF}$ | $\alpha_{j3ERS}$ | $\alpha_{j3BF}$ | $\alpha_{j3ERS}$ | $\alpha_{j3BF}$ | |
| N4    | 0 | $\alpha_{j4BF}$ | $\alpha_{j4ERS}$ | $\alpha_{j4BF}$ | $\alpha_{j4ERS}$ | $\alpha_{j4BF}$ | |

Table 24 indicates which latent dimensions are hypothesized to influence which sub-responses in a given IRTree model. The number “0” indicates that the discrimination parameter of the respective sub-item for a given latent dimension was fixed to 0; and an $\alpha_{jn\theta}$ parameter indicates that the discrimination parameter for latent dimension $\theta$ at node n of item j was freely estimated. $\theta_R$ represents “Non-neutral responding”, $\theta_{BF}$ represents one of the Big Five personality traits (different for each fitted model) and $\theta_{ERS}$ represents “Extreme response style”.

**Results**

Identically to Study 1, all IRTree models were fit using the function `tam.mml.2pl` in the TAM package for IRT for R, utilizing full-information maximum likelihood estimation with EM optimizer. Multidimensional integrals were approximated using Quasi-Monte Carlo integration with 3000 nodes. The GRM model was fit using the `mirt` package for R with full-information maximum likelihood estimation with EM optimizer and 61 Gaussian nodes.
The results below are organized as follows: separate sections are dedicated to each of the Big Five traits (Agreeableness, Extraversion, Conscientiousness, Neuroticism and Openness to Experience) and in each section the results of fitting the models are presented alongside model fit indices and short summaries of the parameter estimates for some of the models. After all models are presented, the overall results are summarized and briefly discussed.

Agreeableness

Data for Agreeableness contained responses of 7884 respondents to 10 items. The results of fitting all models are summarized in Table 25, with EAP reliabilities of the latent traits in Table 26, and the mean values and standard deviations of the discrimination parameter estimates for the two best fitting models are shown in Table 27. The number of parameters for each model was the following: 100 parameters for the 3D and 2D models (40 sub-item $\beta$ parameters, because the data contains 10 items and the tree structure consists of 4 nodes, and 60 $\alpha$ parameters – two for each sub-item except for the sub-items representing the first and second nodes where only a single latent dimension is measured. The number of parameters was equal because the only difference between the two models was that the two response style dimensions in the 3D model were joined into one in the 2D model), 154 parameters for the 1D model (40 sub-item $\beta$ parameters and 30 $\alpha$ parameters – the sub-items representing the first node were not measuring any latent dimension) and 50 parameters for the GRM model (four threshold parameters per item – 40 in total, and one $\alpha$ parameter per item – 10 in total). The number of parameters for the models does not depend on the Big Five trait that is considered.
<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>nPar</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>184269</td>
<td>100</td>
<td>184469</td>
<td>185166</td>
</tr>
<tr>
<td>2D</td>
<td>183810</td>
<td>100</td>
<td>184010</td>
<td>184708</td>
</tr>
<tr>
<td>1D</td>
<td>192593</td>
<td>70</td>
<td>192733</td>
<td>193221</td>
</tr>
<tr>
<td>GRM</td>
<td>186110</td>
<td>50</td>
<td>186210</td>
<td>186559</td>
</tr>
</tbody>
</table>

Table 25. Model fit of the proposed and alternative models (Agreeableness).

*The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.*

<table>
<thead>
<tr>
<th>Model</th>
<th>$\theta_R$</th>
<th>$\theta_A$</th>
<th>$\theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>.57</td>
<td>.69</td>
<td>.66</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>.66</td>
<td>.81</td>
</tr>
<tr>
<td>1D</td>
<td>-</td>
<td>.85</td>
<td>-</td>
</tr>
<tr>
<td>GRM</td>
<td>-</td>
<td>.88</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 26. EAP reliabilities of latent trait estimates (Agreeableness).
### Extraversion

Data for Extraversion contained responses of 7884 respondents to 10 items. The results of fitting all models are summarized in Table 28, with EAP reliabilities of the latent traits in Table 29, and the mean values and standard deviations of the discrimination parameter estimates for the 2D and 3D models are shown in Table 30:

<table>
<thead>
<tr>
<th>Node</th>
<th>$\Theta_R$</th>
<th>$\Theta_A$</th>
<th>$\Theta_{ERS}$</th>
<th>$\Theta_A$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>1.23 (0.44)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.03 (0.33)</td>
</tr>
<tr>
<td>N2</td>
<td>0</td>
<td>2.50 (1.08)</td>
<td>0</td>
<td>2.54 (1.09)</td>
<td>0</td>
</tr>
<tr>
<td>N3</td>
<td>0</td>
<td>0.60 (0.53)</td>
<td>-1.50 (0.42)</td>
<td>0.71 (0.38)</td>
<td>-1.50 (0.53)</td>
</tr>
<tr>
<td>N4</td>
<td>0</td>
<td>0.93 (0.55)</td>
<td>1.50 (0.45)</td>
<td>0.45 (0.41)</td>
<td>1.55 (0.54)</td>
</tr>
</tbody>
</table>

Table 27. Summary of discrimination parameter estimates at each node (Agreeableness). Means are written in bold, while standard deviations are in parentheses and italicized.
<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>nPar</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>204519</td>
<td>100</td>
<td>204719</td>
<td>205416</td>
</tr>
<tr>
<td>2D</td>
<td>204178</td>
<td>100</td>
<td>204378</td>
<td>205075</td>
</tr>
<tr>
<td>1D</td>
<td>212137</td>
<td>70</td>
<td>212277</td>
<td>212765</td>
</tr>
<tr>
<td>GRM</td>
<td>203848</td>
<td>50</td>
<td>203949</td>
<td>204297</td>
</tr>
</tbody>
</table>

Table 28. Model fit of the proposed and alternative models (Extraversion). The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\Theta_R$</th>
<th>$\Theta_E$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>.54</td>
<td>.85</td>
<td>.63</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>.84</td>
<td>.74</td>
</tr>
<tr>
<td>1D</td>
<td>-</td>
<td>.89</td>
<td>-</td>
</tr>
<tr>
<td>GRM</td>
<td>-</td>
<td>.92</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 29. EAP reliabilities of latent trait estimates (Extraversion).
Table 30. Summary of discrimination parameter estimates at each node (Extraversion). *Means are written in bold, while standard deviations are in parentheses and italicized.*

Conscientiousness

Data for Conscientiousness contained responses of 7884 respondents to 10 items. The results of fitting all models are summarized in Table 31, with EAP reliabilities of the latent traits in Table 32, and the mean values and standard deviations of the discrimination parameter estimates for the two best fitting models are shown in Table 33:

Table 31. Model fit of the proposed and alternative models (Conscientiousness). *The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.*
Table 32. EAP reliabilities of latent trait estimates (Conscientiousness).

<table>
<thead>
<tr>
<th>Model</th>
<th>$\theta_R$</th>
<th>$\theta_C$</th>
<th>$\theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>.51</td>
<td>.78</td>
<td>.61</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>.77</td>
<td>.72</td>
</tr>
<tr>
<td>1D</td>
<td>-</td>
<td>.81</td>
<td>-</td>
</tr>
<tr>
<td>GRM</td>
<td>-</td>
<td>.86</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 33. Summary of discrimination parameter estimates at each node (Conscientiousness).

Means are written in bold, while standard deviations are in parentheses and italicized.

Neuroticism

Data for Neuroticism contained responses of 7884 respondents to 10 items. The results of fitting all models are summarized in Table 34, with EAP reliabilities of the latent traits in Table 35, and the mean values and standard deviations of the discrimination parameter estimates for the two best fitting models are shown in Table 36:
<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>nPar</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>208554</td>
<td>100</td>
<td>208754</td>
<td>209451</td>
</tr>
<tr>
<td>2D</td>
<td>208410</td>
<td>100</td>
<td>208610</td>
<td>209307</td>
</tr>
<tr>
<td>1D</td>
<td>217399</td>
<td>70</td>
<td>217539</td>
<td>218027</td>
</tr>
<tr>
<td>GRM</td>
<td>211210</td>
<td>50</td>
<td>211311</td>
<td>211659</td>
</tr>
</tbody>
</table>

Table 34. Model fit of the proposed and alternative models (Neuroticism). The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\Theta_R$</th>
<th>$\Theta_N$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>.54</td>
<td>.83</td>
<td>.65</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>.82</td>
<td>.75</td>
</tr>
<tr>
<td>1D</td>
<td>-</td>
<td>.87</td>
<td>-</td>
</tr>
<tr>
<td>GRM</td>
<td>-</td>
<td>.91</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 35. EAP reliabilities of latent trait estimates (Neuroticism).
Table 36. Summary of discrimination parameter estimates at each node (Neuroticism). Means are written in bold, while standard deviations are in parentheses and italicized.

**Openness to Experience**

Data for Openness contained responses of 7884 respondents to 10 items. The results of fitting all models are summarized in Table 37, with EAP reliabilities of the latent traits in Table 38, and the mean values and standard deviations of the discrimination parameter estimates for the two best fitting models are shown in Table 39:

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>nPar</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>186775</td>
<td>100</td>
<td>186975</td>
<td>187673</td>
</tr>
<tr>
<td>2D</td>
<td>186046</td>
<td>100</td>
<td>186246</td>
<td>186944</td>
</tr>
<tr>
<td>1D</td>
<td>194099</td>
<td>70</td>
<td>194239</td>
<td>194727</td>
</tr>
<tr>
<td>GRM</td>
<td>188192</td>
<td>50</td>
<td>188292</td>
<td>188641</td>
</tr>
</tbody>
</table>

Table 37. Model fit of the proposed and alternative models (Openness). The table shows the model fit of all evaluated models, with the model abbreviation, deviance (-2*Log-likelihood), number of estimated parameters and two information criteria.
<table>
<thead>
<tr>
<th>Model</th>
<th>$\Theta_R$</th>
<th>$\Theta_O$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>.56</td>
<td>.67</td>
<td>.64</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>.66</td>
<td>.79</td>
</tr>
<tr>
<td>1D</td>
<td>-</td>
<td>.77</td>
<td>-</td>
</tr>
<tr>
<td>GRM</td>
<td>-</td>
<td>.83</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 38. EAP reliabilities of latent trait estimates (Openness).

<table>
<thead>
<tr>
<th>N</th>
<th>$\Theta_R$</th>
<th>$\Theta_O$</th>
<th>$\Theta_{ERS}$</th>
<th>$\Theta_O$</th>
<th>$\Theta_{ERS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>1.10 (0.35)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.95 (0.34)</td>
</tr>
<tr>
<td>N2</td>
<td>0</td>
<td>1.84 (0.85)</td>
<td>0</td>
<td>1.88 (1.05)</td>
<td>0</td>
</tr>
<tr>
<td>N3</td>
<td>0</td>
<td>0.75 (0.43)</td>
<td>-1.43 (0.32)</td>
<td>0.93 (0.56)</td>
<td>-1.36 (0.40)</td>
</tr>
<tr>
<td>N4</td>
<td>0</td>
<td>0.79 (0.41)</td>
<td>1.29 (0.39)</td>
<td>0.52 (0.52)</td>
<td>1.38 (0.49)</td>
</tr>
</tbody>
</table>

Table 39. Summary of discrimination parameter estimates at each node (Openness). *Means are written in bold, while standard deviations are in parentheses and italicized.*

The results of the above part of Study 2, together with the results of the following part, are summarized and discussed in the *Summary and discussion of the results* section on page 90 and then further with the results of Study 1 in the *General discussion* section at the end of the thesis on page 93.
As in the previous study, two structural models were fit on the EAP score estimates of each modeled latent variable for each person obtained from either the 2D or the 3D model. The data for each person contained an EAP score estimate for all the Big Five personality traits (used as random predictors in this model), five extreme response style EAP scores and, in case of estimates from the 3D model, five non-neutral responding EAP scores. Both the extreme response style scores and the non-neutral responding scores were used in a measurement model as variables loading on a (higher-order) latent variable. These higher-order latent variables, representing general response style factors, were allowed to correlate (which only applies to the model of the 3D estimates) and were regressed on the Big Five personality trait estimates. Each Big Five personality trait estimate was allowed to correlate with the residuals of response style estimates coming from the same IRTree model (so, for example, the Extroversion EAP estimate was allowed to correlate with the residuals of non-neutral responding and extreme response style estimates from the Extroversion 3D IRTree model). Finally, in case of the SEM model using the 3D IRTree estimates, the residuals of non-neutral responding and extreme response style estimates from the same Big Five IRTree model were allowed to correlate as well. The models were fit using normal maximum likelihood estimation (with conventional standard errors computed using the observed information matrix) in the package lavaan for R.
2D model

For the structural model based on latent score estimates from the 2D IRTree model, the (non-adjusted) minimum function test statistic was $\chi^2(20) = 825.1$, $p < 0.01$; RMSEA = 0.071 (90% CI = 0.067; 0.076), sRMR = 0.041, TLI = 0.86, CFI = 0.94. The path diagram of the model is shown in Figure 9. Note that only the regression parameters, factor loadings, residual variances and the latent variable variance (all standardized) are shown in Figure 9. In order to reduce clutter, the covariance parameters of the exogenous variables are shown in Table 40 below.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>E</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>.07</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.17</td>
<td>.13</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-.04</td>
<td>-.26</td>
<td>-.24</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>-.00*</td>
<td>.00*</td>
<td>.06</td>
<td>-.05</td>
<td>1</td>
</tr>
<tr>
<td>BF_E</td>
<td>.48</td>
<td>.29</td>
<td>-.05</td>
<td>-.07</td>
<td>.43</td>
</tr>
</tbody>
</table>

Table 40. Covariances of the exogenous variables in the structural model, 2D model.
Each of the 5 ERS score residuals was allowed to co-vary only with the Big Five personality trait score from the same IRTree model – thus, the BF,E row contains the covariances of the Big Five personality trait score in the given column with the respective ERS score residual.
* = $p > 0.05$. 

86
Figure 9. Structural model of the EAP estimates from Study 2, 2D model. 

The variables in this model are labeled using the first letter of the respective Big Five personality trait from the model of which they were extracted. The single letter labels refer to estimates of the personality trait in question, while labels ending with “_E” refer to extreme response style (e.g., “A” denotes Agreeableness and “A_E” denotes extreme response style, both as modeled in the Agreeableness model). The latent variable “ERS” refers to extreme response style. For the italicized parameter estimates, p > 0.05.
For the structural model based on latent score estimates from the 3D IRTree model, the (non-adjusted) minimum function test statistic was $\chi^2(59) = 923.3$, $p < 0.01$; RMSEA = 0.043 (90% CI = 0.041; 0.046), sRMR = 0.034, TLI = 0.94, CFI = 0.96. The path diagram of the model is shown in Figure 10. Note that only the regression parameters, factor loadings, residual variances and the latent variable variances (all standardized) are shown in Figure 10. In order to reduce clutter, the covariance parameters of the exogenous variables are shown in Tables 41 and 42 below.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>E</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>.08</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.20</td>
<td>.13</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-.04</td>
<td>-.25</td>
<td>-.24</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>.02*</td>
<td>-.01*</td>
<td>.08</td>
<td>-.06</td>
<td>1</td>
</tr>
<tr>
<td>BF_E</td>
<td>.26</td>
<td>.08</td>
<td>-.03</td>
<td>-.07</td>
<td>.20</td>
</tr>
<tr>
<td>BF_R</td>
<td>.45</td>
<td>.31</td>
<td>-.02*</td>
<td>-.01*</td>
<td>.42</td>
</tr>
</tbody>
</table>

Table 41. Covariances of the personality trait scores in the structural model, 3D model. Each of the 5 pairs of ERS and non-neutral responding score residuals was allowed to covary only with the Big Five personality trait score from the same IRTree model – thus, the BF_E row contains the covariances of the Big Five personality trait score in the given column with the respective ERS score residual, and the same holds for the BF_R row of non-neutral responding score covariances. * = $p > 0.05$. 

88
Figure 10. Structural model of the EAP estimates from Study 2, 3D model. The variables are labeled in an identical way to Figure 9. Additionally, labels ending with “_R” refer to non-neutral responding (e.g., “A_R” denotes non-neutral responding as modeled in the Agreeableness model). The latent variable “R” refers to non-neutral responding. For the italicized parameter estimates, p > 0.05.
Table 42. Covariances of the response style score residuals in the structural model, 3D model.

<table>
<thead>
<tr>
<th></th>
<th>A_R</th>
<th>C_R</th>
<th>E_R</th>
<th>N_R</th>
<th>O_R</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_E</td>
<td>.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_E</td>
<td></td>
<td>.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E_E</td>
<td></td>
<td></td>
<td>.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_E</td>
<td></td>
<td></td>
<td></td>
<td>.25</td>
<td></td>
</tr>
<tr>
<td>O_E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.31</td>
</tr>
</tbody>
</table>

Summary and discussion of the results

Similar to Study 1, the multidimensional models (2D and 3D) were overall superior in fit to the unidimensional models (1D and GRM), while, again, the fit of the 2D model was slightly better than that of the 3D model. The only exception to this pattern was the Extraversion data for which the GRM exhibited the best fit. The reasons for the better fit of the 2D model can be considered the same as in the previous study – although due to the consistently high correlation between the two general response style factors in the 3D structural models (see Study 1 and the section dedicated to the structural models below) the interpretation that it is primarily the forcing of non-neutral responding and ERS to be independent in the IRTree models what hinders model fit becomes ever more plausible.
Overall, the pattern of reliabilities was consistent with the one found in Study 1, with reliabilities of the Big Five trait estimates being higher for the simpler models (1D and GRM). For the ERS estimates, the range of reliabilities is between .65 and .80 and the two-dimensional IRTree model yielded overall more reliable estimates.

The discrimination parameters for the Big Five personality dimensions followed a similar pattern as in the previous study, with their size decreasing as the tree progresses to its bottom. Interestingly, the values of discrimination parameter estimates were consistently larger than those in Study 1, and their values were more dispersed around the mean for Node 2 compared to Nodes 3 and 4. The most likely reason for this pattern was the different response scale format (5-point response scale instead of a 7-point response scale), as the sub-responses in a shorter response scale can be hypothesized to discriminate better between persons of different locations on the latent continuum. Study 2, however, did not allow for comparing the discrimination parameter estimates of higher nodes vs. lower nodes, simply because the number of final response categories (and thus, in our case, the number of nodes) was smaller.

The rationale for the structural equation models of EAP estimates from both the 2D and 3D models was the same as in Study 1. Both structural models fitted well and the fit of each was similar to its Study 1 counterpart. Although the trait-specific ERS estimates still had fairly high loadings on the general ERS factor in both structural models, those loadings were overall lower compared to the previous study – the same can be concluded for loadings of the trait-specific non-neutral responding estimates on the corresponding general factor in the 3D structural model. Again, the pattern of correlations
between the unique variances of the response style EAP estimates and the Big Five personality trait estimates can be observed. In some cases, the unexplained variance of the non-neutral responding estimates was correlated with the Big Five personality trait estimate from the same IRTree model (in case of Agreeableness, Conscientiousness and Openness to Experience), which was less pronounced for the trait-specific ERS estimates in the 3D model. While the unique variances of trait-specific ERS estimates did correlate more with the Big Five personality trait estimates in case of the 2D model, the comparison of the two SEM models indicates that modeling the non-neutral responding and ERS as a single latent variable in the 2D model was the major source of these correlations. Again, as in Study 1, it can be observed that for some Big Five traits (such as for Agreeableness, Conscientiousness or Openness in both the 2D and 3D model, for example), the trait’s relationship with a general response style factor (either ERS or non-neutral responding) differed in size or even in sign from its relationship with the corresponding trait-specific response style estimate residual. Such a difference could reflect a dissimilar relationship between the personality traits and general versus trait-specific response styles.

Finally, the correlation between the two general response style factors was consistent with findings from the previous study – the two factors were correlated, although an independence between their trait-specific counterparts was imposed in the IRTree models. As explained before, this is not impossible, because constraining two latent variables to be independent within a model does not necessarily imply that their estimates will be uncorrelated. Furthermore, Agreeableness and Extraversion exhibited a
relationship with ERS in the 2D model similar to Study 1, but the relationship of Extraversion and ERS in the 3D model is weaker in this study.

General discussion

One of the purposes of the two studies in this thesis was to expand on a similar work done by Khorramdel and von Davier (2014) and Plieninger and Meiser (2014) using a like-minded approach based on item response trees. The approaches employed by these authors have proved fruitful, however, they did suffer from a number of limitations. Perhaps most importantly, the previous approaches assumed that the selection of extreme response categories is purely driven by extreme response style, without any influence from the measured substantive trait. Secondly, Plieninger and Meiser (2014) assumed that all pseudo-items corresponding to extreme responses are identically relevant for measuring ERS (by modeling sub-responses using a Rasch model) and that the choice of the second most extreme response categories ("2" and "6" on a 7-point scale) is not at all indicative of ERS. By modeling responses to pseudo-items with a multidimensional 2PL model, these limitations can be overcome – the relative contributions of extreme response style and the measured substantive trait to the pseudo-item responses can be evaluated via comparison of their respective discrimination parameters. Likewise, responses to the second most extreme response categories (if available) can be modeled as being influenced by ERS and the degree of this influence can be quantified by the discrimination parameters.
Based on the results of both studies in this thesis, it seems that this approach performs well. The relative contributions of ERS and the measured substantive trait to the selection of a particular response category are respectively increasing and decreasing as the response category gets more extreme. Moreover, by including latent variable estimates from multiple IRTree models in a structural model, the extent to which the extreme response style (as measured in each of the separate IRTree models) represents a general, trait-like response tendency can be evaluated.

As can be concluded from results of the above studies, such a general response style was indeed largely modeled in this way, although with a significant specific component. In some cases, the relationship between a particular Big Five personality trait and the general response style factor differed in size or even sign from the former’s relationship with the specific (residual) component of the response style estimate coming from the same IRTree model. This observation could mean that the relationships between a trait, a general response style factor and trait-specific response style factor might differ.

Potentially, though, this difference could also reflect that some proportion of substantive trait variance was incorrectly and systematically modeled as *extreme response style* along with true response style variance under this modeling approach (causing the trait-specific response style residual to correlate with its corresponding trait), which would constitute a methodological artifact. This seems to be more noticeable in the structural model for 2D IRTree estimates, where non-neutral responding was considered to comprise a single latent dimension along with ERS. When modeled independently (as was the case in the 3D models) it became clear that it was primarily the non-neutral responding
estimates’ residual variances that covaried with the substantive traits. This could be considered one of the main justifications for preferring the 3D model over the 2D model for the purposes of ERS modeling. Valid and efficient modeling of a general non-neutral responding tendency overall proved more difficult under this particular approach, since the substantive personality trait could not have been sensibly measured (and thus its variance controlled for) at the sub-item which loaded on non-neutral responding, as explained earlier.

Other limitations of this approach are also obvious – for example, if the tree models are designed in the way as they were in this thesis, imposing a constraint of independence between the substantive personality trait and ERS becomes a necessity for identification purposes. This is not ideal since it represents an artificial assumption of orthogonality of latent variables, although these might in many cases be related to a significant degree. This limitation might be overcome by formulating the discrimination matrix differently (such as not considering the most extreme sub-items to load on the substantive trait), by imposing constraints on discrimination parameters or by modeling a general ERS jointly with all the substantive personality traits. However, none of these potential solutions is flawless and each represents a certain trade-off between different model constraints and assumptions.

Some of the alternative approaches discussed earlier in this thesis do not require the constraint of independence either, however, they do suffer from other limitations. For example, the 2PL model used by De Jong, Steenkamp, Fox and Baumgartner (2008) does not allow for simultaneous modeling of ERS and substantive dimensions. Bolt and Johnson’s (2009) approach is an inherently exploratory use of the multidimensional
nominal response model while its confirmatory counterpart (Bolt & Newton, 2011) requires (arbitrary) constraints on the response category slope parameters. The model proposed by Falk and Cai (2016) allows item (not category) slope parameters to be freely estimated, but requires an a priori formulation of scoring functions. Any model limitations should be pondered and their respective advantages and disadvantages taken into account when deciding what model to use – choosing between models is in many cases ultimately about choosing what assumptions one is willing to make.

One important potential limitation of the IRTree approach lies in the mechanism by which the original responses are expanded into sub-responses. As mentioned before, some sub-responses are coded as not observed (missing) since the corresponding pseudo-items were not reached by the respondent along the way to the final response. This means that for some pseudo-items, the number of observed responses can be greatly reduced and constitute just a fraction of the number of observed responses to the original item. The extent to which this poses an issue in estimating the parameters of such sub-items should be a subject of further scrutiny.

Future directions

Although this thesis presented an attempt to jointly model primarily extreme response style and substantive traits using the IRTree approach, the same approach could be readily employed for modeling other response styles as well – namely, acquiescence, disacquiescence, midpoint responding or social desirability (Baumgartner & Steenkamp,
Midpoint responding (or, rather, its inverse – non-neutral responding) was already modeled in this thesis, however, as explained above, this proved rather difficult due to design choices stemming from the fact that modeling it was only secondary to modeling ERS. A different formulation of the tree model (such as a linear tree) might prove more fruitful in this regard.

Acquiescence and disacquiescence could be modeled fairly easily using a nested tree with multidimensional sub-items, as used in the two studies presented here. In such case, the sub-items representing choice of agreement-indicating or disagreement-indicating response categories would be indicative of acquiescence or disacquiescence, respectively. In order to model social desirability, an a priori decision would need to be made on designating sub-items that represent a choice between a socially desirable response and response that is not socially desirable or neutral in terms of desirability. Clearly, multiple application of item response trees for measurement of response styles are feasible and their investigation will hopefully be encouraged by further findings.
References


101


