Detection, Avoidance, and Compensation – Three Studies on Extreme Response Style

Florian Pargent

München 2017

"Hello darkness, my old friend. I've come to talk with you again."

— psychometrics and data

Detection, Avoidance, and Compensation – Three Studies on Extreme Response Style

Florian Pargent

Dissertation
an der Fakultät für Psychologie und Pädagogik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Florian Pargent
aus Kronach

München, den 27.04.2017

Erstgutachter: Prof. Dr. Markus Bühner
Zweitgutachter: Prof. Dr. Moritz Heene
Tag der mündlichen Prüfung: 12.07.2017

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2PL model	Two Parameter Logistic Item Response Model
A4	Compliance
ACC	Accuracy
ACHI	Achievement Orientation
AMM	Allocentric/Mental Map
ARS	Aquiescence Response Style
BIC	Bayesian Information Criterion
BMI	Body Mass Index
C2	Order
C5	Self-Discipline
CART	Classification And Regression Trees
CI	Confidence Interval
CTT	Classical Test Theory
CV	Cross-Validation
DIF	Differential Item Functioning
E3	Assertiveness
E4	Activity
E5	Excitement-Seeking
EAS	Extremer Antwortstil
ERS	Extreme Response Style

FPI-R	Revised Freiburger Persöhnlichkeitsinventar
FRS	Fragebogen Räumliche Strategien
GEO	Global/Egocentric Orientation
ICC	Item Characteristic Curve
IMP	Impulsivity
IR	Impurity Reduction
IRT	Item Response Theory
LMU	Ludwig-Maximilians-Universität München
MLE	Maximum Likelihood Estimation
MMCE	Mean Misclassification Error
MMPI-2	Revised Minnesota Multiphasic Personality Inventory
MRERS	Extreme Response Style Measure Based on a Polytomous Mixed Rasch Model
MRS	Mid Response Style
MSE	Mean Squared Error
NA	Negative Affect
NEO-FFI	NEO Five-Factor Inventory
NEO-PI-R	Revised NEO Personality Inventory
N2	Angry Hostility
N5	Impulsivity
O3	Openness to Feelings
OOB	Out-Of-Bag
ORD	Order
PA	Positive Affetc
PANAS	Positive and Negative Affect Schedule
PCB	Partial Credit Baum

PCM	Partial Credit Model
PC tree	Partial Credit Tree
PISA	Program for International Student Assessment
pMLE	Penalized Maximum Likelihood Estimation
RF	Random Forest
RIRS	Representative Indicators of Response Styles
RIRSMAC	Representative Indicators Response style Means and Covariance Structure
RMSE	Root Mean Squared Error
SENS	Sensitivity
SHY	Shyness
SPEC	Specificity
SRERS	Self-Reported Extreme Response Style
ZIS	Zusammenstellung Sozialwissenschaftlicher Items und Skalen
ZPID	Leibniz-Zentrum für Psychologische Information und Dokumentation

Zusammenfassung

Einleitung Menschen unterscheiden sich darin, wie sie die Antwortkategorien von Fragebogenitems mit Likert Skalenformat verwenden (Vaerenbergh & Thomas, 2013). Studien legen nahe, dass es sich bei diesen Antwortstilen um Personeneigenschaften handelt, die zeitlich über lange Zeit stabil bleiben (Weijters, Geuens, & Schillewaert, 2010b; Wetzel, Lüdtke, Zettler, & Böhnke, 2016) und gleichermaßen in Fragenbögen zu unterschiedlichen Konstrukten auftreten (Wetzel, Carstensen, & Böhnke, 2013). Unabhängig von der zu messenden latenten Eigenschaft tendieren manche Personen dazu extremere Kategorien anzukreuzen, während andere Personen eher mittlere Kategorien bevorzugen. Man bezeichnet dieses Phänomen als Extremen Antwortstil (EAS). Bleibt EAS unberücksichtigt, kann dies dazu führen, dass die Varianzen von Skalenwerten (Wetzel & Carstensen, 2015) sowie Korrelationen zwischen psychologischen Konstrukten (Baumgartner & Steenkamp, 2001) überschätzt werden.

Die beiden häufigsten Modellierungansätze um EAS zu erkennen, sind das ordinale Mixed Raschmodell (Rost, 1991) und Antwortstilindizes aus heterogenen Items (Greenleaf, 1992). Das Mixed Raschmodell liefert eine attraktive Operationalisierung von EAS anhand unterschiedlicher Schwellenabstände im Partial Credit Modell (Masters, 1982). Es erlaubt eine Trennung von Inhalt und Antwortstil, allein mithilfe der Items aus der betreffenden Skala. Allerdings ist eine Konfundierung des gemessenen Antwortstils durch andere Konstrukte nicht auszuschließen und das Modell nimmt implizit an, dass es sich bei EAS um eine kategoriale Eigenschaft handelt. Antwortstilindizes aus heterogenen Items liefern eine bessere Interpretierbarkeit von EAS durch eine objektive Trennung von Inhalt und Antwortstil, setzen jedoch zusätzliche Items für die Berechnung des Index voraus.

Die vorliegende Arbeit beinhaltet drei Studien zu EAS. Diese bedienen sich moderner Verfahren der Item Response Theorie (IRT) und der prädiktiven Modellierung, um eine neue Sichtweise auf Themen zu liefern, die die Psychometrie seit mehr als 60 Jahren beschäftigen (Cronbach, 1950). In der ersten Studie wird eine neue Methode zur Erkennung von EAS vorgestellt, die viele Vorteile von Mixed Raschmodellen und Antwortstilindizes miteinander vereint. Die zweite Studie untersucht erstmals die häufig getroffene Annahme, dass EAS durch die Verwendung von Items mit nur zwei Antwortkategorien vermieden werden kann. Schließlich wird in der dritten Studie der Versuch unternommen, EAS durch spezielle Instruktionen auszugleichen, die an den individuellen Antwortstil des Probanden angepasst werden.

Studie 1 Partial Credit Bäume (PCBs; Komboz, Strobl, & Zeileis, 2016) sind ein kürzlich entwickeltes Verfahren zur Erkennung von Differential Item Functioning (DIF), basierend auf kategorialen oder metrischen Kovariablen. Durch die Verwendung eines extremen Antwortstilindex aus heterogenen Items als Kovariable, stellen PCBs ein objektives Verfahren zur Entdeckung von EAS dar, das die Operationalisierung von EAS anhand unterschiedlicher Schwellenabstände beibehält. Indirekt erlaubt es das Verfahren empirisch zu überprüfen, ob es sich bei EAS eher um ein kontinuierliches oder um ein diskretes Konstrukt mit mehreren getrennten Personengruppen handelt. Mit PCBs untersucht wurden Persönlichkeitsfacetten der Big Five im deutschen nicht klinischen Normdatensatz des revidierten NEO-Persönlichkeitsinventars (NEO-PI-R; Ostendorf & Angleitner, 2004), sowie Subskalen des deutschen Positive and Negative Affect Schedules (PANAS; Krohne, Egloff, Kohlmann, & Tausch, 1996) und des Fragebogens Räumliche Strategien (FRS; Münzer & Hölscher, 2011), erhoben im Rahmen des Panels des GESIS-Leibniz-Institut für Sozialwissenschaften (GESIS, 2015). Analysiert wurden die Daten von 11714 Personen im NEO-PI-R Normdatensatz und 3835 Personen aus dem GESIS Panel. In allen betrachteten Skalen zeigte sich ein kontinuierliches Muster von EAS. Dieses zeichnet sich durch engere Schwellenabstände und höhere Antwortwahrscheinlichkeiten für die äußeren Kategorien in Gruppen mit hohen Werten im Antwortstilindex aus. Vor allem im NEO-PI-R Datensatz zeigte sich ein prägnantes Muster, mit einer großen Anzahl von identifizierten Antwortstilgruppen. Dieser Befund legt nahe, dass es sich bei EAS um eine kontinuierliche Persönlichkeitseigenschaft handelt. Weiterhin zeigen die Ergebnisse aus den Daten des längsschnittlichen GESIS Panels die zeitliche Stabilität von EAS. Obwohl durch die Verwendung von PCBs in Kombination mit einem Antwortstilindex EAS objektiv erkannt werden kann, eignet sich das Verfahren nicht zur statistischen Kontrolle von EAS. Spezielle IRT Modelle (Jin & Wang, 2014; Tutz, Schauberger, & Berger, 2016), die EAS in Form eines kontinuierlichen Personenparameters modellieren, scheinen dazu besser geeignet. Gegen deren praktische Verwendung spricht momentan jedoch das Vorliegen von DIF bezüglich demografischer Variablen wie Geschlecht und Alter in beiden analysierten Datensätzen.

Studie 2 Die Ergebnisse der ersten Studie legen nahe, dass es sich bei EAS um ein allgegenwärtiges Problem bei Likert Items mit mehrkategoriellem Antwortformat handelt. Items mit dichotomem Antwortformat werden im Rahmen der psychologischen Testkonstruktion als wirkungsvolles Mittel diskutiert um EAS zu vermeiden (Bühner, 2011). Obwohl dies plausibel erscheint, liegen dazu bisher keine empirischen Studien vor. Es kann jedoch nicht augeschlossen werden, dass EAS auch bei dichotomen Itemformaten unbemerkt eine Rolle spielt. Einige Studien deuten an, dass eine hohe Ausprägung von EAS auch eine generelle Präferenz für extreme Antworten in einem inhaltlichen Sinne beschreibt (Naemi, Beal, & Payne, 2009; Zettler, Lang, Hülsheger, & Hilbig, 2015). Es erscheint daher möglich, dass sich EAS auch bei dichotomen Items auf die Schwellen in IRT Modellen auswirkt. Mithilfe von Rasch Bäumen (Strobl, Kopf, & Zeileis, 2013) und DIF Lasso Modellen (Tutz & Schauberger, 2015) kann der Einfluss von EAS auf die Schwellen im dichotomen Raschmodell untersucht werden. In einem Papier und Bleistift Fragebogen beantworteten

429 Psychologiestudenten die dichotomen Skalen Schüchternheit aus dem deutschen revidierten Minnesota Multiphasic Personality Inventory (MMPI-2; Engel & Hathaway, 2003) und Leistungsorientierung aus dem revidierten Freiburger Persönlichkeitsinventar (FPI-R; Fahrenberg, Hampel, & Selg, 2001). EAS wurde auf drei Arten operationalisiert: Durch einen Antwortstilindex aus heterogenen Items mit mehrkatigoriellem Antwortformat, mithilfe der dichotomen Klassifizierung eines ordinalen Mixed Raschmodells basierend auf der Facette Ordentlichkeit des NEO-PI-R und durch eine dichotome Selbsteinschätzung zu extremem Antwortverhalten. Sowohl bei der Analyse mit Rasch Bäumen als auch mit dem DIF Lasso zeigte sich in keiner der beiden dichotomen Skalen ein Einfluss der drei verwendeten Antwortstilindikatoren auf die Schwellen im Raschmodell. Kontrollanalysen bestätigten jedoch einen Effekt von EAS auf die ordinale Ordentlichkeitsskala, wodurch die Validität der drei Antwortstilindikatoren nachgewiesen werden konnte. Wie auch in Studie 1 wurde in beiden dichotomen Skalen für einige Items DIF basierend auf Geschlecht und Alter erkannt. Die Ergebnisse legen nahe, dass die Konstruktion von psychologischen Fragebögen mit dichotomen Itemformaten eine wirksame Strategie darstellt, um EAS zu vermeiden. Da viele Anwender dichotomen Antwortformaten bisher skeptisch gegenüber stehen, werden Vor- und Nachteile diskutiert.

Studie 3 Da sich mehrkategorielle Antwortformate in der Psychologie großer Beliebtheit erfreuen, ist eine Strategie besonders wünschenswert die den Einfluss von EAS bei ordinalen Skalen auszugleichen vermag. Bereits Cronbach (1950) spielte mit dem Gedanken, Antwortstile durch Probandentrainings und spezielle Instruktionen zu kompensieren. Angelehnt an das kognitive Prozessmodell von Tourangeau, Rips, and Rasinski (2000) wird eine Instruktion vorgeschlagen, bei der Probanden ihre Itemantwort in eine extremere oder weniger extreme Richtung anpassen sollen, wenn sie sich zwischen zwei Kategorien nicht entscheiden können. Um die Effektivität der Instruktionen hinsichtlich der Kompensation von EAS zu untersuchen, wird die Güte der Vorhersage von konstruktrelevanten Verhaltensweisen mithilfe von Itemantworten unter verschiedenen Instruktionsbedingungen betrachtet: Wir nehmen an, dass die Kriteriumsvalidität zunimmt, falls der Einfluss von EAS durch passende Instruktionen ausgeglichen werden kann. Es erfolgt eine kurze Einführung in die moderne prädiktive Modellierung mit Random Forest Modellen (Breiman, 2001) einschließlich wichtiger Grundprinzipien wie Kreuzvalidierung und Overfitting (Hastie, Tibshirani, & Friedman, 2009).

In einem Onlinefragebogen beantworteten 788 Probanden dreimal die Facetten Impulsivität und Ordentlichkeit aus dem NEO-PI-R unter verschiedenen Instruktionen. Zunächst absolvierten alle Probanden beide Skalen unter Standardinstruktionen. In randomisierter Reihenfolge erfolgte in der Experimentalgruppe die zweite und dritte Beantwortung entweder unter einer Instruktion zum extremen Ankreuzen oder einer Instruktion zum mittleren Ankreuzen. Eine Kontrollgruppe beantwortete die Skalen in der zweiten und dritten Runde erneut unter neutraler Instruktion. Als Kriteriumsvariablen in prädiktiven Modellen dienten Selbstauskunftsfragen zu konkret beobachtbaren impulsiven bzw. ordentlichen Verhaltensweisen. Weiterhin erhoben wurden heterogene Items zur Berechnung

eines Antwortstilindex und eine dichotome Selbsteinschätzung zu extremen Antworttendenzen. Mithilfe der beiden Antwortstilindikatoren wurde bestimmt, unter welcher Instruktion ein Proband seine eigenen Antworttendenzen kompensiert bzw. verschlimmert. Itemantworten aus der Kompensations- und Verschlechterungsbedingung wurden dann in neue Pseudodatensätze zusammengefasst, um diese hinsichtlich der Vorhersage der Verhaltenskriterien vergleichen zu können. Generell konnte ein Effekt der Instruktionen auf die Itemantworten nachgewiesen werden. Jedoch zeigten sich zwischen der Kompensations-, Verschlechterungs- und Kontrollbedingung keine Unterschiede in der Vorhersagegüte der kreuzvalidierten Random Forest Modelle mit den Impulsivitäts- und Ordentlichkeitsitems als Prädiktoren. In weiteren Analysen mit PCBs konnte auch in der Kompensationsbedingung ein Einfluss von EAS festgestellt werden. Dieser war nicht eindeutig schwächer als in der Verschlechterungsbedingung. In prädiktiven Modellen mit den Itemantworten unter Standardinstruktionen und demografischen Variablen als Prädiktoren zeigte sich keine Verbesserung der Vorhersagegüte durch die Hinzunahme der beiden Antwortstilindikatoren in das Modell. Dasselbe Ergebnis lieferte ein Maß für die geschätzte Bedeutung der beiden Antwortstilindikatoren bei der Vorhersage (Out-Of-Bag Permutation Variable Importance). Individuelle Partial Prediction Plots (Goldstein, Kapelner, Bleich, & Pitkin, 2015) ließen bestenfalls einen kleinen Einfluss des Antwortstilindex auf die Vorhersagen mancher Kriterien vermuten. Obwohl mithilfe der Kontrollgruppe nachgewiesen werden konnte, dass durch die wiederholte Darbietung der Items Reihenfolgeeffekte auftraten, kann dies die Befunde nicht erklären. Die Ergebnisse stehen im Einklang mit zwei Simulationsstudien, die einen geringen Einfluss von EAS auf die Kriteriumsvalidität von psychologischen Fragebögen nahelegen (Plieninger, 2016; Wetzel, Böhnke, & Rose, 2016). Der Einfluss von EAS auf die Kriteriumsvalidität hat große praktische Relevanz und sollte daher in zukünftigen Studien weiter erforscht werden. Außerdem werden mögliche Verbesserungsvorschläge für die verwendeten Instruktionen diskutiert.

Diskussion Die dargestellten Analysen mit PCBs und Antwortstilindizes aus heterogenen Items legen nahe, dass EAS bei Items mit mehrkategoriellem Antwortformat allgegenwärtig ist und sich am besten als stabile Persönlichkeitseigenschaft mit kontinuierlicher Struktur beschreiben lässt. Die Vermutung, dass dichotome Antwortformate geeignet sind um EAS zu vermeiden, konnte mithilfe von Rasch Bäumen und DIF Lasso Modellen erstmals empirisch bestätigt werden. Die Kompensation von EAS mithilfe von speziellen Instruktion war in der hier dargestellten Form nicht erfolgreich. Dabei zeigt sich weiterer Forschungsbedarf hinsichtlich des Einfluss von EAS auf die Kriteriumsvalidität psychologischer Fragebögen.

Für die zukünftige Erforschung von EAS ergeben sich angelehnt an die präsentierten Studien zwei Ausrichtungen: Zum einen spielt die Untersuchung von EAS eine wichtige Rolle für die präzise Messung psychologischer Konstrukte. Um die Itemantworten in psychologischen Fragebögen angemessen statistisch zu beschreiben, müssen systematische Einflüsse wie Antwortstile mitmodelliert werden (Jin & Wang, 2014; Tutz et al., 2016). Dabei lohnt es sich auch Modellansätze zu betrachten, die nicht nur den Einfluss von Antwort-

stilen auf die Itemantworten treffend beschreiben, sondern auch kognitive Aspekte des Antwortprozesses berücksichtigen (Böckenholt, 2012; Zettler et al., 2015). Diese Multiprozessmodelle könnten wichtige Hinweise zu den Ursachen von Antwortstilen liefern. Zum anderen wird derzeit eine stärkere Ausrichtung der Psychologie auf prädiktive Fragestellungen diskutiert (Yarkoni & Westfall, 2017; Chapman, Weiss, & Duberstein, 2016). Dies geht zunehmend mit der Aneignung von Methoden aus dem Bereich der prädiktiven Modellierung einher (Hastie et al., 2009). Werden psychologische Konstrukte als Kriterien in prädiktiven Modellen verwendet, erscheint eine Kontrolle von Antwortstilen mithilfe spezieller IRT Modelle (Jin & Wang, 2014; Tutz et al., 2016) notwendig. Andererseits haben viele Bereiche der Psychologie das Ziel, mithilfe von Fragebögen praktisch relevante Außenkriterien vorherzusagen. In diesem Fall gilt es in Zukunft zu klären, ob die Vorhersagekraft von psychologischen Theorien durch die Aufnahme von Antwortstilindikatoren als Prädiktoren in nicht lineare prädiktive Modelle deutlich gesteigert werden kann.

Abstract

Extreme Response Style (ERS) describes individual differences in selecting extreme response options in Likert scale items, which are stable over time (Weijters et al., 2010b; Wetzel, Lüdtke, et al., 2016) and across different psychological constructs (Wetzel, Carstensen, & Böhnke, 2013). This thesis contains three empirical studies on the detection, avoidance, and compensation of ERS:

In the first study, we introduce a new method to detect ERS which uses an ERS index from heterogeneous items as covariate in partial credit trees (PC trees; Komboz et al., 2016). This approach combines the objectivity of ERS indices from heterogeneous items (Greenleaf, 1992) with the threshold interpretation of ERS known from analyses with the ordinal mixed-Rasch model (Rost, 1991). We analyzed personality facets of 11714 subjects from the German nonclinical normative sample of the Revised NEO Personality Inventory (NEO-PI-R; Ostendorf & Angleitner, 2004), and 3835 participants of the longitudinal panel of the GESIS - Leibniz-Institute for the Social Sciences (GESIS, 2015), who filled out the Positive and Negative Affect Schedule (Krohne et al., 1996), and the Questionnaire of Spatial Strategies (Münzer & Hölscher, 2011). ERS was detected in all analyzed scales. The resulting pattern suggests that ERS reflects a stable trait with a continuous structure.

In the second study, we investigate whether data from items with dichotomous response formats are unaffected by ERS, as has been assumed in the literature (Wetzel, Carstensen, & Böhnke, 2013). In a paper and pencil questionnaire, 429 German psychology students completed the Shyness scale from the Revised Minnesota Multiphasic Personality Inventory (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) and the Achievement Orientation scale from the Revised Freiburger Persöhnlichkeitsinventar (Fahrenberg et al., 2001). ERS was assessed by an ERS index from heterogeneous items, a binary ERS measure based on the classification of an ordinal mixed-Rasch model, and a binary self-report measure of ERS. ERS measures were used as covariates in Rasch trees (Strobl et al., 2013) and DIF Lasso models (Tutz & Schauberger, 2015) of the dichotomous scales. We did not find any effect of ERS on dichotomous item responses. Adopting dichotomous response formats seems to be a reasonable strategy to avoid ERS.

In the third study, we test whether instructions to give more or less extreme responses depending on participants' individual response tendencies, can counterbalance the impact of ERS. In an online questionnaire, 788 German subjects completed the Impulsivity and Order facets of the NEO-PI-R three times under different ERS instructions. In the first round, a standard instruction was used. Participants in the experimental group received

instructions for more or less extreme responses in the second and third round, while subjects in the control group responded under neutral instructions. ERS was measured by an ERS index from heterogeneous items and a self-report measure of ERS. Binary ERS classifications were used to create artificial datasets in which participants received an instruction which should either compensate or aggravate their individual response tendencies. Predictive performance of Random Forest models (Breiman, 2001), in which self-reported impulsive and orderly behaviors were predicted by the item responses, was compared between the compensation, aggravation, and control settings. No differences in predictive performance were observed between the settings. Likewise, PC tree analyses suggest that ERS was still present in the compensation setting. Including ERS measures as predictors did not increase predictive performance when items were answered under standard instructions. Our findings are in line with simulation studies suggesting that ERS has a small impact on applied psychological measurement (Plieninger, 2016; Wetzel, Böhnke, & Rose, 2016).

Future research on ERS could improve psychological measurements by considering continuous models of ERS (Jin & Wang, 2014; Tutz et al., 2016). In light of recent calls to turn psychology into a more predictive science (Yarkoni & Westfall, 2017; Chapman et al., 2016), investigating the impact of ERS on criterion validity should also have high priority.

Chapter 1

General Introduction

Individual tendencies of responding to survey items irrespective of item content, commonly referred to as response styles, still are an important topic in psychometric research. Under the term "response sets", Cronbach (1946) did seminal work on response bias in psychological measurements and raised many issues which bother the psychometric community up to this day. The term "response style" was first introduced by Jackson and Messick (1958), emphasizing that response styles might be trait-like constructs that are stable over time and consistent across situations (Cronbach, 1950).

While rigorous empirical evidence was lacking in Cronbach's time, current research has found that large components of response styles are stable over one (Weijters et al., 2010b) or even eight years (Wetzel, Lüdtke, et al., 2016), are consistent across different traits (Wetzel, Carstensen, & Böhnke, 2013), within a longer questionnaire (Weijters, Geuens, & Schillewaert, 2010a) or across different modes of data collection (Weijters, Schillewaert, & Geuens, 2008), and are comparable for scales with different numbers of response categories (Kieruj & Moors, 2013).

The two kinds of response styles which are the focus of current research are Extreme Response Style (ERS) and Aquiescence Response Style (ARS). ERS refers to the observation that people differ in their tendency to choose extreme response categories of Likert scale items, independent of their value on the primary latent trait that is measured by the items. ARS describes varying tendencies to agree with questionnaire items regardless of the item content. In this dissertation we examine ERS.[1] All methodology could in principle be applied to ARS or to any other response style postulated in the literature (for a current review on response styles, see Vaerenbergh & Thomas, 2013).

This thesis contains three empirical studies on ERS: In the first study, a new methodological approach to detect ERS is introduced. Two large datasets are used to illustrate the benefits of the technique. In the second study, we empirically investigate the widely held assumption that ERS can be avoided by using questionnaire items with dichotomous

[1] Henceforth the personal pronoun "we" will be used. Although all work was done by the sole author of this thesis, he hates forcing his writing into passive speech even more than referring to himself all the time. Since all of his work was heavily influenced by frequent discussions with his amazing supervisors, colleagues, and friends, writing in this form feels most natural to him.

response format. In the third study, we explore the possibility to compensate for the impact of ERS by giving subjects specific instructions which work against their own response tendencies.

Before presenting those studies, we discuss different definitions of ERS, mention fundamental challenges when studying ERS, describe the two most applied methods to detect ERS, and talk about the impact of ERS on psychological measurement as well as different strategies to mitigate those effects.

1.1 Definitions of Extreme Response Style

Conceptualizations of ERS found in the literature differ on at least three important attributes, which are rarely explicitly discussed: The meaning of extreme responding refers to which item responses are considered a symptom of ERS, the dimensionality of extreme responding refers to whether the phenomenon of ERS reflects one or more latent processes, and the continuity of extreme responding refers to whether ERS is described as a discrete or as a continuous variable.

Meaning of Extreme Responding The definition of an extreme response is not consistent across the literature. The majority of researchers define an extreme response as selecting the highest or lowest response category on a Likert scale (Vaerenbergh & Thomas, 2013). With this definition, ERS reflects a special preference for the most extreme categories and can be interpreted as a person's unconditional probability of selecting the highest or lowest category of a Likert scale (Greenleaf, 1992). Thus, the person's value on the latent trait measured by a specific item is not taken into account. ERS can also be thought of as the unconditional general extremeness of a person's response on Likert scale items (Jin & Wang, 2014; Tutz et al., 2016). In this view, the extremeness of an item response can be visualized as the location of the average item response, compared to the mid category of the response scale. Thus, selecting the highest or lowest category reflects a more extreme response than choosing the second highest or lowest category and so on. However, no specials role is assigned to the most extreme categories or to the middle category if it exists. Any item response carries information about ERS.

Dimensionality of Extreme Responding The majority of research distinguishes ERS from midpoint response style (MRS), the tendency to endorse the middle category of a rating scale (Baumgartner & Steenkamp, 2001; Weijters et al., 2008). In other conceptualizations, extreme response tendencies are considered a unidimensional trait. This can be unipolar with a balanced response pattern on the one side, but high preferences for the most extreme categories on the other side of the spectrum (Greenleaf, 1992; Wetzel, Böhnke, Carstensen, Ziegler, & Ostendorf, 2013). It can also be bipolar, ranging from a high preference for the middle category, to high preferences for the most extreme response categories (Jin & Wang, 2014; Tutz et al., 2016).

This differentiation is partly dependent on the definition of extreme responding described in the last paragraph. If ERS is defined as the general extremeness of a person's responses, MRS is also interpreted as "mild response style", and thought to be the exact opposite of ERS (Jin & Wang, 2014). In this line of thought, MRS does not need to be treated as a separate concept. However, if ERS is defined as a person's tendency to choose the highest or lowest response category, MRS should be considered as a unique response style, which might be driven by different factors in the response process. For example, Weijters et al. (2010b) argue for a separation of ERS and MRS, given evidence that a subsample of old, less educated subjects shows high proportions of mid and extreme responses at the same time. If such a pattern were replicated in larger samples, this would suggest that modeling ERS as a unidimensional construct is inappropriate.

Some authors even go one step further and consider responses to a Likert scale item as series of cognitive decisions that correspond to different aspects of extreme responding (Böckenholt, 2012; Plieninger & Meiser, 2014; Zettler et al., 2015; Böckenholt & Meiser, 2017; De Boeck & Partchev, 2012). For example, when dealing with an item with seven response categories, the first decision process controls if the mid category is chosen or not. This is comparable to the concept of MRS. If the mid category is not selected, a second decision process is responsible for the direction of the item response. This should be the content process, only determined by the primary trait measured. The result of a third decision process is whether the most extreme category is chosen or not. This is comparable to the definition of ERS as the tendency to choose the highest or lowest response category. If the extreme category is not chosen, a fourth process controls which of the remaining two categories is chosen, which also reflects a certain extremeness of the response. Notably, the last process would not exist for an item with five or less response categories and the first process would not exist if there is no mid category. Thus, in this conceptualization of ERS, the number of dimensions depends on the response scale of the respective items.

Continuity of Extreme Responding In one line or research, ERS is viewed as discrete differences in item responding, best described by only two distinct groups of people (Rost, 1990; Wetzel, Böhnke, Carstensen, et al., 2013; Ziegler & Kemper, 2013; Meiser & Machunsky, 2008; Eid & Rauber, 2000). The groups are described slightly differently, depending on the two attributes introduced above. Some define "extreme responders" with high preference for the most extreme categories, and "midpoint responders" with high preference for the mid category. This constitutes a bi-dimensional concept of ERS, defined by special preferences for the middle and most extreme categories (Ziegler & Kemper, 2013). Others use a unidimensional definition with "extreme responders" who prefer the most extreme categories, and "non-extreme responders" who avoid the most extreme categories but do not show a special preference for the mid category (Wetzel, Carstensen, & Böhnke, 2013). Another conceptualization assumes one group with balanced response behavior while the other is either characterized by a special preference for the most extreme categories (Eid & Rauber, 2000), or by an avoidance of those (Meiser & Machunsky, 2008). In contrast to these discrete conceptualizations, other researchers propose that ERS is a continuous trait

(Greenleaf, 1992; Weijters et al., 2008; Jin & Wang, 2014; Tutz et al., 2016; Baumgartner & Steenkamp, 2001). Also here, definitions of ERS differ with respect to the other two ERS attributes. Some authors use continuous unidimensional definitions of ERS, which may represent the general extremeness of responses (Jin & Wang, 2014; Tutz et al., 2016) or the tendency to select the most extreme categories (Greenleaf, 1992). Others consider separate continuous versions of ERS and MRS, representing special preferences for the most extreme and mid categories (Baumgartner & Steenkamp, 2001; Weijters et al., 2008)

As already mentioned, the three attributes of ERS are rarely discussed explicitly in the literature. In most cases, they have to be inferred from stated definitions. This is because the highlighted taxonomy is closely related to the methodological approach by which ERS is investigated in a specific line of research: So far, all researchers defining ERS as a discrete trait used mixed Rasch modeling (Rost, 1990; Wetzel, Böhnke, Carstensen, et al., 2013; Ziegler & Kemper, 2013; Meiser & Machunsky, 2008; Eid & Rauber, 2000). On the other hand, continuous concepts of ERS based on the special preference for the most extreme response categories have traditionally been studied by computing response style indices from heterogeneous items (Greenleaf, 1992; Baumgartner & Steenkamp, 2001; Weijters et al., 2008). For a more detailed description of these methodological approaches see chapter 1.3.

In previous studies, it is mostly unclear whether the choice of a methodological approach was guided by the researchers' theoretical conceptualization of ERS, or whether they adapted their definition of ERS to their preferred methodological approach. This ambiguity is important when thinking about the empirical basis of different concepts of ERS. In particular, the issue whether ERS is a continuous construct seems to be a principal question about the true nature of ERS, that is accessible to empirical testing. This issue has been mentioned by Austin, Deary, and Egan (2006), but directed to future research as the methodology to study ERS as a continuous trait was considered to little developed at the time. Meanwhile, ERS has been detected repeatedly with sophisticated techniques that treat it as a continuous trait (Weijters et al., 2008; Jin & Wang, 2014; Tutz et al., 2016). However, to test empirically whether ERS is better described as a discrete or a continuous trait, a technique is needed that will deliver different results depending on the nature of ERS. In study 1, we introduce an approach to detect ERS, which for the first time offers the necessary characteristics to shed light on this important question.

Moreover, the new method allows us to investigate whether extreme responding is limited to the highest and lowest response category of a rating scale, and possibly the mid category. If this is the case, subjects with high ERS scores should be expected to show higher response probabilities on these most extreme response categories, only. If ERS is related to the general extremeness of responses, we should observe higher response probabilities for categories next to the most extreme ones, too, as long as the measure of ERS takes this definition of general extremeness into account. Study 1 also investigates this issue.

Considering the dimensionality of extreme responding, it is less clear how to compare the different conceptualizations based on data, or whether they are open for empirical test-

ing at all. The predominant question is whether the most extreme and the mid categories of a rating scale are qualitatively different from all other response options. Multiprocess models (Böckenholt, 2012; Plieninger & Meiser, 2014; Böckenholt & Meiser, 2017; De Boeck & Partchev, 2012) may be a good way to approach this issues among others. Although we did not apply multiprocess models here, they will be mentioned several times as they provide an interesting alternative approach to investigate some of our research questions.

In the original work presented in this dissertation, we conceptualize ERS as a stable, continuous, bipolar, unidimensional trait, that can be interpreted as the unconditional general extremeness of a person's responses to Likert scale items, and ranges from high preference for the mid category to high preferences of the most extreme categories. In the introduction, the term ERS is used for various definitions of extreme response style, to simplify the discussion of previous research. However, contrasting definitions of ERS will be clarified, when necessary.

1.2 Challenges When Studying Extreme Response Style

When studying the impact of ERS on a psychological scale, it is difficult to separate a high value of ERS from a high value on the primary trait of interest. Baumgartner and Steenkamp (2001) call this "not to confound stylistic variance with substantive variance". The most naive way to measure ERS is to calculate the number of responses on the most extreme categories in a psychological scale of interest. However, this measure of ERS is highly confounded with the primary content of the scale. Subjects with high values on the primary trait will frequently choose the most extreme categories. Therefore, we would expect some relationship between the ERS score and the trait score derived from the questionnaire, that results from correlating the primary trait with itself. This fallacy can be avoided by either measuring ERS based on a separate set of items that were not used to measure the primary content of interest, or by using psychometric models that incorporate more sophisticated ways to assess response style and content from the same set of items. In the next chapter, we describe the most commonly used methodological approach for each strategy.

The difficulty of separating response style and content is most important when studying how ERS is associated with other variables. As ERS appears to be stable across time and situations (Weijters et al., 2010b; Wetzel, Lüdtke, et al., 2016; Wetzel, Carstensen, & Böhnke, 2013), it seems likely that ERS is a behavioral manifestation of a combination of certain person characteristics (Naemi et al., 2009). Linking ERS to demographic variables has yielded mixed results (for a review see Vaerenbergh & Thomas, 2013). While high levels of ERS have been found to be negatively correlated with education, findings on sex differences remain inconclusive. Reported results range from no effect, to higher levels of ERS for both women and men. For age, suggestions range from no effect, over positive or negative linear effects, to quadratic effects, with both young and old persons showing high levels of ERS. A possible reason for these inconsistencies is that measures of ERS might be confounded with primary content in a large number of studies (Beuckelaer, Weijters, &

Rutten, 2010). If so, the reported associations may vary depending on how demographic variables are related to the confounded content.

Studying the relations between ERS and demographic variables is straightforward from a methodological standpoint, as long as ERS measures are not confounded with content variance. The situation gets more complicated when investigating the association between ERS and personality traits. To this day, the established approach to measure personality traits are psychological questionnaires. Correlating self-report measures of personality with ERS is highly problematic, as any personality score is likely contaminated with response style variance. The study by Plieger, Montag, Felten, and Reuter (2014) represents an example in which claimed associations between ERS and personality traits are hard to interpret. The authors report a positive association between ERS and neuroticism. However, they used the number of extreme responses on a Big Five inventory as their ERS measure, with one fifth of these items also being used to compute the neuroticism score.

Even when efforts are undertaken to obtain an independent ERS variable that is largely free of confounding primary content, this does not solve the problem that ERS is somewhat correlated with itself due to contaminated personality scores. This was demonstrated by the study by Austin et al. (2006), who report higher extraversion and conscientiousness scores for subjects in the high ERS class of mixed Rasch models. However, raw personality scores were used to estimate this relation, instead of person parameters from the two class solution, which would have been somewhat corrected for ERS. These results should therefore be interpreted with caution. The literature documents only two serious attempts to circumvent these limitations. Naemi et al. (2009) determined personality scores using peer ratings instead of self reports. They have found that subjects' ERS was positively related to intolerance of ambiguity, preference for simplistic thinking, and decisiveness rated by their peers. Lately, another study also used other ratings to show that ERS, conceptualized as the intensity process in a multiprocess model, is negatively related to humility (Zettler et al., 2015).[2] As other ratings suffer from possible selection bias of the primary and secondary raters and apparent flaws of using human judgments, Cabooter (2010) decided to avoid direct measurement scales completely. She constructed a set of scale-free personality measures, and found ERS to be positively associated to a concept termed promotion focus.[3] Unfortunately, this work which was part of her doctoral thesis has not been taken up on since.

In light of this discussion, we think that investigating the relationship between ERS and personal traits is not a very fruitful endeavor at this point. Instead, all three studies presented here seek to foster our understanding of the true nature of ERS. Hopefully, this will contribute to the development of better psychological measures which are less affected by ERS (see chapter 5.2), or alternatively enable the community to model ERS directly sometime in the future (see chapter 2.4.4).

Until this is achieved, we agree with Cabooter (2010) that ERS should better be related to personality without using self-report scales. A promising way would be to substitute

[2] Humility constitutes the sixth factor in the HEXACO model of personality (Ashton & Lee, 2007).

[3] Promotion focused people are likely to take risks and make active choices.

personality questionnaires by predictions of personality traits based on digital footprints of personality and behavior. It has been shown that self ratings of personality can be predicted by Facebook likes with comparable accuracy than by other ratings of spouses (Youyou, Kosinski, & Stillwell, 2015). They used techniques derived from predictive modeling, demonstrating the great potential of these methods for psychological science. Remarkable performance was achieved by a simple regularized linear model. Even better predictions may be possible when using non-linear methods like we did in study 3. Combining direct modeling of ERS with predictive models of personality, might yield the first reliable insights into how ERS is associated with personality traits.

1.3 Methods of Detecting Extreme Response Style

1.3.1 Mixed Rasch Modeling

One of the two most popular approaches to detect ERS are mixed Rasch models (Rost, 1990), a combination of Rasch models and latent class analysis. Generally, the polytomous mixed Rasch model is used in this context (Rost, 1991), because rating scale items are analyzed to detect ERS. Polytomous mixed Rasch models assume that the target population consists of several latent classes. In each class, item responses follow a partial credit model (PCM; Masters, 1982). When fitting a mixed Rasch model, the number of latent classes must be specified by the researcher. The estimation algorithm then yields latent classes which maximally differ with respect to their respective model parameters. As with any discrete mixture model, threshold parameters for the PCMs in all latent classes are estimated jointly, and a probability that a subject belongs to a class is estimated for each of the latent classes. Mixed Rasch models can be thought of as a way to detect differences in the item response patterns, resulting from a combination of unobserved covariates of the subjects included in the analyzed sample. Therefore, mixed Rasch models can be used as a rigorous test of measurement invariance in the PCM. Mixed Rasch model solutions with different numbers of classes can be compared against each other, based on information criteria. For the special case of one latent class, the polytomous mixed Rasch model is identical to the PCM. If a mixed Rasch model with two or more classes shows better fit to a set of item response data than the PCM, the unidimensionality assumption of the primary trait measured by the scale must be rejected. However, the threshold pattern in the latent classes can be used to gain useful insights into the source of multidimensionality. While different primary traits might be measured within each class, another possibility for the emergence of multiple classes is that subjects differ in their usage of the response scale. This feature makes the polytomous mixed Rasch model a useful tool to investigate ERS.

Mixed Rasch models have been originally used to investigate measurement invariance by Rost, Carstensen, and von Davier (1997). Thereby, they discovered signs of ERS in a personality inventory, the German version of the NEO Five-Factor Inventory (NEO-FFI; Borkenau & Ostendorf, 2008). They found two latent classes in a mixed Rasch analysis of the NEO-FFI. In one class, item thresholds where close together and the probability to

respond on the highest or lowest scale category was high even for subjects with moderately high or low values on the latent trait. In contrast, threshold parameters were widely spaced in the second class, with higher probability to use less extreme categories for moderate values on the latent trait. This pattern is highly intuitive when considering how ERS should manifest in item responses of psychological questionnaires. The introduction of mixed Rasch models made it possible to investigate ERS within the sophisticated modeling framework of item response theory. Similar patterns of ERS have meanwhile been detected in the English version of the NEO-FFI (Borkenau & Ostendorf, 2008) by Austin et al. (2006) and in its German long version, the Revised NEO Personality Inventory (NEO-PI-R; Ostendorf & Angleitner, 2004) by Wetzel, Böhnke, Carstensen, et al. (2013). Two class patterns of ERS are prevalent not only in the NEO family of personality inventories, but have also been detected in inventories on anger expression (Gollwitzer, Eid, & Jürgensen, 2005) or leadership performance (Eid & Rauber, 2000).

These findings have had a big impact on describing ERS as a trait that differentiates two distinct classes of people, one with a high tendency to chose extreme response categories and the other with a high tendency to chose the mid category. Although this might be a reasonable conceptualization, it does not mean that these two distinct groups with qualitatively different response tendencies actually exist. In contrast, they could likewise be an artifact of the mixed Rasch methodology. Because of its discrete structure, a mixed Rasch model can only differentiate between a small number of latent classes. For higher numbers of classes, differentiation gets harder and convergence problems are common (Böckenholt & Meiser, 2017). If ERS is better described as a continuum ranging from high to low tendencies to give extreme responses, that structure is insufficiently depicted by a discrete model like the mixed Rasch model (Bolt & Johnson, 2009). Indeed, mixed Rasch models yield inappropriate trait estimates when item responses are modeled by a multidimensional item response theory (IRT) model with a continuous ERS variable (Wetzel, Böhnke, & Rose, 2016). In practice, more than three latent classes in mixed Rasch analyses are rare. Even when three classes are found, authors usually report the threshold pattern of the third class to be closely comparable to either the class of extreme or mid responders (for an example, see Wetzel, Carstensen, & Böhnke, 2013).

The ability of mixed Rasch models to find differences in response patterns resulting from an unobserved subpopulation, is a great feature when testing for measurement invariance in the PCM. However, when studying response styles, interpretability of the resulting multi-class solutions is limited, as response styles are only one possible factor inducing differences between classes. Even if a pronounced pattern of wide compared to narrow thresholds emerges, response styles can be confounded with differences in the level of the primary trait, or by other personality characteristics that were not explicitly observed in the dataset. Moreover, mixed Rasch models have been found to produce spurious latent classes under certain conditions (Alexeev, Templin, & Cohen, 2011).

To address these problems, Wetzel and colleagues (Wetzel, Carstensen, & Böhnke, 2013; Wetzel, Böhnke, Carstensen, et al., 2013) use a constrained version of the mixed

Rasch model, in which item locations are restricted to be equal between latent classes.[4] These models provide higher confidence that threshold patterns between latent classes can be interpreted as differences in response scale usage. Model fit of the constrained mixed Rasch model can be compared to its unconstrained version based on information criteria. If the model fit of the constrained version is better, one can be more confident that the same primary construct is measured in all latent classes. Different threshold patterns between classes can then be attributed more clearly to differences in response styles.

A big advantage of mixed Rasch models is that ERS can be detected based on the same items which are used for measuring the construct of interest. In contrast to the approach described in the next section, no additional items are needed to investigate ERS. As a consequence, mixed Rasch models can be applied to assess effects of ERS in any study using psychological scales based on a homogeneous set of rating scale items. More generally, they can be considered as a flexible, exploratory, all-purpose tool to deal with a variety of response styles (Böckenholt & Meiser, 2017).

1.3.2 Response Style Indices

The other widely used approach to study response styles focuses on an optimal separation of response style and content of interest. This is achieved by computing a response style index from a set of heterogeneous items. The method was originally introduced by Greenleaf (1992) and extensively developed by Weijters and colleagues (Weijters et al., 2008; Weijters et al., 2010a; Weijters, Cabooter, & Schillewaert, 2010; Weijters et al., 2010b) under the label Representative Indicators of Response Styles (RIRS). The basic idea is that an index representing extreme responding in a set of items is minimally confounded with trait variance when the items measure widely different psychological constructs.

Such indices for ERS typically estimate a subject's probability to give an extreme response by computing how often the highest or lowest response category is chosen on average for some heterogeneous item set (Greenleaf, 1992). However, different coding schemes are possible, that also give lower weights to less extreme response categories. Such a weighting scheme was applied in all three studies presented in this thesis.

Previous studies used different numbers of items to compute response style indices, ranging from 16 (Greenleaf, 1992) to 112 (Weijters et al., 2010a). As a rough measure, most studies report the mean absolute inter-item correlation for the items contained by the index, ranging from 0.071 (Greenleaf, 1992) to 0.12 (Baumgartner & Steenkamp, 2001).

While Weijters et al. (2008) urge researchers to use a minimum of 30 items when response styles are of major interest, Greenleaf (1992) argues, based on analytical reasoning, that reducing the inter-item correlations is more effective to maximize the accuracy of a response style index than increasing the number of items. As items are never completely uncorrelated in practice, low inter-item correlations are easier to achieve with a smaller number of items.

An obvious disadvantage of response style indices is that additional items are needed.

[4] The location of an item is defined as the sum of its threshold parameters.

Moreover, as these items have to be highly uncorrelated, they should not be taken from the small number of scales included in a typical psychological study. In most situations, a set of heterogeneous items has to be deliberately included in the study design in order to detect response styles with the index approach. This has been done repeatedly by Weijters and colleagues (Weijters et al., 2008; Weijters et al., 2010a; Weijters, Cabooter, & Schillewaert, 2010; Weijters et al., 2010b).

Weijters et al. (2008) highly recommend that additional items are randomly sampled from a relevant item population. This ensures that findings on response style are generalizable to other items, and that items are heterogeneous in content, thus showing only small inter-item correlations. In their studies, they sample from a large set of marketing scales, and then draw one item from each selected scale. We use a similar approach in study 2 and 3. There, we collected our own data, including a set of heterogeneous items that we sampled from publicly available item banks.

Another possibility to investigate response styles are large scale comprehensive surveys, that often contain a variety of scales used to study a number of different research questions. In these datasets, a set of uncorrelated items can be found from scales that were primarily included for content reasons, as in Baumgartner and Steenkamp (2001), Greenleaf (1992) or Plieninger and Meiser (2014). However, when working with secondary data, ensuring a high level of heterogeneity can be challenging (Beuckelaer et al., 2010). By using an algorithm that automatically selects an uncorrelated set of items from all available items in a dataset, heterogeneity can be further improved. We used such a procedure in both analyses of study 1. In the first analysis, we selected items to compute an ERS index from a large personality inventory that contains several hundred items, measuring constructs which are theoretically independent. In the second analysis, we analyzed longitudinal panel data, where we selected a set of uncorrelated items from a large number of questionnaires that investigate a wide variety of research questions.

Response style indices can be used in different analysis frameworks, to detect response styles or to study the relationship between response styles, demographic person characteristics, and psychological constructs of interest. The most common approach is to estimate a structural equation model, including latent response style variables composed of multiple parcels of response style indicators (Weijters et al., 2008). In study 1 we present a new procedure, using an ERS index as a covariate in partial credit trees (PC trees; Komboz et al., 2016). This combines the rigorous separation of response style and psychological content from the response style indices framework with the intuitive threshold interpretation of ERS known from mixed Rasch modeling.

1.4 Mitigating the Impact of Extreme Response Style

Response styles have been declared "an enemy to validity" by Cronbach (1950). Although some researchers remained skeptical (Rorer, 1965), it is widely assumed today that response styles have detrimental effects on psychological measurement:

ERS has been shown to affect means and variances derived from Likert scale mea-

sures. Correcting for ERS in cross-cultural research changes the ranking of countries in mean conscientiousness (Mõttus et al., 2012). Sex differences in leadership types disappear when considering different levels of ERS between men and women (Moors, 2012). In facets of the NEO-PI-R, ERS explains 25% of the variance of item responses (Wetzel & Carstensen, 2015). Moreover, ERS influences the estimated relationship between variables. Controlling for response styles often changes the correlations between scales (Baumgartner & Steenkamp, 2001). For example, estimated latent correlations between two dimensions of Personal Need for Structure decrease when accounting for ERS (Böckenholt & Meiser, 2017).

Although a large body of empirical research supports that ERS poses a threat to psychological measurement, two recent simulation studies present different results. The authors claim that as long as the response style is not strongly correlated to the primary construct of interest, response styles might have a minor impact in practice (Plieninger, 2016; Wetzel, Böhnke, & Rose, 2016).

Wetzel, Böhnke, and Rose (2016) simulated item response data contaminated with ERS. They used an ordinal mixed-Rasch model with two response style groups and a bi-dimensional IRT model with a continuous ERS variable. Trait and ERS were either uncorrelated or weakly correlated ($r = 0.2$) in the bi-dimensional model. Simulated data was analyzed with PCMs, mixed-Rasch models, bi-dimensional IRT models, and a simple approach in which sumscores were regressed on an ERS index computed from the same items. Trait recovery was evaluated with the correlation and the average absolute difference between true and estimated trait parameters. In all conditions, trait recovery with the PCM which ignored ERS was comparable to trait estimates from the true data generating model that took ERS into account. The sample size and the number of items in the scale did not affect these results. When the PCM was used to simulate data without ERS, trait recovery with the same model was highly biased. This raises some concerns about the interpretability of the simulation results in general. One possibility for this unexpected finding might be that implausible priors were used in the expected a posteriori estimates of the person parameters.

Plieninger (2016) simulated data from multidimensional IRT models extended for ERS and ARS. This is the same model class which was also used in the simulation study by Wetzel, Böhnke, and Rose (2016). The correlation between the trait and the response style variable in the model was varied systematically. To evaluate the influence of ignoring response styles, several measures were considered: Cronbach's Alpha estimates (Cronbach, 1951) were compared to a version of Alpha in which true response styles are partialled out (Bentler, 2016). Correlations between sumscores from two scales were compared to partial correlations which control for the true response style values. Correlations of true trait parameters with sumscores were compared between conditions with and without response style. Finally, dichotomous trait classifications based on different percentiles of the sumscores were compared to the corresponding classifications based on the true trait values. The author concluded that ERS induced little bias, as long as the correlation between trait and response style is small. The simulation study suffers from several limitations: First, the effect of ERS on Cronbach's Alpha is irrelevant. Reliability always depends on

a certain measurement model. As soon as ERS is present in the data, the essentially tau-equivalent model of classical test theory (CTT) is violated and Cronbach's Alpha cannot be interpreted as reliability. It is not clear whether the corrected version of Cronbach's Alpha is based on an appropriate definition of reliability in the ERS model which was used for the simulation. Only then would it make sense to interpret the observed differences between the standard Cronbach's Alpha estimates and the corrected versions. Moreover, the comparison between the correlation of sumscores and the corresponding partial correlation cannot be meaningfully interpreted. The partial correlation is not an unbiased estimator of the true correlation between the latent variables in the data generating IRT model, as ERS does not have a linear effect on the sumscore of the scale.

A common caveat of both simulation studies is that only the average effect of ERS was investigated. When averaging across the distribution of both the latent trait and ERS, the impact of ERS might be small as reported in both simulations. The classification accuracies reported by Plieninger (2016) suggest that the impact of ERS is stronger for subjects with a high or low value on the content trait. Trait and response style are supposed to follow a multivariate normal distribution with small correlation, which seems like a reasonable assumption. However, this implies that even for extreme trait values, the majority of subjects have moderate values on ERS, explaining why bias is still small on average. Ignoring response styles should have the largest impact for people with extreme values on both the content trait and on ERS. As this applies only to a small number of people, the impact of ERS appears negligible when averaged across all subjects.

While ERS might be less relevant in studies focusing on correlations and mean differences between groups, its impact on individual diagnostics could be devastating. In many applications like clinical psychology and personnel selection, one is genuinely interested in subjects with extremely high or low values on some latent trait. Under these circumstances, not controlling for ERS might yield strongly biased trait estimates for people with high or low response style values. This can lead to judgment errors that might have serious consequences for the examined individual (e.g. a patient is denied insurance) or the client of the psychological measurement (e.g. a company hires an incompetent employee).

If applied researchers want to mitigate possible effect of ERS in their psychological measurements, two kinds of strategies can be applied (Podsakoff, MacKenzie, & Podsakoff, 2011). Statistical remedies try to correct trait scores for ERS within statistical models, while procedural remedies try to avoid the confounding of item responses with ERS in the first place.

1.4.1 Statistical Remedies

In theory, any statistical technique which is capable of detecting ERS can also be used to correct for the impact of the response style. How the statistical control works in detail, depends on the applied model class:

In mixed Rasch models, person parameters are regarded as a measure of the latent trait, which is corrected for differences between the latent classes (Rost et al., 1997). If visual inspection suggests that classes capture different levels of extreme responding, estimated

person parameters are supposed to be free of ERS. For this interpretation, it is essential that the same latent trait is measured within each class, which can be assumed if item difficulties are the same for all classes (Wetzel, Carstensen, & Böhnke, 2013).

When a response style index from heterogeneous items is used to measure ERS, corrected scores can be obtained by regressing trait scores on the ERS index. In the simplest approach, sumscores of the scale of interest are regressed on the ERS index in ordinary linear regression, and the resulting residuals are used as corrected trait scores (Baumgartner & Steenkamp, 2001). In the representative indicators response style means and covariance structure (RIRSMAC) approach (Weijters et al., 2008), a linear structural equation model is estimated. Item responses from the scale of interest load on a common content factor, as well as a latent response style factor. The response style factor additionally loads on several identical ERS indices, which are computed based on separate item parcels. Estimated factor scores for the content factor are then considered trait scores which are corrected for ERS.

How well the correction works depends on the fit between the true structure of the response style, and the structure which is assumed by the applied ERS model. Both the regression residuals and the RIRSMAC approach assume a linear relationship between ERS and individual item responses, which makes no theoretical sense. Thus, both methods should perform poorly when used to correct for ERS in practice. If ERS is best described by a continuous trait (Greenleaf, 1992; Weijters et al., 2008; Jin & Wang, 2014; Tutz et al., 2016; Baumgartner & Steenkamp, 2001), mixed Rasch models which implicitly assume a categorical structure can be expected to perform poorly as correction method (Wetzel, Böhnke, & Rose, 2016). In contrast, more simple models might yield more robust corrections for ERS, in particular if the sample size or the number of items in the analyzed scale are small. Wetzel, Böhnke, and Rose (2016) report good performance of using residuals from linear regressions with an ERS index as predictor, even when the assumed linear relationship between ERS and item responses was violated in the simulation process. This leads them to the conclusion that model-based approaches can also do more harm than good under some circumstances. In addition to model complexity, the degree of structure in the model might be important as well. Possibly due to their high flexibility, mixed Rasch models performed worse in the correction of ERS than multidimensional IRT models with a clear structural representation of ERS (Wetzel, Böhnke, & Rose, 2016). This structural aspect might explain, why the mixed Rasch model even performed worse than the other models, when the mixed Rasch model was used to simulate the data.

These findings lead to the conclusion that ERS should be modeled directly with IRT models that assume an appropriate structure for the response style. When a good model is found, derived trait estimates should substantially reduce the impact of ERS in psychological applications. Exploratory studies on the real structure of ERS are important to choose between the models which have been proposed in the literature and to stimulate further refinements. In study 1, we introduce a new approach to explore ERS, which is well suited to detect and depict ERS in an objective way but is not a good candidate to correct for ERS due to its high flexibility and lack of structure. However, the proposed method can yield important insights into the structure of ERS. Based on our empirical

results, we discuss which models might be most promising for direct modeling of ERS in chapter 2.4.4.

1.4.2 Procedural Remedies

Instead of removing ERS from item response data with statistical techniques, procedural remedies try to avoid the confounding of item responses with ERS, by adjusting the design of the measurement process.

One aspect of psychological measurements which is supposed to be related to the impact of ERS, is the response format of items in psychological questionnaires. The literature on this topic is mixed. It has been suggested that participants tend to choose more intermediate item responses when using fully labeled response scales (Weijters, Cabooter, & Schillewaert, 2010). If only endpoints are labeled, Weijters, Cabooter, and Schillewaert (2010) find that increasing the number of response categories leads to a smaller rate of extreme responses, and a higher rate of midpoint responses. In contrast, other researchers report that the responses rates for the midpoint and the most extreme categories are unrelated to the number of response categories (Kieruj & Moors, 2010, 2013).

Instead of investigating how ordinal item formats should be designed to minimize the impact of ERS, the most natural strategy to avoid ERS is to use dichotomous items. Although this advice has been voiced in the psychometric literature multiple times (Bühner, 2011; Cabooter, Millet, Weijters, & Pandelaere, 2016; Wetzel, Carstensen, & Böhnke, 2013), empirical evidence is not available. In study 2, we argue that contrary to the standard definition of ERS which relies on items with ordinal response formats, extreme responding might also affect dichotomous item responses. Recently introduced IRT models are used in combination with ERS indices from heterogeneous items with ordinal response formats, to investigate whether dichotomous item responses are unaffected by ERS.

Psychological questionnaires do not only consist of the items and their response format. Another important aspect is how respondents are instructed to answer the questions. In study 3, we will investigate specific instructions, which urge participants to give more or less extreme item responses. The appropriate direction of the instruction is selected for each participants, based on some measure of ERS. As a consequence, individual response tendencies should be compensated during the response process, and statistical control for ERS might be unnecessary. Cronbach (1950) also advocated researchers to reduce response bias by increasing the testwiseness of participants through specific training sessions. This should be most successful if subjects have some conscious knowledge about their own response tendencies. In studies 2 and 3, we also explore if the self-reported tendency to give extreme responses in questionnaires carries valid information about the impact of ERS. If this is the case, simple training procedures could be developed which focus on increasing participants' awareness of their individual response styles, and help them in adjusting their item responses accordingly.

Chapter 2

Study 1: Detecting Extreme Response Style with Partial Credit Trees

2.1 Introduction

2.1.1 An Introduction to Partial Credit Trees

An item in a psychological test is said to display differential item functioning (DIF) if the probability of item responses under a particular item response model varies between people from different subgroups (e.g. based on age or sex). Rasch trees (Strobl et al., 2013) and PC trees (Komboz et al., 2016) were recently introduced as procedures to automatically detect DIF in item response models on the basis of flexible combinations of person covariates. Both methods rely on the model-based recursive partitioning framework by Zeileis, Hothorn, and Hornik (2008), and are conceptually related to the decision trees introduced in chapter 4.1.3.

If no DIF is present in a dataset, fitting a joined item response model for the whole sample is appropriate. Consequently, when the sample is split based on a covariate and an item response model is computed separately for each subsample (e.g. men and women), the estimated model parameters should be approximately the same. However, if DIF is induced by available covariates, there is a partition of the sample based on these covariates, in which model parameters differ. Rasch and PC trees do not only provide a global test for DIF (Strobl et al., 2013; Komboz et al., 2016), they also automatically detect subgroups for which DIF is present based on available covariates.

In Rasch trees, the underlying item response model is the dichotomous Rasch model, whereas the PCM is used in PC trees. An excellent resource that includes all commonly used models from IRT is Van der Linden (2016). A Rasch or PC tree takes the responses on a set of dichotomous or polytomous items that belong to a psychological scale, as well as a set of person covariates as input. Then the algorithm repeatedly performs the following three steps (Strobl et al., 2013; Komboz et al., 2016):

1. The item response model with all its item parameters is estimated for the current

sample. In the first round, the whole sample is provided.

2. Statistical tests are performed to examine the stability of model parameters with respect to each covariate.[1]

3. If significant parameter instability is found, the covariate responsible for the strongest instability is chosen to split the current samples into two subsamples. Then, the split is performed at the value of the covariate that results in the highest improvement of model fit in the two subsamples compared to the combined sample.[2]

These steps are recursively repeated for each subsample until either no significant parameter instability is detected, or the sample size of the subsamples falls below a specified threshold. The statistical tests for detecting parameter instability are Bonferroni adjusted for the number of covariates, to avoid excessive false-positive rates when a large number of covariates are available. Furthermore, the selection of a splitting variable is made independently of choosing the optimal cut point within the splitting variable. This prevents a phenomenon called variable selection bias, in which covariates with a higher number of possible cut points have an advantage in being selected as splitting variables due to multiple comparisons (Strobl et al., 2013). As a result, Rasch and PC trees provide a global test for DIF, that preserves a prespecified level of significance. Owing to their recursive nature, Rasch and PC trees are powerful tools to detect DIF caused by complex interactions of covariates, as long as a sufficiently large sample is provided. Naturally, the power of the statistical tests in step 2 depend on sample size. Thus, fewer splits can be expected if the total sample size is small, considering that subsamples get approximately halved by each split. When the recursive splitting algorithm has converged, the same type of item response model (dichotomous Rasch model or PCM) is estimated independently for each of the final subsamples.

A Rasch or PC tree analysis can be visualized as a tree-like structure, giving the procedure its name. The tree plot provides a compact summary containing all necessary information for interpreting the results. Starting with the full sample on the top of the tree (this is also called the root), all splits are displayed together with the splitting variable, the respective cut point, and the corrected p-value of the corresponding test of parameter instability. At the bottom of the tree when no more splits are justified, a graphical display of the estimated item response model is presented for each final subsample. These are called the leaves of the tree. In the leaves of Rasch trees, item difficulties are shown for each item of the scale. In PC trees, more informative region plots are used in PC trees, which are based on effect displays by Fox and Hong (2009). In a region plot, each item is represented by a vertical column, consecutively numbered on the x-axis. The y-axis represents the common latent trait, measured by all items. Within each item column, the latent variable

[1] To assess parameter stability, advanced statistical tests from the class of generalized M-fluctuation tests are used, which form the basis of the model-based recursive partitioning framework. Further information for the interested reader can be found in Strobl et al. (2013) and Zeileis et al. (2008).

[2] The optimal cut point is chosen by maximizing the partitioned log-likelihood over all possible cut points of the selected covariate (Strobl et al., 2013).

on the y-axis is partitioned into regions where different response categories have the highest response probability. These regions are depicted in greyscale. If the threshold parameters of an item are in correct order, the borders of the regions of highest category probability coincide with the threshold parameters of the PCM. However, if thresholds are unordered, there is no region of highest probability for one or more response categories. Then, those thresholds not corresponding to borders of highest category probability are indicated by red lines (Komboz et al., 2016). Above each region plot, the number of subjects contained in the leaf is reported.

An important caveat of Rasch and PC trees is that DIF can only be detected if the variables containing DIF were actually observed in the dataset. This is in sharp contrast to mixed Rasch models, which also detect DIF with respect to unobserved subgroups. However, their high level of interpretability is a very appealing aspect of Rasch and PC trees. When a covariate that induces DIF is found, the nature of DIF can be directly observed in the thresholds patterns of the item response models in the leaves of the tree. This feature is especially meaningful when using PC trees to detect ERS.

2.1.2 Aim of Study and Research Questions

In the following analyses, we show that PC trees can be used to detect ERS in a psychological scale of interest, by applying an ERS index from heterogeneous items as described in chapter 1.3 as a covariate in the PC tree. This approach combines advantages of both mixed Rasch model and ERS index approaches. A minor disadvantage is that additional items and a large sample are needed to detect ERS. If the analyzed sample is split based on the ERS index, measurement invariance is rejected, as a single PCM is not appropriate for the whole sample (Komboz et al., 2016). In this case, an effect of ERS on the analyzed scale is detected as model misfit and can be directly linked to the objective ERS index. More importantly, region plots in the leaves of the PC tree immediately reveal how subjects differ in their response patterns, contingent on their value of the ERS index. Thus, it is possible to interpret ERS in terms of threshold patterns as known from mixed Rasch modeling.

The main aim of this study is to present the advantages of our PC tree approach to detect ERS, by analyzing two different datasets. First we use a well known dataset from a large Big Five personality inventory, the German NEO-PI-R (Ostendorf & Angleitner, 2004), which has been found to be influenced by ERS using mixed Rasch model analysis. In a second step, we analyze new data from the longitudinal GESIS panel (GESIS, 2015). In the NEO-PI-R dataset, the ERS index contains items from the same questionnaire, but from different factors than the analyzed scales. In the GESIS dataset, the ERS index contains heterogeneous items from different waves of the longitudinal panel, which were completed several months apart.

Specific research questions concerning our approach to use PC trees in combination with an ERS index from heterogeneous items for detecting ERS were:

Ability of PC trees to detect ERS As previous studies found strong evidence of ERS in the NEO-PI-R dataset, we expect PC trees to split the sample based on an ERS index computed from heterogeneous items. It is unclear, if PC trees should split the sample based on an ERS index in the GESIS panel dataset. On the one hand there is no theoretical reason why ERS should be restricted to certain personality dimensions or psychological scales. On the other hand, the ERS index in the GESIS analysis was computed in a more conservative way, that should make it more challenging to detect extreme responding. However, if an effect of ERS is found in this setting, stronger conclusions about the nature of ERS can be drawn.

Validation of the Threshold Interpretation of ERS Our approach to combine PC trees with an ERS index allows us to validate the interpretation of ERS based on the threshold pattern within the leaves. As stated in chapter 1.3, this understanding of ERS stems from analyses using mixed Rasch models. Although theoretically persuasive, the mixed Rasch methodology can not guarantee that the characteristic threshold pattern stems from another phenomenon than ERS. If it is appropriate to interpret the threshold pattern in a mixed Rasch model as an indication of ERS, we expect a similar pattern in the region plots of a PC tree when an ERS index was selected as splitting variable. In leaves containing subjects with high ERS, the distances between thresholds in the PCM should be smaller than in leaves with low ERS. Moreover, in leaves with high ERS, extreme response categories should have the highest probability for larger regions of the latent variable, compared to leaves with low ERS.

Validation of the ERS index If the PC tree analyses confirm the ERS interpretation of threshold patterns in mixed Rasch analyses, this validates the use of ERS indices based on heterogeneous items to study ERS. The rationale that an ERS index signals differences in extreme responding relies on the heterogeneity of the items used for its computation. Although using totally uncorrelated items should eliminate other possible systematic influences, items are never completely uncorrelated in practice. Furthermore, zero correlation does not rule out nonlinear dependencies between items that do not represent response tendencies. If the threshold patterns resulting from splits based on the ERS index can be convincingly interpreted as different degrees of extreme responding, it strengthens the argument to use an index from heterogeneous items to investigate extreme responding in the first place.

In addition to the validation of the method, the features of PC trees in combination with the characteristics of our datasets allow us to investigate the following questions about the general nature of ERS:

Impact of ERS Compared to DIF Induced by Sex and Age One positive feature of PC trees is that they can handle multiple covariates at the same time. When more than one covariate is available, the variable responsible for the strongest parameter instability

is used for splitting first. Thus, the impact of ERS in a dataset can be compared to potential DIF, induced by other demographic variables. For all analyses, we consider sex and age as additional covariates. These variables are known to have a big impact on the measurement of psychological constructs and are routinely investigated when considering measurement invariance. If it turns out that an ERS index can have a stronger impact on the measurement model than sex and age, which are considered very important factors in psychometric research, ERS should be more stringently controlled for in psychological measurement. On a side note, we also register how the ERS index from heterogeneous items is associated with sex and age, contributing to a debate based on highly inconsistent findings in the literature.

Suitability of the Categorical Interpretation of Extreme and Mid Responders The investigation of ERS with mixed Rasch models stresses a conceptualization of ERS as two separate groups of people, one preferring extreme and the other mid responding. This is in contrast to concepts that depict ERS as a continuous attribute, ranging from mid to extreme responding. Our approach can differentiate between these two ideas. If ERS is appropriately described by a dichotomous model, we expect PC trees to split based on an ERS index only once. Thresholds in the resulting two groups should mirror the pattern of two latent classes repeatedly found by mixed Rasch methodology. If ERS is better described as a continuum, we would expect to see several splits revealing varying degrees of extreme and mid responding.

Consistency of ERS Across Scales and Across Time We expect to detect ERS in all scales of the NEO-PI-R, in which ERS has already been detected by mixed Rasch models. Additionally, we investigate scales from two different psychological questionnaires in the panel dataset. These scales differ in their measured construct, their general item design, and their response format. If we also detect ERS in these scales, we can rule out that ERS shows a distinct pattern in measures of Big Five personality factors, or depends on the specific questionnaire format of the NEO inventories. Due to the longitudinal nature of the GESIS panel dataset, we can investigate the stability of ERS over time. The ERS index is computed from items of different panel waves. If our analyses detect ERS in the panel dataset, this means that response patterns in a questionnaire at a certain wave, are associated with extreme responding at earlier and later time points. This would corroborate findings, suggesting that the ERS response pattern is consistent over time.

Meaning of Extreme Responding While most researchers believe that ERS reflects a person's probability to endorse the highest or lowest category of a rating scale, ERS can also be thought to reflect the general extremeness of a person's item responses. If we observe that subjects with higher values of the ERS index are more likely to endorse not only the most extreme response categories but also categories next to them, this would be evidence against the notion that the impact of ERS is limited to the lowest and highest response categories.

2.2 Methods

2.2.1 Description of the Datasets

NEO-PI-R

In the first analysis, we used the German nonclinical normative sample of the NEO-PI-R (Ostendorf & Angleitner, 2004), which contains 11724 subjects. This is the same dataset that was analyzed by Wetzel, Carstensen, and Böhnke (2013). They found strong evidence for ERS in the NEO-PI-R with mixed Rasch models (see chapter 1.3.1).

The original NEO-PI-R by Costa and MacCrae (1992) as well as its German adaptation are widely used to measure the Big Five factors of personality: neuroticism, extraversion, openness, agreeableness, and conscientiousness. The NEO-PI-R consists of 240 items. Each Big Five factor is divided into 6 more homogeneous facets, measured by 8 items respectively. Items are rated on a fully labeled five-point Likert scale ("strongly disagree", "disagree", "neutral", "agree", "strongly agree").[3]

Due to missing values, we could only use a subset of the whole normative sample. 11714 subjects had complete data for at least one of the PC tree analyses. This subsample contained 4216 men and 7498 women between the age of 16 and 91 ($M = 29.92$, $SD = 12.08$).

GESIS

In the second analysis, we used data from the longitudinal GESIS panel (GESIS, 2015)[4], hosted by the GESIS - Leibniz-Institute for the Social Sciences in Mannheim, Germany. The panel is based on a representative sample of the German population, recruited during 2013 in face-to-face interviews. These were followed by welcome surveys, conducted between June 2013 and February 2014. Since February 2014, panelists attend bimonthly surveys of about 20 minutes. They either attend online (62% of panelists) or by mail (38% of panelists), depending on which mode they chose during recruiting. With each survey, panelists receive five euros, regardless of completion. Surveys contain a longitudinal core study of the GESIS institute, as well as external studies that researchers have to submit to a peer-review process in order to get accepted.

Panel data can be requested from the GESIS institute for scientific use. At the time of our data request in February 2015, panelists had completed five regular survey waves. During the recruiting phase there were three additional waves in 2013, conducted with only a subset of the final sample. We do not have access to these waves as they contain variables under privacy protection not included in the standard version of the dataset.

For the first regular wave in February 2014, 4888 active panelists (2356 men and 2532 women), between the age of 18 and 70 ($M = 45.13$, $SD = 14.58$) were invited to participate. Due to missing values, we could only use a subset of these participants for our

[3] German labels: "starke Ablehnung", "Ablehnung", "neutral", "Zustimmung", "starke zustimmung"

[4] For more information, visit http://www.gesis.org/en/services/data-collection/gesis-panel/.

analyses. Complete data for at least one of our PC tree analyses was available for 3835 subjects. This subsample contained 1839 men and 1996 women between the age of 18 and 70 ($M = 46.34,\ SD = 14.24$).

2.2.2 Scales

NEO-PI-R

We analyzed all scales of the NEO-PI-R for which Wetzel, Carstensen, and Böhnke (2013) showed superior fit of the constraint mixed Rasch model compared to the unconstraint mixed Rasch model. We further reduced our analysis to a subset of facets for which two latent classes, that could be interpreted as extreme and non-extreme response style, were found to be most appropriate based on information criteria. This resulted in nine scales: Angry Hostility (N2), Impulsivity (N5), Assertiveness (E3), Activity (E4), Excitement-Seeking (E5), Openness to Feelings (O3), Compliance (A4), Order (C2), and Self-Discipline (C5). Notably this selection includes at least one scale from each of the Big Five factors.[5]

GESIS

In the GESIS dataset, we considered all multi-item scales with Likert response format in our analyses. The measured psychological constructs had to be one-dimensional at least in theory, to make sure that fitting a PCM was reasonable. Only two questionnaires in our dataset fit these criteria:

Positive and Negative Affect Schedule The Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) is the most widely used questionnaire to measure self-reported state or trait affect. The PANAS contains two theoretically uncorrelated factors, although this conceptualization remains controversial (Crawford & Henry, 2004; Schmukle, Egloff, & Burns, 2002): the Positive Affect scale (PA), involving pleasant emotions like feeling enthusiastic, active, or alert, and the Negative Affect scale (NA), a general dimension of subjective distress involving feeling nervous, afraid, or upset. In the GESIS panel, the German translation of the PANAS by Krohne et al. (1996) was used. Each scale consists of ten adjectives rated on a fully labeled five-point Likert scale ("very slightly or not at all", "a little", "moderately", "quite a bit", "extremely").[6] The PANAS can be used with different time instructions. In our dataset, subjects were asked about their feelings in general. Notably, the study documentation states that the PANAS was included in the GESIS panel to investigate whether measurement invariance holds for different groups of sex, age, and education in a general sample of the German population. PANAS items were included in the second regular panel wave, collected from April to June 2014.

[5] We limited our analysis to this subset of scales because we were not interested in comparing our PC tree approach to the mixed Rasch methodology directly. As a result, we do not address the possibility that our PC tree approach might detect ERS also in those scales for which the evidence from mixed Rasch methodology has been inconclusive.

[6] German labels: "gar nicht", "ein bisschen", "einigermaßen", "erheblich", "äußerst".

Questionnaire of Spatial Strategies The German Questionnaire of Spatial Strategies (Fragebogen Räumliche Strategien, FRS; Münzer & Hölscher, 2011) gathers self-reports about spatial navigation. Its final version consists of three correlated dimensions: The ten-item scale Global/Egocentric Orientation (GEO), assessing the ability to orient oneself based on knowledge of directions and routes, the seven-item scale Allocentric/Mental Map (AMM), capturing navigation strategies that rely on the formation of a mental map, and the two-item scale Cardinal Directions, requesting participants to judge their ability to guess cardinal directions. In our dataset, all items were rated on a seven-point Likert scale with labeled extreme categories ("fully disagree", and "fully agree").[7] The study documentation states that the FRS was included in the GESIS panel to obtain normative data about self-reported sense of direction and spatial strategies in the German general population. Notably separate norm data for men and women were planned as men reported higher orientation abilities as women in the original publication (Münzer & Hölscher, 2011). The authors reasoned that this does either reflect actual ability differences, or is the result of gender stereotypes. The finding might also indicate that the construct validity is not invariant between men and women. FRS items were included in the first regular wave of the GESIS panel, collected from February to April 2014.

2.2.3 Indices of Extreme Responding

ERS indices in both datasets were computed as follows: The first step was to find a set of 30 items that correlated among themselves as weakly as possible. This was determined by the test statistic of the Bartlett test of sphericity (Bartlett, 1951).[8] The Bartlett test compares the correlation matrix from a set of variables to the identity matrix, which implies zero correlations between variables. The smaller the test statistic, the closer the observed correlation matrix is to the identity matrix. For our analyses, the final item sets were selected by minimizing the Bartlett test statistic with a handcrafted version of simulated annealing. Technical details are described in Appendix A.

In a second step, each of the 30 selected ERS items was recoded to a response style scale, where extreme categories were coded with one, the middle category was coded with zero and all other categories were coded equally spaced in-between. Note that item responses on the new scale reflect how extreme the original responses are in terms of ERS. In the new scale, the direction of responses is ignored. Thus it does not matter if the original responses reflect agreement or disagreement with the question. Finally, an ERS score for each subject was computed as the mean of all item responses on the recoded ERS items. An ERS index was not computed for a participant if more than 10 responses on the 30 ERS items were missing. We did not require participants to have complete responses to avoid unnecessary reduction of sample size.

The result of this procedure is an approximately continuous index about the extremeness of subjects' responses on a variety of comparatively heterogeneous items. An ERS

[7] German labels: "trifft überhaupt nicht zu", "trifft vollkommen zu".

[8] In the computation of the correlation matrix for the Bartlett test, missing values were deleted pairwise.

index takes values between 0 and 1. It is 0 if the subject chose the mid category on all of the 30 ERS items. It is 1 if one of the most extreme categories was chosen on all ERS items. For example, an ERS index of exactly 0.5 would result if the second or forth response category was endorsed for all ERS items.

NEOPIR Dataset

When computing any response style index, it is essential that the included items are not only uncorrelated among themselves, but are also unrelated to the items of the primary scale. As the NEO-PI-R dataset does not contain any additional items except demographic variables, the ERS index had to be based on NEO-PI-R items. Obviously, items from the scale to be analyzed should not be included into the ERS index. Considering that facets belonging to the same Big Five factor also highly correlate given factor construction, all items of the same Big Five factor as the scale to be analyzed should not be included into the ERS index. As a result, five different ERS indices were computed, one for each Big Five factor. From now on, we refer to these indices as neuroERS, extraERS, openERS, agreeERS and conscERS. For example, when the subscale A4 of the factor Agreeableness was analyzed, the agreeERS index was used, which only contained items of the Big Five factors neuroticism, extraversion, openness, and conscientiousness. As all NEO-PI-R items were rated on a five point Likert scale, the recoded values on the response style scale were: 1, 0.5, 0, 0.5, 1.

GESIS Dataset

We identified 264 items as possible candidates to be selected for the ERS index. These included items from the welcome survey and the five regular panel waves that were at our disposal. All items were measured on a Likert scale with at least four response categories, so that both extreme and non-extreme responses were possible. Items could have an even or odd number of response categories. For some items all response categories were labeled, for others only the most extreme response categories were labeled. Items were ignored if the response categories contained precise quantifications of time, amount, or frequency. We did not include questions from the face-to-face recruiting interview, as we were only interested in the effect of ERS in self report questionnaires. Notably, as the number of response categories were not the same for all ERS items, the recoded values on the response style scale also differed. Moreover, as not all items contained a mid category, an ERS index of zero was not possible in this dataset.

2.2.4 Statistical Analyses

For all ERS indices, histograms and descriptive statistics were computed. These included the mean and the standard deviation of the index, the mean absolute correlation between items within the index, the maximum absolute correlation between items within the index,

and 95% confidence intervals for the correlations of the ERS index with sex and age. Moreover, correlations between the five ERS indices were computed. All reported correlations are Pearson product-moment coefficients.

Two PC trees were computed for all scales described in the section above. In the first tree, the corresponding ERS index was the only covariate. In the second tree, sex and age were included as additional covariates. PC tree analyses are only presented graphically, as more detailed information is unnecessary to answer the research questions detailed in section 2.1.2.

Pairwise deletion was used for all PC tree analyses and sample sizes are reported for each tree separately.[9] Although listwise deletion would be preferable for simplicity reasons, it would have resulted in a non-acceptable loss of observations (several thousand subjects in the NEO-PI-R and several hundreds in the GESIS dataset). Descriptive analyses are reported based on all subjects that could be included in at least one of the PC tree analyses. Note that as a result, computations of single coefficients in tables 2.1 and 2.2, correlations between ERS indices in the NEO-PI-R, as well as descriptions of demographic variables sex and age in the methods section, are based on slightly different samples. Because of the large overall sample size, descriptive statistics were highly stable. Therefore, different sample sizes for descriptive analyses can be safely ignored. For easier understandability we thus report the exact number of subjects only for the main analyses.

The significance level for parameter instability tests in all PC trees was set to 0.05. The minimum number of observations, required in a tree node to permit further splitting, was set to 200. Fitting PCMs to smaller samples would lead to threshold parameter estimates with too low precision.

A grid of barplots was computed for each PC tree, to better understand the response patterns in different leaves. Relative response frequencies of each response category are shown for each combination of leaf and item. In this way, item response patterns can be directly observed without the additional abstraction layer of PCM parameters.

To further illustrate the structure of threshold patterns in the PC tree analyses of the NEO-PI-R dataset, we additionally present threshold plots for the respective PC trees. In these line plots, the mean ERS index in each tree leaf is plotted against the estimated thresholds of the respective PCMs. Thresholds are represented by colored lines, while items are depicted in separate plot panels. Threshold plots were not created for the GESIS dataset, as the number of leaves in the computed PC trees were too small for an insightful visualization. Moreover, threshold plots were only created for a small subset of all analyzed NEO-PI-R facets. It can occur that not all response categories are chosen at least once for each item in the leaves of a PC tree. With the software package we used, it is currently not possible to fit all threshold parameters of an item if the highest or lowest response categories are missing. This happened frequently in both datasets. In such cases, items with missing categories were automatically treated as if their response scale consisted only of the non-missing categories. As a result, some threshold parameters were not estimated.

[9] Complete responses on all items of the analyzed scale and the included covariates are needed for PC tree analyses, which cannot deal with missing values.

This poses no problem for our interpretation of the PC trees, as we were interested in the general patterns. However, missing thresholds are problematic for the interpretation of threshold plots, as we were interested in the distribution of thresholds across leaves. Thus, threshold plots were not computed for PC trees with missing threshold parameters. All statistical analyses were conducted in `R` (R Core Team, 2016b). PC trees and tree plots were computed with the `psychotree` package (Zeileis, Strobl, Wickelmaier, Komboz, & Kopf, 2016). All remaining plots were created with the `ggplot2` package (Wickham & Chang, 2016). For the selection algorithm of uncorrelated items, Bartlett test statistics were computed with the `psych` package (Revelle, 2017), and beta-binomially distributed random numbers were drawn with the `emdbook` package (Bolker, 2016). The `irace` package (López-Ibáñez et al., 2015) was used for tuning the selection algorithm. Additionally, we used the packages `haven` (Wickham & Miller, 2016), `plyr` (Wickham, 2016a), `reshape2` (Wickham, 2016b), `tidyr` (Wickham, 2017), `mvtnorm` (Genz, Bretz, Miwa, Mi, & Hothorn, 2016), `progress` (Csárdi & FitzJohn, 2016), `knitr` (Xie, 2016), and `xtable` (Dahl, 2016).

2.3 Results

2.3.1 NEO-PI-R Dataset

Table 2.1: Descriptive Statistics of ERS Indices in the NEO-PI-R Dataset

ERS Index	M	SD	MeanAC	MaxAC	r_{Sex} CI	r_{Age} CI
neuroERS	0.50	0.11	0.04	0.17	[0.02, 0.06]	[0.00, 0.03]
extraERS	0.49	0.10	0.05	0.16	[0.02, 0.06]	[-0.02, 0.02]
openERS	0.49	0.10	0.05	0.18	[0.04, 0.07]	[0.03, 0.06]
agreeERS	0.49	0.10	0.05	0.18	[0.00, 0.04]	[-0.01, 0.03]
conscERS	0.48	0.11	0.05	0.17	[0.03, 0.07]	[-0.01, 0.03]

Note. 95% confidence intervals are presented for the correlations of the ERS indices with sex and age. Females are coded as one, males are coded as zero. Pairwise deletion was used to compute absolute correlations between items within ERS indices. Descriptive statistics were computed for all subjects that appear in at least one of the PC tree analyses ($N = 11714$). MeanAC = Mean of absolute correlations between all items within an ERS index. MaxAC = Highest absolute correlation between items within an ERS index.

As illustrated in Table 2.1, descriptive statistics were highly similar for all five ERS indices of the NEO-PI-R dataset. Means of all indices were close to the midpoint of possible ERS values. Figure 2.1 shows a histogram of the neuroERS index. Histograms for the remaining ERS indices looked very similar and can be found in Appendix B.2. Distributions of ERS indices were clearly unimodal, symmetric, and approximately normal. Our automatic algorithm was able to find sets of highly uncorrelated items. The mean of absolute bivariate correlations between all items in an ERS index was very low, as was the highest absolute bivariate correlation. ERS indices showed only negligible correlations

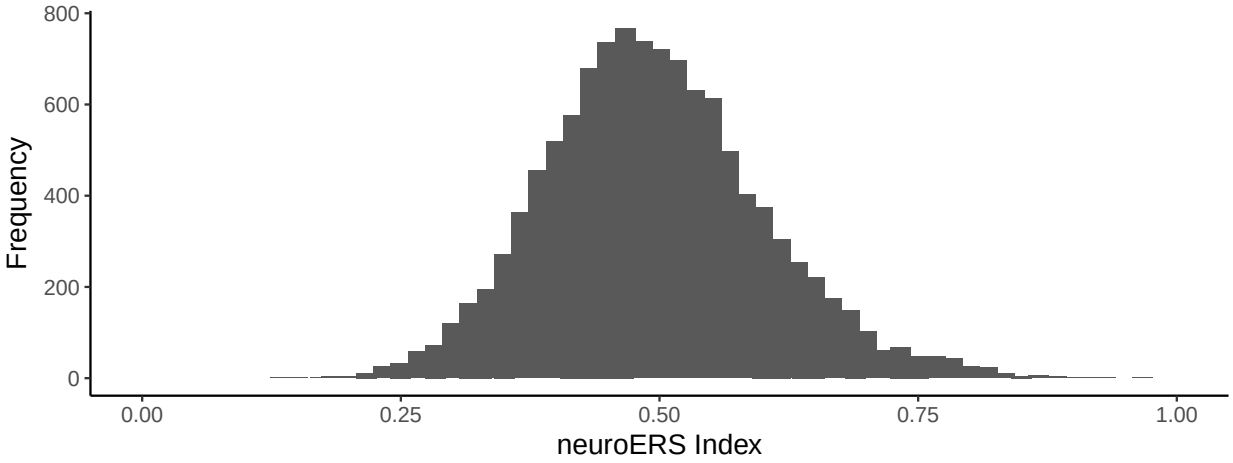

Figure 2.1: Histogram of the neuroERS index computed from heterogeneous items (except neuroticism) in the NEO-PI-R dataset.

with sex and age.[10] The five ERS indices were highly correlated, with correlations ranging between 0.89 and 0.93. For the majority of ERS items in all indices, response distributions were left-skewed.[11] The largest imbalance was noted for the extraERS index with 21 left-skewed and 9 right-skewed items.

[10] Scatterplot smoothing revealed no sign of a non-linear relationship between ERS indices and age.
[11] All items were coded in line with their measured Big Five trait.

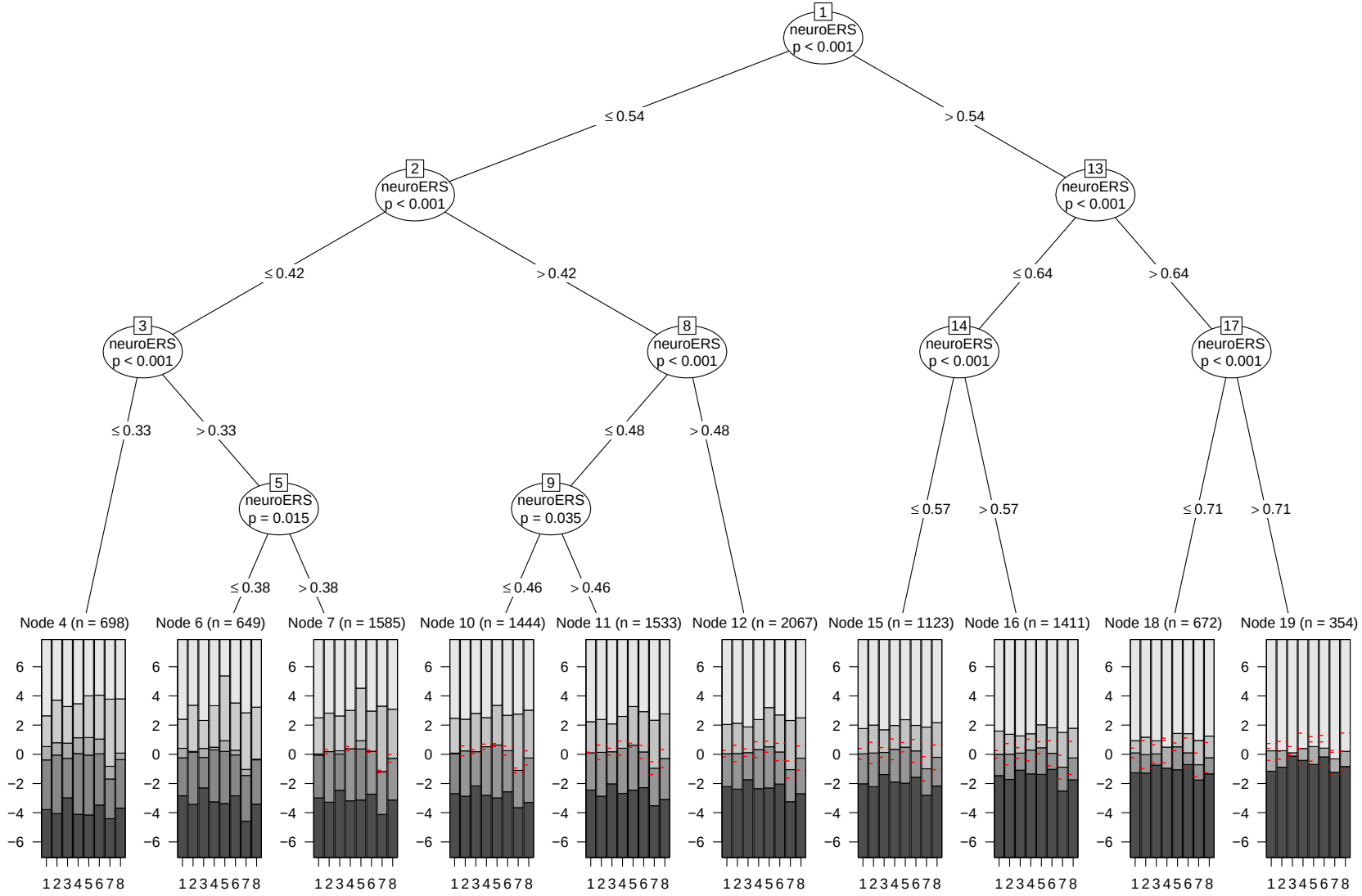

Figure 2.2: PC tree of the NEO-PI-R facet Angry Hostility (N2) with the neuroERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11536$.

The PC tree for the facet Angry Hostility (N2) with neuroERS as single covariate is presented in Figure 2.2. PC trees of the remaining facets looked very similar and can be found in Appendix B.3. For all analyzed facets, a single PCM was not appropriate to fit the item response data. The respective ERS index was repeatedly selected to split the dataset into smaller subsamples, resulting in PC trees with a very high number of leaves. Note that for the first split in each tree, the split point for the ERS index corresponding to the highest parameter instability in the PCM was close to the mean ERS index of about 0.5. Regarding the region plots in the leaves of the PC trees, a clear pattern emerged: Leaves on the left side of the tree contained subjects with low ERS values, while leaves on the right contained high ERS participants. The higher subjects' ERS values in a leaf, the smaller the distances between thresholds in the respective PCM. Moreover, regions on the latent trait where the highest and lowest response categories are most likely to be chosen are wider in leaves with higher ERS values. On the right side of the PC trees, the most extreme categories dominate the latent trait, showing the highest response probability across almost the whole range. However, this does not mean that the most extreme categories were chosen for the majority of item responses within these leaves. Figure 2.3 shows the actual relative frequency of response categories of the N2 facet, separately for each item and each leaf of the corresponding PC tree. Plots for the remaining NEO-PI-R facets can be found in Appendix B.4.

For all NEO-PI-R facets, a high number of unordered threshold parameters was observed. Threshold crossings were most severe in leaves with high ERS, whereas few or no unordered thresholds were found in leaves with very low ERS values. This can also be concluded from the threshold plot of the N2 facet in Figure 2.4. In this facet, thresholds are in the right order only for the leaves with very low ERS. In leaves with medium ERS values the second and third thresholds are interchanged and in leaves with very high ERS, even the first and fourth thresholds have the wrong rank order. Threshold plots for the facets Assertiveness (E3), Activity (E4), and Excitement-Seeking (E5) can be found in Appendix B.5. For all remaining facets, PC trees contained leaves with missing response categories, so no threshold plots were computed.

The pattern of narrowing thresholds and increasing regions of highest probability for the most extreme response categories was very consistent with regard to the increase in ERS across leaves. This can be more easily observed in the threshold plots. For all thresholds, an approximately continuous relationship between the mean ERS index in the leaves and threshold parameter estimates emerged. The functional relationship seemed close to linear for the first and fourth threshold. For the second and third threshold, graphs had a non-linear shape and did not monotonically increase or decrease for all PC trees.

When sex and age were added to the analyses as additional covariates, PC trees grew larger, such that leaves were nearly impossible to interpret. In each PC tree, both sex and age were selected as splitting variables at least once. However, the respective ERS index always contained the highest parameter instability and was selected for the first split in each PC tree. Figures of all PC trees including sex and age can be found in Appendix B.6.[12]

[12] Unfortunately, trees are too large to visualize the threshold patterns in single nodes on a standard A4

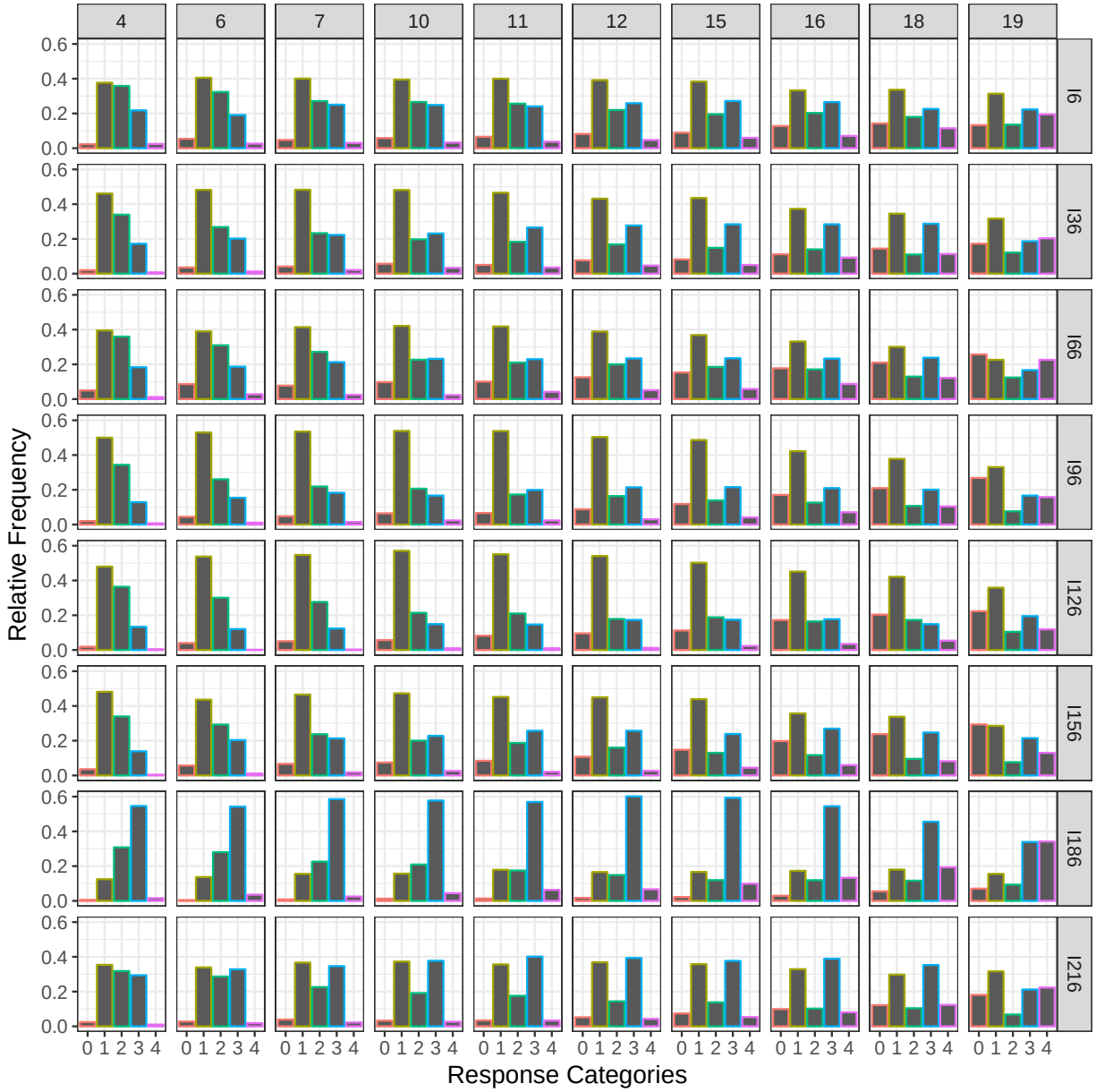

Figure 2.3: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the NEO-PI-R facet Angry Hostility (N2). Rows represent the items of the scale, numbered according to their position in the NEO-PI-R. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure 2.2. Response frequencies sum up to one in each cell.

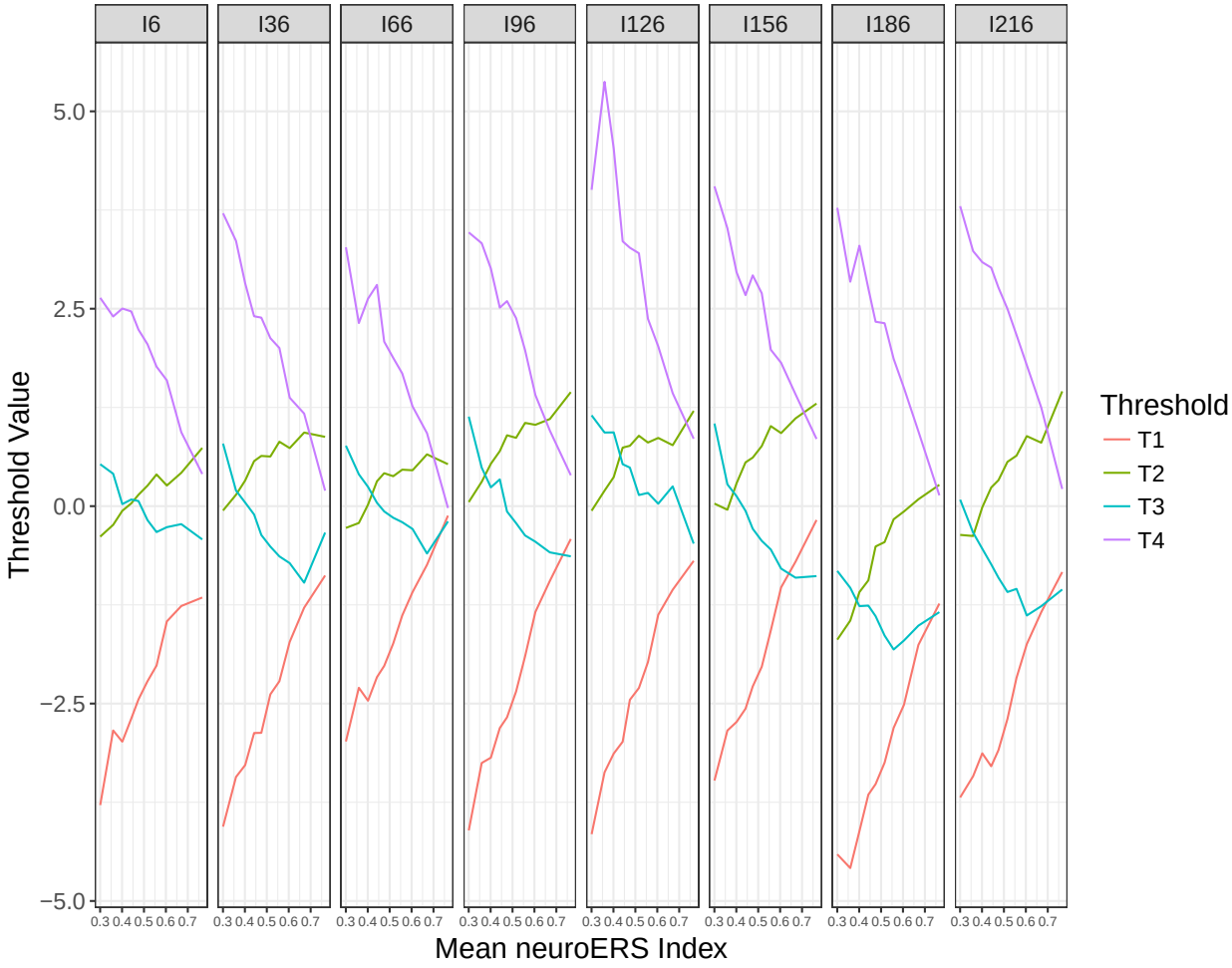

Figure 2.4: Threshold plot for the PC tree of the NEO-PI-R facet Angry Hostility (N2) with the neuroERS index from heterogeneous items as single covariate. Mean neuroERS values for each leaf of the PC tree in Figure 2.2 are plotted against parameter estimates of the respective partial credit model. Items are shown in separate panels. Thresholds are depicted as lines. $N = 11536$.

2.3.2 GESIS Dataset

Descriptive statistics of the ERS index in the GESIS dataset are presented in Table 2.2, while Figure 2.5 shows the corresponding histogram. Overall, the results are similar to the NEO-PI-R dataset. The mean of the ERS index was slightly higher than 0.5. The mean absolute correlation as well as the highest absolute correlation were lower than in the NEO-PI-R dataset. To get a better feeling for the items included in the ERS index, a list with the names of all items selected by the automatic algorithm can be found in Appendix B.1. ERS items from the welcome survey as well as all five regular panel waves were selected for the ERS index. Only one ERS item was from the same wave as the GEO and AMM scales (first wave), and two ERS items were from the same wave as the PA and NA scales (second wave). A majority of 22 ERS items were included in the regular panel waves three to five. Distributions of item responses were left-skewed for 15 ERS items, and right-skewed for 15 ERS items. ERS items had four different numbers of response categories (four-point scale: 3 items, five-point scale: 17 items, six-point scale: 2 items, seven-point scale: 8 items). Similar to the NEO-PI-R, associations of the ERS index with sex and age were negligible.[13]

Table 2.2: Descriptive Statistics of the ERS Index in the GESIS Dataset

M	SD	MeanAC	MaxAC	r_{Sex} CI	r_{Age} CI
0.54	0.09	0.03	0.13	[-0.03, 0.04]	[0.06, 0.12]

Note. 95% confidence intervals are presented for the correlations of the ERS index with sex and age. Females are coded as one, males are coded as zero. Pairwise deletion was used to compute absolute correlations between items within the ERS index. Descriptive statistics were computed for all subjects that appear in at least one of the PC tree analyses ($N = 3835$). MeanAC = Mean of absolute correlations between all items within the ERS index. MaxAC = Highest absolute correlation between items within the ERS index.

Figures 2.6 and 2.7 show the PC trees of the two PANAS scales, while tree plots of the two FRS scales are presented in Figures 2.8 and 2.9. Barplots showing the relative frequencies of response categories in different leaves can be found in Appendix B.4. For all scales in the GESIS dataset, less splits were performed than for the NEO-PI-R facets. The number of leaves per PC tree ranged between three for the NA scale of the PANAS and five for the AMM scale of the FRS. Despite the different number of leaves, the same threshold pattern emerged as in the NEO-PI-R dataset: the higher the ERS values in a leaf, the smaller the distance between thresholds, and the wider the regions of highest probability for the highest and lowest response categories. Compared to the NEO-PI-R

page. However, even under closer inspection, no consistent response patterns were found that would be relevant for the topics discussed here.

[13] Again, we found no sign for a non-linear relationship between the ERS index and age.

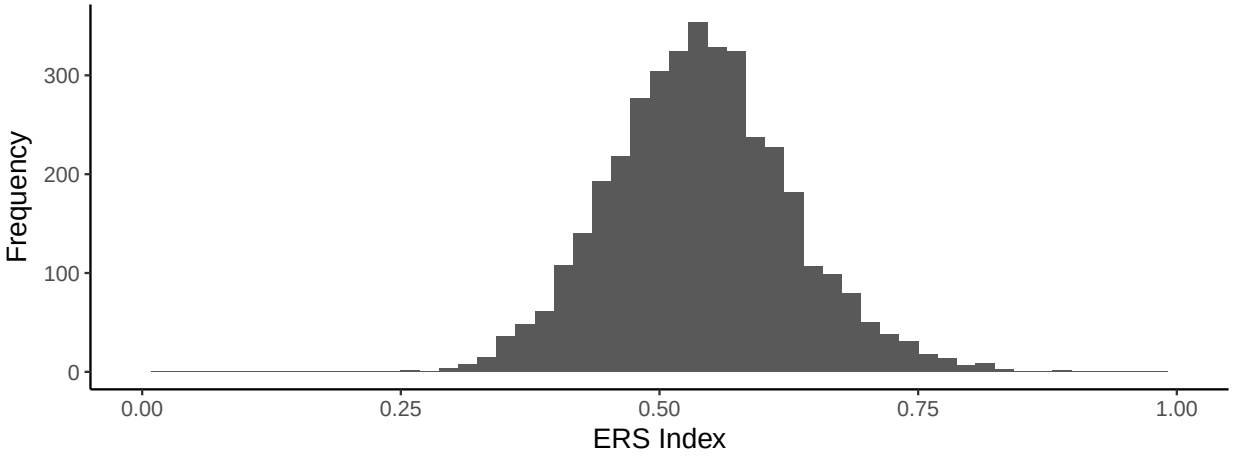

Figure 2.5: Histogram of the ERS index computed from heterogeneous items in the GESIS dataset.

facets, a smaller number of unordered thresholds was observed. Unordered thresholds were again more frequent in leaves with higher ERS values.

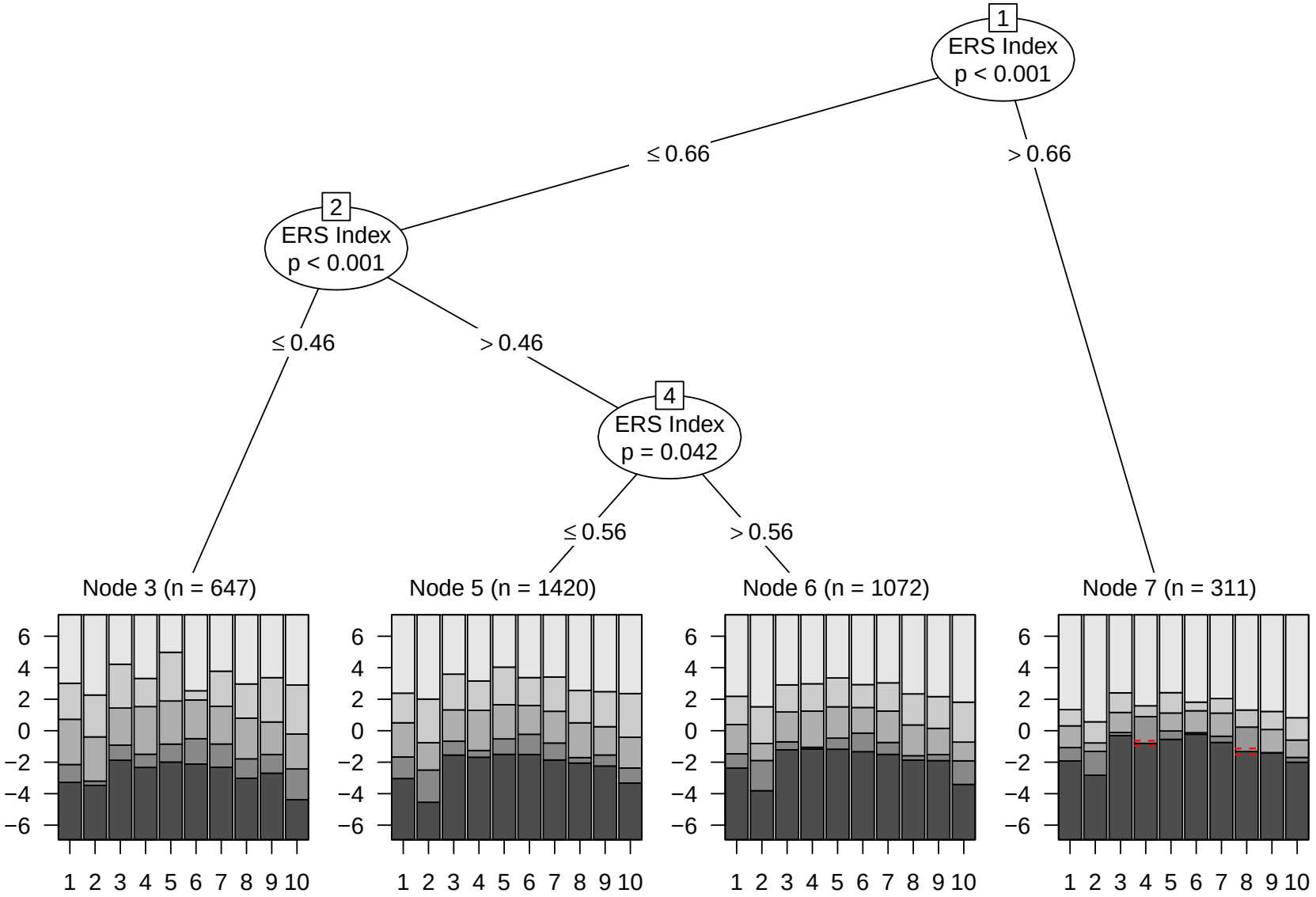

Figure 2.6: PC tree of the Positive Affect (PA) scale with the ERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 3450$.

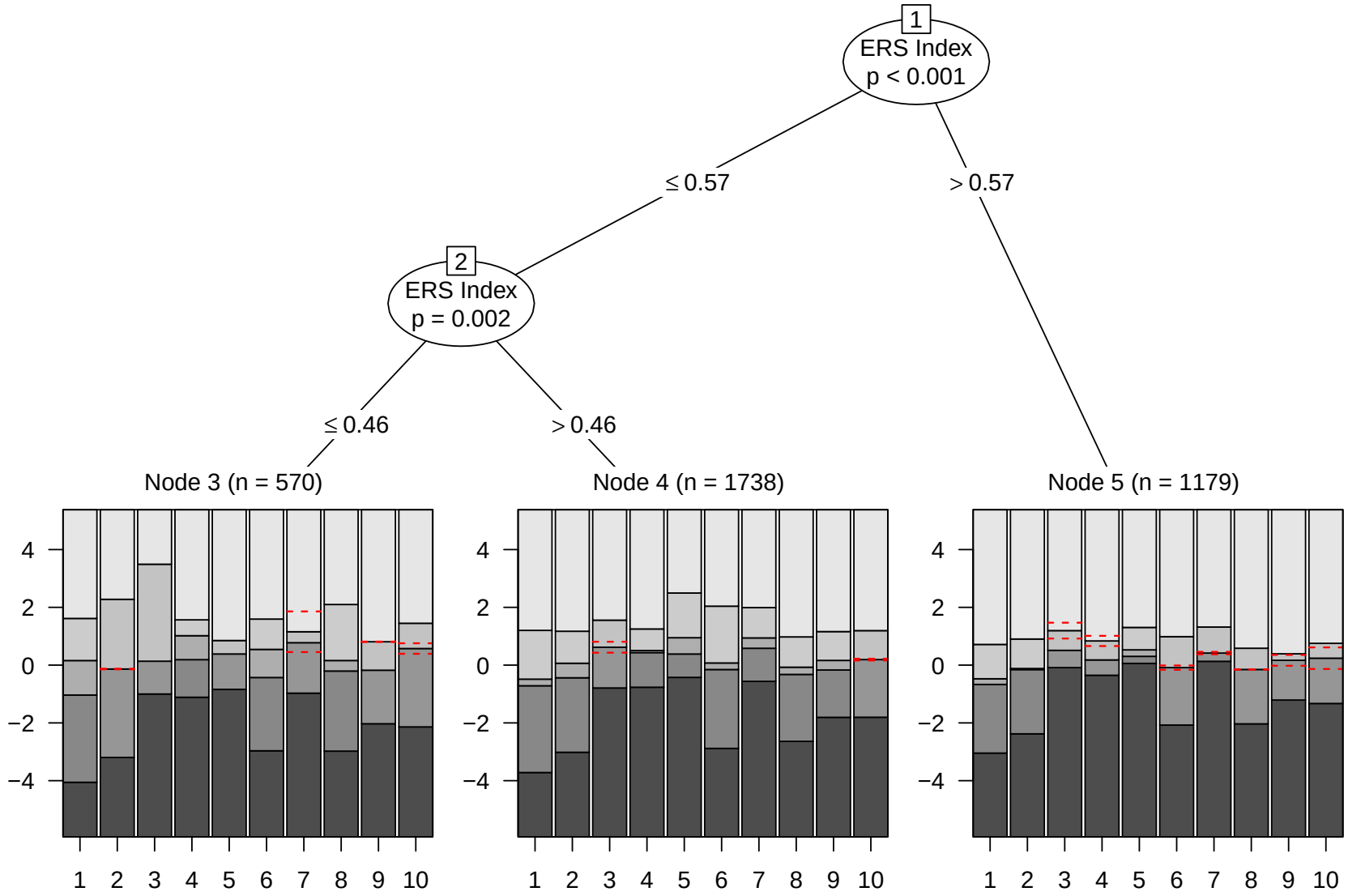

Figure 2.7: PC tree of the Negative Affect (NA) scale with the ERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 3487$.

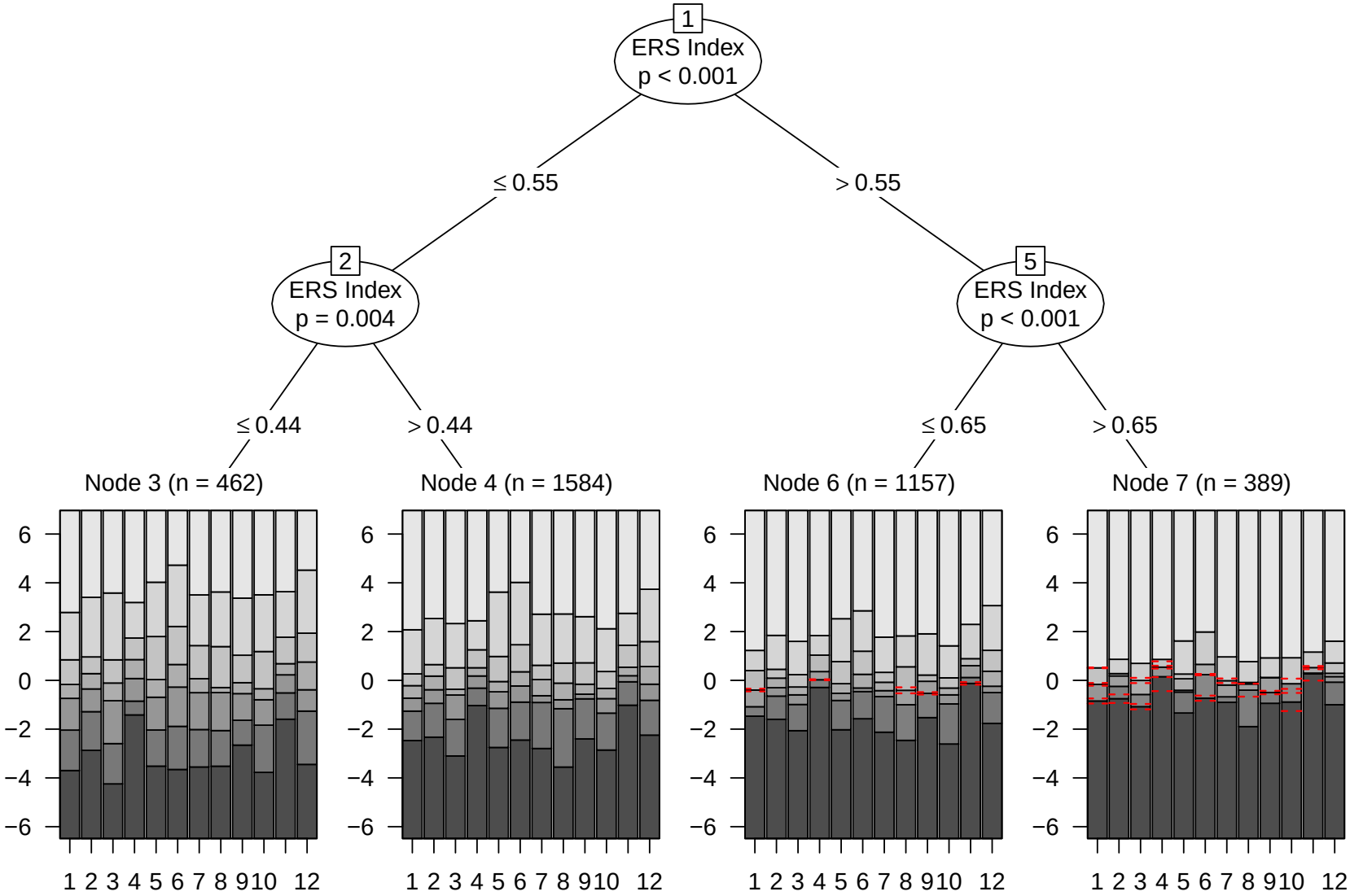

Figure 2.8: PC tree of the Global/Egocentric Orientation (GEO) scale with the ERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 3592$.

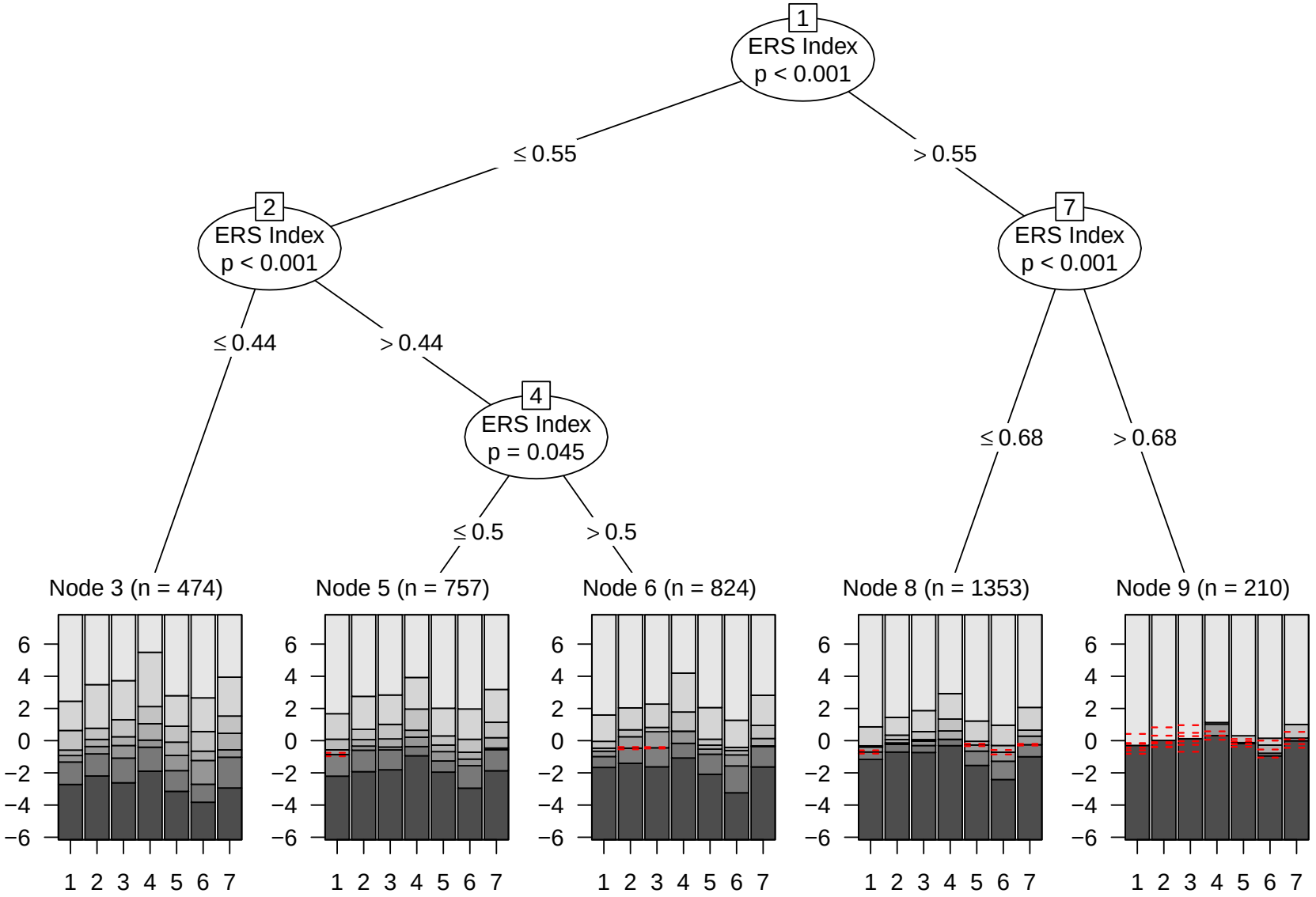

Figure 2.9: PC tree of the Allocentric/Mental Map (AMM) scale with the ERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 3618$.

When sex and age were added to the analyses as additional covariates, both demographic variables were selected as splitting variables in each PC tree except for the AMM scale of the FRS in which age was not selected. However, contrary to the NEO-PI-R dataset, the ERS index was selected for the first split only in the PC tree of the AMM scale. For the remaining scales, age contained the highest parameter instability. Again, figures of all PC trees including sex and age can be found in Appendix B.6.

2.4 Discussion

2.4.1 Summary of Results

Overall, results were highly similar for all analyzed scales. In the NEO-PI-R and in the GESIS dataset, sets of largely uncorrelated items could be found, leading to ERS indices with an unimodal, symmetric distribution. In both datasets, ERS indices were only negligibly related to sex and age. For PC trees with the ERS index as a single covariate, multiple splits were performed. Here, the final number of leaves was much higher in the NEO-PI-R than in the GESIS dataset. In all PC trees, the same characteristic threshold pattern emerged. The higher the ERS index, the smaller the distance between thresholds in the respective PCMs and the larger the regions of highest probability for the most extreme response categories. The threshold pattern was reflected in the actual item responses, showing that extreme responses were indeed more prevalent for subjects with high ERS values. When sex and age were added as additional covariates, at least one demographic variable was selected as splitting variable within each PC tree. For all PC tree analyses in the NEO-PI-R, the respective ERS indices were selected in the first splitting process. In the GESIS dataset, the ERS variable was selected in the first splitting process only once. For the remaining three analyses, age had the highest parameter instability.

2.4.2 Partial Credit Trees Are a Valid Method to Detect Extreme Response Style

Our analyses of two different datasets illustrate how PC trees (Komboz et al., 2016) in combination with an ERS index from heterogeneous items can be used to detect extreme responding in a variety of psychological measures. An effect of ERS was detected in both datasets and in all analyzed scales: the ERS indices were repeatedly selected as splitting variables in the PC trees. Our ERS indices from heterogeneous items were very similar to the ones previously reported in the literature (Greenleaf, 1992; Baumgartner & Steenkamp, 2001; Weijters et al., 2008). We could not compute ERS indices based on heterogeneous items that were included specifically for this purpose, as suggested by Weijters et al. (2008) and Greenleaf (1992). However, we used an optimization algorithm to successfully select sets of highly uncorrelated items from suitable items in our datasets. This is illustrated by the mean absolute correlations of ERS items, which are the smallest reported so far. ERS indices included left-skewed as well as right-skewed items. While the majority of ERS

items were left-skewed in the NEO-PI-R dataset, skewness was balanced in the GESIS data. Thus, indices indeed captured ERS and not acquiescence. If all ERS items were left-skewed, one could argue that our indices mainly represent a subject's tendency to agree with rating scale items (ARS), despite our continuous coding scheme of ERS items. With balanced numbers of left- and right-skewed items, this becomes unlikely. Additionally, bar plots with relative frequencies of response categories in each leaf clearly show that subjects with high ERS were more likely to give extreme responses compared to subjects with low ERS on both sides of the response scale. This was true even for items with heavily skewed response distribution. PCMs in the leaves of all PC trees showed a parameter pattern with narrow thresholds if ERS was high, and wide thresholds if ERS was low. This is reminiscent of the pattern repeatedly observed in investigations of ERS using the mixed Rasch modeling approach (Rost et al., 1997; Gollwitzer et al., 2005; Eid & Rauber, 2000).

Thus, our results strengthen the validity of both the ERS index or RIRS method to investigating ERS, as well as the mixed Rasch model approach. In their review on ERS, Vaerenbergh and Thomas (2013) stress that little is known about the convergent validity of different kinds of response style analyses. Our results suggest that ERS indices and mixed Rasch models both carry information about extreme responding. Moreover our approach contributes to the question of how both are related. When mixed Rasch models are used on their own, an important caveat is that the familiar threshold pattern may be confounded by processes unrelated to ERS. In the worst case, ERS groups might even be considered spurious classes (Alexeev et al., 2011). In our study, the same threshold pattern can be linked to an ERS index from heterogeneous items. Therefore, the present findings increase our confidence that threshold patterns in mixed Rasch models indeed reflect extreme responding. As explained earlier, ERS indices are a conservative way to avoid confounding of ERS measures by unrelated trait variance. However, if used on their own, they give little insight into the structure of ERS, compared to the theoretical understanding provided by the threshold interpretation of the mixed Rasch model. With our approach, we could not only verify that ERS indices are related to personal differences in responding. The threshold patterns in the leaves were also in line with our theoretical understanding of ERS. In this way, results confirm that combining PC trees with an ERS index from heterogeneous items retains the intuitive threshold interpretation from mixed Rasch models, and places it on a more solid foundation by linking it to an ERS index that guarantees minimal confounding of the response style measurement.

Another useful feature of PC trees is the ability to detect which covariate is responsible for the largest difference in response patterns. Recall that the covariate used in the first split carries the highest overall parameter instability. In the NEO-PI-R dataset, ERS indices induced larger model instability compared to demographic variables. This is in line with previous analyses suggesting that extreme responding is extremely prevalent in NEO inventories (Austin et al., 2006; Wetzel, Böhnke, Carstensen, et al., 2013; Rost et al., 1997). In the GESIS dataset, age seemed to have a larger effect on model stability than ERS. Unfortunately, this discrepancy is hard to interpret. One explanation might be the different sample characteristics of the two datasets. The normative sample of the NEO-PI-R is composed of a variety of convenience samples, with a high percentage of young,

female students. In contrast, the GESIS panel is based on a representative sample of the German population. This is reflected in a balanced sex distribution, a much higher age mean, and a higher age variance. It seems reasonable, that age effects in the NEO-PI-R were smaller due to this restricted age distribution. Note that in both datasets, ERS indices were only weakly associated with sex and age. In this regard, our results add to the highly inconsistent findings on the relationship of ERS with demographic variables in the extant literature(Vaerenbergh & Thomas, 2013).

Smaller effects of ERS in the GESIS panel were also observed with respect to the number of leaves produced by the ERS index. Both could be a result of different characteristics of ERS indices between the analyzed datasets. In the GESIS panel, items from different survey waves were included into the ERS index and few ERS items were selected from the same wave as the analyzed scales. While it is remarkable that ERS could be detected, it is reasonable to assume that the effects should be weaker under these circumstances. It has been previously found that ERS is highly consistent within a single questionnaire (Weijters et al., 2010a), with consistency increasing if items are closer together. Although ERS was found to be stable over large periods of time (Weijters et al., 2010b; Wetzel, Lüdtke, et al., 2016), we would expect ERS effects to be stronger in the NEO-PI-R dataset, where items were answered in close succession. Moreover, the design of the ERS items and all items from the analyzed scales was identical in the NEO-PI-R analysis. In the GESIS dataset, ERS items came from different questionnaires with varying response formats and principles of item wording. Another point worth mentioning is the difference in sample size between both datasets. Due to the included hypotheses tests, the PC tree approach to investigate ERS is a very conservative procedure. Its main appeal stems from good interpretability and high objectivity. As explained in the introduction, we can expect more splits for bigger samples, where the hypothesis tests of parameter stability have more power. Smaller effects of ERS in the GESIS dataset are in line with these considerations, as the sample size in the NEO-PI-R was nearly four times larger.

2.4.3 Extreme Response Style Is a Continuous Trait

PC tree analyses support the commonly held believe that threshold patterns in mixed Rasch analyses reflect ERS. However, our results cast doubt on a discrete nature of ERS, as inferred from the mixed Rasch methodology (Rost, 1990; Ziegler & Kemper, 2013). As stated in the general introduction, the conceptualization of ERS as discrete or continuous should be an empirical question. The current approach can be used to test this. Mixed Rasch analyses typically detected two ERS groups (Böckenholt & Meiser, 2017): a group of non-extreme responders with about two-thirds of participants, and a smaller group of extreme responders. If ERS were best described by this dichotomous structure, we would expect a one-split solution based on the ERS index in our PC tree approach. The two leaves should then mirror the latent groups of extreme and midpoint responders in the mixed Rasch model. Yet we clearly observed more than one split in all PC trees. Still, PC trees of NEO-PI-R facets had a much higher number of leaves compared to the analyzed scales in the GESIS dataset. We already mentioned several reasons why a smaller impact

of ERS was observed in the GESIS dataset, with smaller sample size probably being the most deciding factor. We would expect to observe a comparable number of leaves for the same sample size. The PC trees of the NEO-PI-R facets suggest that ERS has a continuous structure. The number of leaves was high and differences in threshold patterns for adjacent leaves were extremely small, suggesting a continuous functional relationship between ERS and thresholds in the PCM. Thus, it is unlikely that ERS is based on a high number of discrete classes. Patterns like the one observed here can be expected when continuous relationships are modeled with trees. If decision trees are used to model a linear function, the number of leaves is known to grow infinitely as sample size is increased (Tutz, 2011). Note that PCMs from different leaves of a PC tree are estimated independently and on separate subsamples. Therefore, PC trees can capture any functional shape of the thresholds across leaves in a data driven fashion (Komboz et al., 2016). Without an underlying continuous functional relationship, we would not expect the smooth relationship between ERS and threshold estimates which was revealed by the threshold plots.

The PC tree approach provides a conservative test against the two-class structure of ERS. In a strict sense it does not provide a certain test for whether ERS is continuous, because PC trees are discrete models by definition. However, we are convinced that our visualizations of the NEO-PI-R facets provide the best evidence for the continuity of ERS presented in the literature so far. Among others, continuous IRT models for ERS were proposed by Jin and Wang (2014) and Tutz et al. (2016). In both modeling approaches, it is possible to test whether ERS was detected. This indirect test for continuity, however, would be inferior to our analyses. Continuous models can be expected to detect signs of ERS, even if the structure of ERS were in fact discrete. After all, mixed Rasch models detected ERS in the NEO-PI-R (Wetzel, Carstensen, & Böhnke, 2013), although our analyses clearly provide evidence for a continuous structure.

It has been discussed to what extent ERS represents a personal trait (Aichholzer, 2013; Vaerenbergh & Thomas, 2013). The continuous IRT models by Jin and Wang (2014) and Tutz et al. (2016) share a trait-like interpretation of ERS. One requirement for ERS to be considered a trait is that the impact of ERS should be consistent across different content scales. Austin et al. (2006) showed that class membership in mixed Rasch analyses of the NEO-FFI was positively correlated between Big Five traits, with Spearman correlations up to 0.48. Wetzel, Carstensen, and Böhnke (2013) found ERS to be consistent across nine attitudes scales within the 2006 PISA assessment. They used a second order latent class analysis to show that subjects had a high probability to fall in the same ERS group of constrained mixed Rasch models in all scales. Additionally, they analyzed the same NEO-PI-R dataset like us, finding high consistency across NEO-PI-R facets in their latent class model. We also observed high consistency of ERS across facets in our PC trees. We can directly compare PC trees of facets belonging to the same Big Five factor, as they used the same ERS index. Not only was the general threshold pattern the same for all PC trees. The partitioning of the ERS index leading to the specific subsamples included in the leaves was also highly similar. These observations provide strong evidence for the consistency of ERS across NEO-PI-R facets, as subjects within a certain ERS range are included in leaves representing highly similar response patterns for all PC trees. For example, subjects with

neuroERS values of 0.65 were included in the second leaf from the right in the N2 tree and the third leaf from the right in the N5 tree. Both leaves show a comparable extreme response pattern, illustrated by both the threshold patterns in the tree plots as well as actual response frequencies. Moreover, ERS indices in the NEO-PI-R dataset were highly correlated at about 0.9. Thus, we can expect that roughly the same subjects are contained in leaves of PC trees from different Big Five factors.

Our NEO-PI-R analysis as well as previous studies about the consistency of ERS across traits mostly compared different traits within the same measurement instrument. Under these conditions one can expect higher consistency, as scales usually share many design characteristics. Moreover, traits within one instrument are connected by a common theory, otherwise they would not be part of the same measure. These limitations are not present for the GESIS analyses, which included two unrelated questionnaires, the PANAS and the FRS, with two subscales each. In addition to investigating consistency of ERS across traits within the same instrument, we can also compare ERS across traits from different questionnaires that vary strongly with respect to response format and question wording. Comparisons are facilitated by the fact that the same ERS index was used here. Due to the smaller number of splits, subsamples from individual leaves do not match across scales as well as in the NEO-PI-R data. Nonetheless, we find evidence for consistency of ERS in the GESIS dataset. Subjects belonging to subsamples showing extreme response patterns in one scale also did so in the other scales. For example, subjects with an ERS index below 0.44 were part of the left leaf in all four PC trees, showing the least extreme response pattern within each tree. Note that this consistency is invariant with respect to the actual distribution of item responses for the respective scale. Whereas relative response frequencies for the left leaf in the PA scale are highly symmetric with a high percentage of mid responses, items in the GEO scale are skewed to the left, and the left leaf of the AMM scale contains left as well as right-skewed items. Most interesting was perhaps the NA scale, where all items were heavily right-skewed with the lowest category being endorsed most frequently in all leaves. However, conditional on the skewed distribution of the NA scale, the left leaf still reflected a less extreme response pattern, illustrating that our approach delivers meaningful results even under extreme conditions. On a side note, our results also provide indirect evidence for the consistency of ERS across response scales, since the ERS index in the GESIS dataset included items with different numbers of response categories.

Another requirement of a trait is stability over time. One study found that an ERS index in a representative high school survey correlated at about 0.4 over a four year time period (Bachman & O'Malley, 1984). More recently, Weijters et al. (2010b) have presented evidence for one year stability of ERS, MRS, and ARS, based on a complex structural equation model including response style indicators from two sets of randomly sampled items. Applying the same methodological approach on a different dataset, Wetzel, Lüdtke, et al. (2016) even report stability of ERS over an eight year period. Although not our main focus, our analyses of the GESIS dataset indirectly support the stability of ERS over intermediate time intervals. Items from the welcome survey and all five regular study waves of the ERS panel were included in the ERS index, with consecutive waves being separated by two months. Only a small number of ERS items were from the same waves

as the analyzed scales. The majority of ERS items were collected in later waves than the PANAS and the FRS. We found robust patterns of ERS in all scales. This means that responses to heterogeneous items from earlier and later waves carried information about extreme responding in our homogeneous scales of interest. This provides further evidence for the stability of ERS.

2.4.4 Direct Modeling of Extreme Response Style

Our results provide strong evidence, that ERS should be regarded as a continuous trait. Moreover, the findings clearly suggest a smooth relationship between ERS and item thresholds, best illustrated by the presented threshold plots.

This structure is compatible with two continuous ERS models recently proposed by Jin and Wang (2014) and Tutz et al. (2016). Both are bi-dimensional IRT models, that extend the PCM by a second continuous person parameter, reflecting extremeness of responding. Threshold parameters in these models still reflect the general characteristics of each item. However, item response functions differ depending on the ERS parameter. The ERS parameter is directly linked to item thresholds of the model. In this sense, item responses are controlled by subject-specific thresholds, which are a combination of the item thresholds and the subject's ERS parameter. For subjects with high ERS, this leads to larger distances between adjacent thresholds compared to subjects with low ERS[14]. While the ERS parameter has an additive effect on item thresholds in the model by Tutz et al. (2016), it has a multiplicative effect in the model by Jin and Wang (2014). This poses different assumptions about the nature of ERS. In the additive model, ERS has the same absolute impact on all item thresholds. In the multiplicative model, ERS has a bigger effect on outer compared to inner thresholds, as the degree of expansion depends on the absolute threshold values.

In both models, ERS has an effect on all thresholds and no exclusive effect on the most extreme response categories. Therefore ERS is implicitly defined as the general extremeness of responses. Our empirical results support this design. When considering the actual relative response frequencies in the leaves of the PC trees, there is evidence that ERS is not a phenomenon only affecting the lowest and highest response categories of the rating scale. If this were the case, we would expect an increase of responses on the most extreme categories as ERS increases, accompanied by a consistent decrease of all other responses. We observed a steady increase in the most extreme response categories along with a decrease in midpoint responses. However, the remaining categories showed a non monotonic trend. Highest relative response frequencies emerged for medium ERS values. Thus, the overall pattern suggests that an increase of ERS shifts the whole response

[14] The domain of the ERS parameter is different between both models. In the model by Jin and Wang (2014), the parameter is strictly positive. Values below one reflect a higher tendency towards extreme categories while values greater than one reflect higher tendency towards the midpoint category. In the model by Tutz et al. (2016), the parameter lies on the real line. Negative numbers reflect a higher tendency towards extreme categories, while positive values reflect higher tendency towards the midpoint category

distribution towards both poles of the response scale, but does not have a specific effect on the most extreme response categories.

In general, we agree with Jin and Wang (2014) and Tutz et al. (2016) that it would be preferable to model ERS more directly. Our study shows that PC trees are a great exploratory tool to investigate the nature of ERS, providing insights about its continuity, among other things. However, PC trees are discrete models by construction, without useful structural assumptions between model parameters and covariates in the tree. As a result they are not well suited to control for ERS in psychometric applications. In principle, one could get estimates for the trait of interest which are adjusted for ERS, by estimating person parameters within the PCMs of the leaves. This would be reasonable when facing a discrete relationship between model parameters and ERS (Rost et al., 1997). However, for the continuous ERS effect we observed, this is highly unsatisfactory. Theoretically one would need one PCM for each separate value of the ERS index. To get a useful correction, a high number of leaves would be necessary. This requires an extremely large sample, as illustrated by our analyses. Also, it would be unclear how to choose the minimum amount of subjects within leaves. On the one hand, higher subsamples in the leaves would increase precision of parameter estimates. On the other hand, this would lead to less leaves thereby decreasing the effectiveness of the ERS correction. Even if a large sample were available, the approach would still lead to low precision in parameter estimation, as PCMs are estimated independently within each tree leaf, leading to an enormous number of model parameters, each estimated based on small subsamples. Continuous models directly implement assumptions about the relationship between ERS and threshold parameters, which is not possible for PC trees. As a result, these models are more parsimonious. The model is estimated in the whole sample, with only one additional parameter per subject, compared to the standard PCM. Most important, no additional items are required and ERS is estimated directly from the primary items of interest.

Our PC tree approach to investigate ERS provides promising evidence that the continuous structure in the models by Jin and Wang (2014) and Tutz et al. (2016) is a good model for ERS. From our analyses, it is not clear whether the additive or the multiplicative model yields a better approximation of ERS. Nevertheless, both models can be readily tested against each other, as they have the same number of parameters. For example, information criteria can be used, when estimating both models within the same marginal likelihood or Bayesian framework. Despite their high theoretical appeal, there is one strong caveat, that speaks against using these continuous models to control for ERS at this point: Using PC trees, we detected DIF on top of the confounding influence of ERS for all analyzed scales. Although we only included two demographic variables, at least one carried parameter instability within each PC tree. Like all multidimensional IRT models, continuous ERS models are highly flexible. If ERS is estimated from the same items as the primary trait, the second dimension might also capture multidimensionality, stemming from other sources than ERS. As a consequence, inferences about the primary trait are hard to interpret, especially if we already know that multidimensionality in addition to ERS is present.

Psychometric methods to model item responses are highly developed (Van der Linden,

2016). Unfortunately, measuring instruments in psychology lag behind the psychometric literature in ensuring unidimensionality of the primary trait. We are convinced that real improvements in psychometric measurement require more research about how to construct better measuring instruments. Questionnaires that precisely measure a psychological trait with little confounding by demographic variables and other psychological traits are required to further investigate ERS in a meaningful way. This issue will be continued in the general discussion in chapter 5.2. Considering that ERS has been repeatedly detected in the literature for a variety of psychological traits and item designs (Vaerenbergh & Thomas, 2013), extreme responding may be unavoidable, when using items with Likert scale formats to measure psychological constructs. Therefore, ERS may still be present for improved questionnaires and should be corrected for.

In this case, we propose the following procedure:

1. When validating a newly developed questionnaire, also collect possible DIF variables as well as a set of heterogeneous items to compute an ERS index.

2. Use the PC tree approach presented here with the ERS index from heterogeneous items and additional DIF variables as covariates.

3. If the ERS index is the only variable chosen to split the sample and the threshold pattern in the leaves is reminiscent of ERS, use continuous models of ERS like the ones by Jin and Wang (2014) or Tutz et al. (2016) to compute parameter estimates of the primary trait, which are corrected for differences in ERS between subjects.

Chapter 3

Study 2: Avoiding Extreme Response Style with Dichotomous Items

3.1 Introduction

3.1.1 Extreme Response Style and Dichotomous Response Formats

The previous study together with the large literature on response styles (Vaerenbergh & Thomas, 2013) might suggest that ERS is inevitable in psychological measurement when ordinal response formats are used. One possibility to avoid ERS which has been repeatedly voiced by experts in psychological test construction is to use questionnaires with dichotomous response format (e.g. Bühner, 2011). However, this recommendation is not guided by empirical evidence. It is rather based on the notion that an effect of ERS on dichotomous items seems "inconceivable" (Plieninger & Meiser, 2014) or even "impossible" (Cabooter et al., 2016). At first glance, it appears that extreme item responses do not exist with only two response categories. In a dichotomous item, a given statement can only be agreed or disagreed with. It is not possible to state different nuances of agreement. On the other hand, it has been suggested that response styles might just be harder to detect in dichotomous items (Dolnicar & Grün, 2007). Note that item response models for dichotomous items are also based on the concept of thresholds. The famous Rasch model (see for example (Van der Linden, 2016)) assumes that the probability to agree with a dichotomous item is only a function of the person ability and the item difficulty, which are both measured in the same unit. If the difficulty of an item is equal to a person's ability, the probability to agree with the item is exactly 0.5. The item difficulty can be interpreted as a threshold: if the person ability is higher than the item difficulty, agreement becomes more likely than disagreement and vice versa. The crucial point in the unidimensional models is that item difficulties are the same for all persons. If item difficulties depend on some characteristics of the respondent, we talk about differential item functioning (DIF). How could ERS have an effect on dichotomous item responses in this framework? Some

studies on the relationship between ERS and personality characteristics might suggest that ERS does not describe a purely methodological preference for specific categories of a rating scale, but reflects a person's general tendency to give extreme responses in a more general sense. It has been indicated that ERS is more prevalent for people who have been rated by their peers as high on intolerance for ambiguity, preference for simplistic thinking, and decisiveness (Naemi et al., 2009). Moreover, other ratings of humility have been linked to low ERS (Zettler et al., 2015). Using scale-free personality measures, extreme responding has been shown to be more frequent for promotion focused people (Cabooter, 2010), which includes a higher preference for making decisions and taking actions. Midpoint responses were linked to prevention focus, which includes tendencies to avoid failures and act conservatively. Cabooter et al. (2016) report that priming subjects with individual pronouns in contrast to collective pronouns leads to more extreme responses on a set of heterogeneous items, especially if items are related to how subjects view themselves.

We consider again the situation of a person whose value on the latent variable is equal to the item difficulty, as conceptualized in the Rasch model, and discuss how an effect of ERS might come into play: For a person with average ERS, the probability to choose each category would still be 0.5. However, for a respondent with high ERS, choosing the extreme category might be more likely than the less extreme category, while a subject with low ERS might be more likely to chose the less extreme category. The magnitude of the ERS effect should not be the same for all items, but depend on the difference in extremeness between the two response options. This still leaves the question, how extremeness could be defined in dichotomous item responses. The easiest conceptualization is that extremeness of a dichotomous item response is reflected by its relative response frequency in the population. For an item with average difficulty, none of the two response options is more extreme and ERS should have no effect. With excessively high or low item difficulties, the distinction which response is more extreme gets more pronounced. As a result, the effect of ERS should increase. One could argue, that the relationship between the general item difficulty and the magnitude of the ERS effect should be non-linear. For excessively high or low item difficulties, the attraction of the extreme response category might decline again. The extremeness of a dichotomous item response might also be determined by more complicated aspects. The literature on social desirability and evaluative neutralization emphasizes the fact that many items in personality inventories are highly evaluatively loaded. An item is considered evaluatively loaded, if its content heavily complies or disagrees with societal norms (Bäckström & Björklund, 2014). This might be intrinsically due to the described behaviors and attributes, or simply be a result of the chosen item wording. For items with high evaluative load, the less popular response might be considered more extreme. Although the most popular responses are often chosen most frequently, this is not necessarily so. It should always be possible to construct items about an unaccepted behavior which is so common, that a majority of people still endorses the undesirable response category. Thus, when considering evaluative load, ERS effects might be in the other direction than would be expected based on the relative response frequencies. Assuming that the evaluative loading of a dichotomous item is related to how the item is affected by ERS does not necessarily mean that ERS has to be a symptom of

social desirable responding. In particular when ERS is measured based on uncorrelated items with different skewness, it seems unlikely that the ERS index is confounded with social desirability. Also, we know of no studies suggesting a reliable association between social desirability and ERS.

Based on the above discussion several assumptions could be made: It seems theoretically possible that dichotomous items are also affected by ERS. Yet, it is not completely clear how the relationship should look like if it does exist. The most straightforward assumption within a relative simple IRT framework would be that item difficulties vary for people depending on their ERS values. While ERS seems to have an effect on all ordinal items, not all items of a dichotomous scale might be affected. If and how strong an item is affected might be non-linearly related to the general difficulty of the item, to its evaluative load or to some other content characteristics. While it would be most plausible that people with high ERS favor that item response which is less chosen by the respective population, it could also be the other way round. Furthermore, the direction of ERS for a dichotomous item might be guided by other aspects of extremeness like social desirability.

To investigate the speculative question if dichotomous items can be affected by ERS, a methodological technique is needed which can link item response probabilities to person covariates. The technique has to be flexible enough to capture the variety of structural aspects discussed above. In chapter 2.1.1, Rasch trees (Strobl et al., 2013) were introduced as a powerful tool to detect DIF based on person covariates. By using an ERS measures as covariate, the Rasch tree methodology makes it possible to investigate an effect of ERS in dichotomous items. In alignment with study 1, our primary operationalization of ERS is an index based on heterogeneous items with ordinal response format. As independent Rasch models are estimated for subsamples defined by a combination of covariates, the approach is flexible enough to model complex effects of ERS on the item difficulties. No functional relationship is assumed for which items are affected by ERS and which are not. The direction in which item difficulties are shifted for each item is also not restricted in any way. If an effect of ERS is found, item characteristics of the affected items can be easily checked against the item wording of the relevant items. This might give insights into the mechanisms of ERS in dichotomous items. Another psychometric technique which is well suited to investigate the effect of ERS in dichotomous items is the DIF Lasso by Tutz and Schauberger (2015). Before the DIF Lasso model will be introduced in the following chapter, we shortly address another view on how dichotomous item responses might be affected by ERS .

Thus far, our discussion assumed that ERS in dichotomous items can be conceptualized as an effect on the item difficulties. This yields a multidimensional model in which item difficulty varies by person. However, it could be deemed more plausible that item discriminations are affected instead of item difficulties. In the standard parametrization of the two parameter logistic item response model (2PL model; for a description see Van der Linden, 2016), the slope of an item characteristic curve (ICC) is determined by an item specific discrimination parameter. In the Rasch model, discrimination is the same for all items. Supposing that high ERS might reflect less ambiguity and clearer decision making, it is reasonable to assume higher item discriminations for high ERS respondents. Unfortu-

nately, we know of no existing IRT model which specifies an effect of person covariates on item discriminations. However, there is an indirect way to investigate this: By splitting a sample based on an ERS measure and fitting a 2PL model to each subsample, an effect of ERS on the item discriminations can be graphically explored. For example a median split of an ERS index based on heterogeneous items could be used as splitting criterion. If high ERS is connected to better item discrimination, ICCs in the high ERS subsample should be steeper. This approach will be used as a supplemental analysis in the present study.

3.1.2 An Introduction to the DIF Lasso

The DIF Lasso by Tutz and Schauberger (2015) is based on the following general DIF model,

$$\log\left(\frac{P(X_{pi}=1)}{P(X_{pi}=0)}\right) = \theta_p - (\beta_i + \boldsymbol{x_p^T}\boldsymbol{\gamma_i}) \tag{3.1}$$

which can be considered an extension of the dichotomous Rasch model. Similar to the Rasch model, the random variable $X_{pi} \in \{0, 1\}$ represents the response of person p on a dichotomous item i, and θ_p stands for the latent ability parameter of person p. However, the item parameter of an item i in the Rasch model is replaced by $\beta_i + \boldsymbol{x_p^T}\boldsymbol{\gamma_i}$, the sum of the items specific intercept β_i and the dot product $\boldsymbol{x_p^T}\boldsymbol{\gamma_i}$. The vector $\boldsymbol{x_p}$ contains the values of person p on a series of C covariates.[1] The item specific DIF vector $\boldsymbol{\gamma_i}$ quantifies the effect of the covariates on the difficulty of item i. The covariate vector $\boldsymbol{x_p}$ contains possible candidate variables which might induce DIF. Covariates can be both metric or categorical. If an item j does not contain DIF, the item specific DIF parameter vector $\boldsymbol{\gamma_j}$ contains only zeros. Then, the difficulty of the item is the same for all persons and represented by β_j. If none of the I items contains DIF, the general DIF model reduces to the Rasch model. Although the β_i are generally called difficulty parameters, the familiar interpretation from the Rasch model only holds for items without DIF. If an item j contains DIF, the item difficulty is given by $\beta_j + \boldsymbol{x_p^T}\boldsymbol{\gamma_j}$, which is not the same for two persons differing on a covariate c with $\gamma_{j,c} \neq 0$. For example, the difficulty of an item might be positively related to age or be higher for men than for women.

The general DIF model contains many parameters which complicates parameter estimation. Tutz and Schauberger (2015) fit the model with penalized maximum likelihood estimation (pMLE). While ordinary MLE would maximize the log-likelihood $l(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ denotes the total vector of parameters $\boldsymbol{\alpha} = (\boldsymbol{\theta^T}, \boldsymbol{\beta^T}, \boldsymbol{\gamma_1^T}, ..., \boldsymbol{\gamma_I^T})$, pMLE maximizes a penalized log-likelihood of the form

$$l_p(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda \sum_{i=1}^{I} \|\boldsymbol{\gamma_i}\| = l(\boldsymbol{\alpha}) - \lambda \sum_{i=1}^{I} (\gamma_{i1}^2 + ... + \gamma_{im}^2)^{1/2} \tag{3.2}$$

A penalty term $\sum_{i=1}^{I} \|\boldsymbol{\gamma_i}\|$ is subtracted from the log-likelihood, thereby punishing high values of the DIF parameters during maximization. The regularization parameter λ determines the impact of the penalty. For high values of λ, the penalty is strong and only

[1] Vectors are depicted in boldface.

small DIF parameters are acceptable. The penalty is a modification of the grouped lasso by Yuan and Lin (2006). It has the effect that for the optimal parameter configuration which maximizes the penalized log-likelihood, $\boldsymbol{\gamma_i}$ contains only zeros for some items. For the remaining items, all entries in $\boldsymbol{\gamma_i}$ are unequal to zero. Tutz and Schauberger (2015) call their approach the DIF Lasso, as it is capable to single out items which contain DIF. DIF items are those for which pMLE yields $\boldsymbol{\gamma_i} \neq \mathbf{0}$.

How many items are labeled to contain DIF depends on the regularization parameter λ. The higher λ, the smaller the number of DIF items. Consequently, finding the optimal value for λ is crucial for the success of the DIF Lasso. If λ is chosen too small, some items are falsely identified as DIF items. If λ is chosen too high, some true DIF items are not detected. Tutz and Schauberger (2015) fit the model for a series of decreasing λ parameters and choose the parameter configuration with the smallest Bayesian Information Criterion (BIC). This requires a measure for model complexity, which is a non trivial task in penalized estimation. In the original publication, model complexity in the BIC is operationalized with degrees of freedom, computed by the method of Yuan and Lin (2006). The software implementation offers an additional approach, called the L2-norm method. This method always leads to less or equal degrees of freedom than the method by Yuan and Lin (2006). Thus, the L2-norm method is the less conservative approach, as a higher number of items are potentially labeled to contain DIF. Since no method is more appropriate in general, it makes sense to always consider both solutions in practice.

After choosing the optimal value for the regularization parameter, the estimated item difficulties $\hat{\beta}_i$ of the final model have to be centered on one item j for which $\hat{\boldsymbol{\gamma}}_{\boldsymbol{j}} = \mathbf{0}$. It ensures that the model is identified. This approach in combination with the properties of the group lasso penalty and the decision to punish high model complexity by using the BIC, reflects the implicit assumption that most items of the analyzed scale are indeed Rasch-compatible. Only few items potentially contain DIF with respect to the applied covariates. Although this might be a sensible assumption in most circumstances, we will come back to this when discussing how the DIF Lasso can be used to detect effects of ERS.

To find out which items contain DIF with regard to which covariates, an obvious procedure is to take the model with the optimal value for the regularization parameter, and focus on the identified DIF parameter estimates $\boldsymbol{\gamma_i}$ of those items with $\hat{\boldsymbol{\gamma}}_{\boldsymbol{j}} \neq \mathbf{0}$. By standardizing all covariates in the DIF Lasso, the importance of any covariate can be measured by the absolute values of its DIF parameter estimates (Tutz & Schauberger, 2015).

As with all penalization approaches, the penalized DIF parameter estimates for true DIF items are downwardly biased in magnitude. To decrease this bias, Tutz and Schauberger (2015) propose an additional refit procedure. In a second step, they estimate the general DIF model with unpenalized MLE, but only include DIF parameters $\boldsymbol{\gamma_i}$ for those items which were diagnosed to contain DIF by the pMLE fit. However, the refit procedure increases the variability of the DIF parameter estimates, especially for truly Rasch-compatible items. DIF analyses are mostly used to improve the items of psychological questionnaires, or to justify the use of unidimensional models which are widely applied in practice. One is mostly interested in which items contain DIF, which covariates induce DIF, which covariates are more important than others, and what is the direction of the

various DIF effects. The numerical values of the DIF parameters are of minor importance, as the DIF model will not be used for measurement of the latent ability anyway. In the described scenario, only the penalized solution is necessary, as it provides all required information.

Due to the structure of the general DIF model and the penalization approach to select DIF items, the DIF Lasso can be used to model the hypothesized mechanisms of ERS discussed in chapter 3.1.1. When measures of ERS are used as covariates, the ERS effect on items difficulties does not have to be in the same direction for all items. Therefore, no assumptions about which response category is more extreme have to be made, as no functional relationship is assumed. Furthermore, it is not necessary that ERS has an effect on all items, or that the effect is equally strong for all selected DIF items.

The DIF Lasso serves a similar purpose to Rasch trees. Both methods are designed to detect DIF in scales of dichotomous items, based on a set of person covariates and can be used to investigate the effect of ERS on dichotomous item responses. However, Tutz and Schauberger (2015) discuss some important differences: Due to their tree structure, Rasch trees can easily detect interactions between covariates and highly non-linear relations. In the presented form, the linear predictor for the item difficulties in the DIF Lasso contains only main effects. Moreover, the DIF Lasso is currently not available for Likert scale items, which is the reason why DIF Lasso models were not used in study 1. However, it would be possible to add linear interactions, replace the linear predictor with smooth additive functions like splines, or extend the general model to more than two response categories (Tutz & Schauberger, 2015). A bigger difference between both approaches is on which level ERS is detected. The DIF Lasso explicitly selects DIF items but estimates DIF parameters on these items for all covariates. In contrast, Rasch trees identify combinations of DIF inducing covariates, but do not single out items which are primarily affected by these covariates. As a consequence, the tree approach might be most sensitive to detect DIF variables, which have a consistent effect on a majority of items. This was the case for the effect of the ERS index in study 1. The DIF Lasso might be more sensitive to detect DIF items, which are affected by a variety of covariates. This is most useful to single out bad items during the construction process of psychological test.

3.1.3 Aim of Study and Research Questions

The main goal of the present study is to investigate, whether ERS affects dichotomous item responses. Two dichotomous scales from widely applied psychological questionnaires are analyzed with DIF Lasso models and Rasch trees. In addition to sex and age, three measures of ERS are used as covariates in both modeling approaches to assess the impact of extreme responding. The primary measure is an ERS index from heterogeneous items with ordinal response format. Additionally, a binary self-report measure of ERS, and a measure based on the binary classification of a constraint polytomous mixed Rasch model are used. The polytomous mixed Rasch model is fitted to a scale with ordinal response format which has been shown to be affected by an ERS index in study 1. The ordinal scale is also used to assess the effectiveness of the ERS measures. Specific research questions

can be structured as follows:

Effectiveness of ERS Measures

Comparing the three ERS variables might give important insights into different aspects of extreme responding: We expect the ERS index of heterogeneous items to be correlated with the class assignment of a constraint mixed Rasch model which is based on a scale with ordinal response format. Study 1 suggested that both variables measure ERS but with different granularity. As a secondary research question, we are interested if subjects have conscious access to their own extreme response style. As it seems plausible that people might have at least some understanding if they rather have a tendency to choose midpoint or extreme categories, we also expect self-reported ERS to be related to the other ERS measures.

To make sure, that our ERS variables actually capture information about extreme responding, PC trees of the scale with ordinal response format are computed with the ERS measures as covariates. Effective ERS measures should induce splits in the PC tree if used as a single covariate. When both the ERS index from heterogeneous items as well as the self-report measure of ERS are included in the PC tree, their impact can be meaningfully compared, by noticing which one is selected as splitting variable first. Trivially, we would always expect that the ERS measure based on a constraint mixed Rasch model of the same scale induces the strongest parameter instability, as the mixed Rasch model is supposed to optimize the differences in response patterns between latent classes. For the ERS index from heterogeneous items, general threshold patterns in the leaves of the PC tree should be comparable to study 1. However, as the sample size is lower in the current study, a smaller number of leaves can be expected. Assuming that respondents are capable to report their own response tendencies, the binary self-report measure of ERS should also be selected as splitting variable in the PC tree. If the self-report measure indeed captures ERS instead of another DIF inducing attribute, threshold patterns in the resulting leaves should reflect groups of extreme and midpoint responders, comparable to the pattern revealed by the binary ERS measure of the constraint mixed Rasch model.

ERS in Dichotomous Scales

If dichotomous items do not suffer from effects of ERS, we expect that no DIF items are detected by the DIF Lasso for a dichotomous scale, when only ERS measures are used as covariates. When other potential DIF variables like sex and age are included, DIF parameter estimates of the ERS measures should be close to zero for all selected DIF items. Similarly, no measure of ERS should be selected as a splitting variable in Rasch trees, no matter if ERS measures are the only covariates or used in combination with sex and age.

Rasch trees of the dichotomous scales are a straightforward test of which ERS measure has the largest impact, as the covariate inducing the highest amount of parameter instability gets selected as splitting variable first. In the corresponding DIF Lasso analyses, the

absolute estimates of the DIF coefficients can be compared between ERS measures. In both modeling approaches, the impact of ERS can be compared to the demographic variables sex and age, which are likely candidates to induce DIF in any psychological measurement.

Depending on the true nature of ERS in dichotomous item responses, either Rasch trees or the DIF Lasso should be more sensitive to detect extreme responding, especially if ERS measures are the only covariates. If all dichotomous items are affected by ERS, as seems to be the case for ordinal response formats, an ERS measure might be selected as splitting variable in the Rasch tree but no DIF items might be detected by the DIF Lasso. If ERS does not have a consistent effect on all dichotomous items but interacts with characteristics of specific items, as discussed in chapter 3.1.1, DIF items might be detected by the DIF Lasso but no DIF might be detected by the Rasch tree. Moreover, if each of the three ERS measures has a non-zero but rather small effect, the DIF Lasso might be more sensitive compared to Rasch trees which evaluate each covariate separately.

3.2 Methods

3.2.1 Scales and Variables

A paper and pencil questionnaire was constructed, which contained two dichotomous psychological scales as well as several items which were used to compute three different measures of ERS:

Shyness

The 14 item scale Shyness (SHY) was introduced by Ben-Porath, Hostetler, Butcher, and Graham (1989) as a content-homogeneous subscale for the Revised Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher et al., 1989). It was created by factor analyzing the Social Introversion scale of the MMPI-2, and selecting items based on item-total correlations and an increase in Cronbach's Alpha (Cronbach, 1951). The SHY scale is quite homogeneous in content. It captures tendencies to act shy and feel uncomfortable in the company of other people, including an incapacity to speak with strangers or in front of others. Estimates for Cronbach's Alpha have been reported around 0.80 (Ben-Porath et al., 1989; Sieber & Meyers, 1992), making it one of the most homogeneous scales in the MMPI-2, despite its comparatively small number of items. Of all MMPI-2 scales, the SHY scale was considered most appropriate to be analyzed by a Rasch model, which assumes a single latent variable for all items. In our study, the German translation of the MMPI-2 by Engel and Hathaway (2003) was used. SHY items were rated on the original dichotomous scale with categories labeled as "wrong" and "right".[2] Six items of the SHY scale are negatively keyed.

[2] German labels: "falsch", "richtig".

Achievement Orientation

The 12 item scale Achievement Orientation (ACHI) of the German Revised Freiburger Persöhnlichkeitsinventar (FPI-R; Fahrenberg et al., 2001) contains statements about competitive behavior, approaching tasks with high ambitions and energetic vigor, and favoring challenges at work over leisure activities. The manual reports a Cronbach's Alpha for the ACHI scale of 0.78, which is an average value within the FPI-R. Nevertheless, the ACHI scale was considered most homogeneous in content of all FPI-R scales. In general, FPI-R scales are designed to be rather diverse (Fahrenberg et al., 2001). Analyses from the manual suggest that the Rasch model is generally not appropriate for any of the FPI-R scales, thus we can expect to detect DIF when analyzing the ACHI scale. It is interesting whether ERS might play a part in the reported Rasch model misfit. ACHI items were rated on the original dichotomous scale with categories labeled as "not true" and "true".[3] All items of the scale are positively keyed.

ERS Index

The questionnaire contained 50 items of heterogeneous content, which were used to compute an ERS index similar to study 1. ERS items were taken from two open access online databases: the "Zusammenstellung sozialwissenschaftlicher Items und Skalen" (ZIS) provided by the GESIS - Leibniz-Institute for the Social Sciences,[4] and the test archive of the "Leibniz-Zentrum für Psychologische Information und Dokumentation" (ZPID).[5] To ensure heterogeneous content and low inter-item correlations, we randomly sampled items from these databases, as has been suggested by (Weijters et al., 2008). On the 5th of January 2015, 210 scales were contained in the ZIS and 171 in the ZPID archive.[6] In a two step procedure, we sampled from the 381 scales, then randomly selected one item of each chosen scale. Subsequently, the content of each selected item was checked for suitability. An item was discarded if the same scale coincidentally was sampled twice, the item asked about clinical symptoms or was otherwise unsuitable for our population of students, the item wording was too long to be included in a standard questionnaire, the item required a special instruction, appeared highly out of context when presented on its own, or could not be used in combination with a Likert response scale. Moreover, items were excluded if their content was highly similar either to another ERS item or to any item from the ORD, SHY, and ACHI scales. In fact, 90 items had to be sampled until 50 suitable items were found. The wording of the final set of ERS items as well as their original source can be found in Appendix C.1.

A seven-point Likert scale with labeled extreme categories was used for all ERS items, irrespective of their original response format. Labels for the extreme categories were taken from the original item scale. For almost all ERS items, we retained the original item

[3] German labels: "stimmt nicht", "stimmt".

[4] The ZIS can be accessed under http://zis.gesis.org/

[5] The test archive of the ZPID can be accessed under https://www.zpid.de/index.php?wahl=products&uwahl=frei&uuwahl=userlog.

[6] Note that an unknown number of scales were included in both databases.

wording. Only in a small number of cases, slight modifications were made to ensure that the items were easier to understand when encountered out of context.

Some research suggests that in contrast to labeled extreme categories, a fully labeled response format leads to higher preferences for intermediate response categories (Weijters, Cabooter, & Schillewaert, 2010). ERS items are used to detect differences in response tendencies, which should be simplified by a high variance of extreme and midpoint responses. Therefore, labeling only the endpoints seemed most appropriate. It is not clear, whether the impact of ERS is affected by the number of response categories. When only endpoints are labeled, Weijters, Cabooter, and Schillewaert (2010) found that more response categories (up to seven) lead to a smaller rate of extreme responses, and a higher rate of midpoint responses. However, other researchers report that the preference for extreme categories is unrelated to the number of response categories, and that midpoint responses increase only for more than seven response categories (Kieruj & Moors, 2010, 2013). The main reason we opted for a seven point Likert scale is that Weijters and colleagues (Weijters et al., 2008; Weijters et al., 2010a; Weijters, Cabooter, & Schillewaert, 2010; Weijters et al., 2010b) always use this format for their randomly sampled response style items.

Self-Reported ERS

We used a dichotomous item to let participants judge their own response tendencies. We call this measure Self-Reported Extreme Response Style (SRERS). The item stated that some people often tend to choose one of the outer response options in questionnaires, while others often tend to choose response options close to the midpoint. Subjects were asked which of two response options provided a better description for them: "I often tend to choose one of the outer response options." or "I often tend to choose response options close to the midpoint."[7]

To our knowledge, this is the first study which includes a self-report measure of ERS. If it turns out that people have some conscious insight into their own response tendencies, this might have important consequences for developing procedural approaches to avoid ERS (see chapter 1.4). For example, it might be possible to give subjects a specific survey training, which reduces the detrimental effects of ERS on item response data quality.

Mixed Rasch ERS Based on Order

The eight item scale Order (ORD) is a facet of the Big Five factor conscientiousness in the NEO-PI-R (see chapter 2.2.1). It captures orderly behavior like cleaning, appearing

[7] German wording: "Manche Personen tendieren in Fragebögen häufig dazu, eine der äußeren Antwortmöglichkeiten anzukreuzen, während andere Personen häufig dazu tendieren, eine Antwortmöglichkeit nahe der Mitte anzukreuzen. Wie würden Sie sich spontan am ehesten einschätzen?"

"Ich tendiere häufig dazu, eine der äußeren Antwortmöglichkeiten anzukreuzen."

"Ich tendiere häufig dazu, eine Antwortmöglichkeit nahe der Mitte anzukreuzen."

meticulous and demanding, as well as tendencies to structure your daily life and plan things ahead. It is one of the nine NEO-PI-R facets for which Wetzel, Carstensen, and Böhnke (2013) showed superior fit of the constraint mixed Rasch model, suggesting that it is a good candidate to reveal effects of ERS. This notion was further supported by the PC tree analysis of the ORD scale in study 1 (for a reminder, see Figure B.11 in Appendix B.3). ORD items were rated on the original fully labeled five-point Likert scale of the German NEO-PI-R (Ostendorf & Angleitner, 2004).

The ORD scale was included in the questionnaire for two reasons. On the one hand, it was used to compute an ERS measure (MRERS). A constraint polytomous mixed Rasch model was fitted to the ORD scale, and participants were classified into the two resulting ERS classes. MRERS could than be used as an additional covariate in the analyses of the dichotomous scales. On the other hand, as extreme responding has been detected for the ORD scale in study 1, it serves as a control scale to assure that the ERS index from heterogeneous items as well as the self-reported ERS measure contain information about extreme responding. If we do not find an effect of the ERS measures in the ORD scale, we cannot assume that our method is capable of detecting ERS in the dichotomous SHY and ACHI scales.

3.2.2 Instructions and Questionnaire Design

At the beginning of the paper and pencil questionnaire, subjects were thanked for their participation. They were informed that their data would be treated anonymously and only be used for instructional or scientific purposes. This was followed by questions on demographic variables: Participants reported their age, sex, field of study, their final educational degree, and the number of completed semesters. Subsequent instructions informed them that the questionnaire consisted of 85 items on a variety of topics and with different response options. They were ensured that there are no right or wrong answers. For all questions, they should simply choose the single response most applicable to them. They were further instructed to answer all questions one by one, without skipping any question or browsing through the pages.

The order of all items (SHY, ACHI, ORD, and ERS items) was randomized once before the construction of the final questionnaire. For two items, the position was changed after the randomization, to avoid three consecutive items of the same content scale. The SRERS question was the last item in the questionnaire. To avoid that subjects who skimmed the questionnaire immediately noticed the special SRERS item, the same design was used for all items. Response categories were depicted as small squares without numbers. To our knowledge, there are no studies investigating the impact of numbered categories on ERS. Items were numbered and followed in direct succession without any blocks. To increase readability, consecutive items were separated by horizontal lines. The final questionnaire with 85 items consisted of five DIN-A4 pages. In our sample of psychology students, completion time was about 10-15 minutes.

3.2.3 Participants

Throughout the year 2015, a sample of 493 German speaking students was collected in psychology lectures and seminars at Ludwig-Maximilians-Universität (LMU) and Fresenius University of Applied Sciences in Munich. Subjects participated in exchange for course credit. All analyses are reported for the subsample of 429 subjects (76 men and 353 women) between the age of 17 and 52 ($M = 22.73$, $SD = 6.44$) who gave complete answers on ORD, SHY, ACHI, SRERS, answered enough ERS items for the computation of the ERS index, and reported their sex and age.

3.2.4 Statistical Analyses

All 50 ERS items were used for the computation of the ERS index. As in study 1, ERS items were first recoded to a response style scale. Extreme categories were coded with 1, the middle category was coded with 0, and all other categories were coded equally spaced in-between. All ERS items were rated on a seven point Likert scale. Thus, the recoded values on the response style scale were 1, 0.67, 0.33, 0, 0.33, 0.67, 1. For each subject, the ERS index was computed as the mean response on the recoded ERS items. The ERS index was still computed if a subject skipped up to two ERS items.[8]

For the SRERS measure, subjects were coded as one, if they expressed a tendency to choose extreme categories and as zero if they expressed a tendency to choose the midpoint category. To compute the MRERS measure, a constraint mixed Rasch model with two latent classes was estimated for the items of the ORD scale with the software `WINMIRA` (Von Davier, 2001). As suggested by (Wetzel, Carstensen, & Böhnke, 2013), the item location of each item was fixed between the latent classes. This should make sure that classes mainly captured diverging response patterns instead of different levels of the primary construct. The converged model revealed the typical ERS threshold pattern which is familiar from previous applications of the mixed Rasch model (Rost et al., 1997). To create the MRERS index, subjects were coded as one if the mixed Rasch model estimated a higher class probability for the group of extreme responders, and as zero if the estimated class probability was higher for the group of midpoint responders.

Similar to study 1, we computed the mean and standard deviation of the ERS index, the mean absolute correlation and the maximum absolute correlation between items within the index, as well as 95% confidence intervals for the correlations of the ERS index with sex and age. Furthermore, 95% confidence intervals for the correlations of the ERS index with the SRERS and MRERS variables were computed. All reported correlations are Pearson product-moment coefficients.

To assure that ERS measures captured information about extreme responding a PC tree was estimated for the ORD scale with the ERS index and SRERS as covariates. MRERS was not included in this analysis, as the constraint mixed PC model was based on the

[8] This decision was made in order to avoid unnecessary reduction of sample size. The distribution of missing values for the ERS items revealed, that 55 subjects skipped one or two ERS items while only 5 participants skipped more.

same scale. Trivially, MRERS contains parameter instability when used as a covariate, as the mixed Rasch model detects latent subgroups with maximum differences in response patterns (Rost, 1990). In a second step, separate PC trees of the ORD scale with only one of the three ERS measures as covariate were estimated to further investigate the effectiveness of the ERS measures.

Relative response frequencies, mean sum scores, and Cronbach's Alpha (Cronbach, 1951) were computed for the dichotomous SHY and ACHI scales. On the one hand, this served to detect potential irregularities in our sample by comparing our estimates with the normative sample. On the other hand, response frequencies might be related to which dichotomous items are affected by ERS.

For the main analyses, DIF Lasso models were estimated for the dichotomous SHY and ACHI scales with different sets of covariates. In the first model, the ERS index, sex, age, SRERS, and MRERS were included, while the second model only contained the three ERS measures. Additionally, Rasch trees were estimated for the SHY and ACHI scales with the same sets of covariates. In each DIF Lasso, the optimal value for the regularization parameter was determined by the BIC. The degrees of freedoms were once computed based on the method by Yuan and Lin (2006) and once based on the alternative L2-norm method. The group penalty was used in DIF Lasso models, and all covariates were standardized before entering the analysis. We graphically present estimates of the DIF parameters in the penalized solution. DIF items labeled by the less conservative L2-norm method are shown. The importance of the DIF inducing covariates decreases from left to right. The refit procedure is not used, as we are not interested in the actual values of the DIF parameters, as discussed in chapter 3.1.2.

For PC and Rasch trees, a significance level of 0.05 was used. For the minimum number of observations in a tree node, the default value equal to ten times the number of estimated parameters was used. The main reason for this decision was the generally small sample size. Similar to study 1, the results of the PC and Rasch trees analyses are presented graphically.

Following the idea that ERS might affect the discrimination of dichotomous items instead of their difficulty, we conducted a supplemental analysis of the SHY and ACHI scales which relied on the 2PL item response model: First, a median split was performed based on the ERS index. Then 2PL models of both scales were estimated for each of the two resulting ERS groups. ICCs were plotted for each 2PL model. If ERS had a considerable effect on item discrimination in dichotomous scales, differences should emerge in this graphical display. Due to the clearly exploratory nature of this analyses, we refrained from using a statistical test to examine if the discrimination parameters in the 2PL models differed between the two ERS groups.[9]

Except for the constraint mixed Rasch model of the ORD scale, all statistical analyses were conducted in `R` (R Core Team, 2016b). PC and Rasch trees as well as the corresponding tree plots were computed with the `psychotree` package (Zeileis et al., 2016). All

[9] For transparency, it should be noted that the idea that ERS might affect item discriminations arose after no effect of ERS was found in the Rasch tree and DIF Lasso analyses of the dichotomous scales.

remaining plots were created with the `ggplot2` package (Wickham & Chang, 2016). DIF Lasso models were estimated with the `DIFlasso` package (Schauberger, 2016). The `ltm` package (Rizopoulos, 2013) was used to estimate 2PL models. In addition to those already cited in study 1, the packages `foreign` (R Core Team, 2016a), `gtools` (Warnes, Bolker, & Lumley, 2015), and `gridExtra` (Auguie, 2016) were used.

3.3 Results

3.3.1 Effectiveness of ERS Measures

Table 3.1: Descriptive Statistics of the ERS Index

M	SD	MeanAC	MaxAC	r_{Sex} CI	r_{Age} CI	r_{SRERS} CI	r_{MRERS} CI
0.57	0.09	0.08	0.47	[-0.10, 0.09]	[-0.03, 0.16]	[0.22, 0.39]	[0.17, 0.35]

Note. 95% confidence intervals are presented for the correlations of the ERS index with sex, age, SRERS, and MRERS. Females are coded as one, males are coded as zero. Self-reported extreme responding is coded as one, self-reported mid responding as zero. Extreme responders classified by the mixed Rasch model are coded as one, midpoint responders as zero. Pairwise deletion was used to compute absolute correlations between items within the ERS index. MeanAC = Mean of absolute correlations between all items within the ERS index. MaxAC = Highest absolute correlation between items within the ERS index. SRERS = self-reported extreme response style. MRERS = extreme response style classified by a constraint mixed Rasch model of the Order scale.

Descriptive statistics for the primary ERS index from heterogeneous items are shown in Table 3.1. The ERS index was unrelated to sex and age. In the SRERS question, 32% of participants reported a tendency to choose extreme categories. In the constraint mixed Rasch model of the ORD scale (MRERS), 15% of participants were classified as extreme responders. Higher values on the ERS index were related to the extreme response groups of both the SRERS and MRERS with a medium effect size. While self-reported extreme responding in the SRERS was weakly correlated with the group of extreme responders in MRERS ($r_{SRERS,MRERS}$ 95% CI: $[0.05; 0.23]$), both measures were unrelated to sex and age.

The PC tree of the ORD facet with the ERS index, sex, age, and SRERS as covariates is presented in Figure 3.1. Parameter instability was only detected with respect to the ERS index, which was chosen to split the sample once at a value close to the mean. In the two leaves, the familiar pattern emerged: In the subsample with higher ERS values, thresholds were more narrow and regions of highest probability for the most extreme response categories were larger. Although self-reported ERS was not selected as splitting variable in addition to the ERS index, SRERS was selected to split the sample if it was the only available covariate. This is shown in Figure 3.2. Trivially, a strong pattern resulted if MRERS was used as a single covariate. The corresponding PC tree can be found in Appendix C.3.

3.3.2 Analyses of the Shyness Scale

Relative response frequencies of the SHY items are presented in Appendix C.2. Response rates were quite low for most items with a median relative response frequency of 0.31. Mean sum scores were 4.54 for male and 5.00 for female subjects. Compared to the German normative sample with mean sum scores of 4.89 and 5.18, our participants described

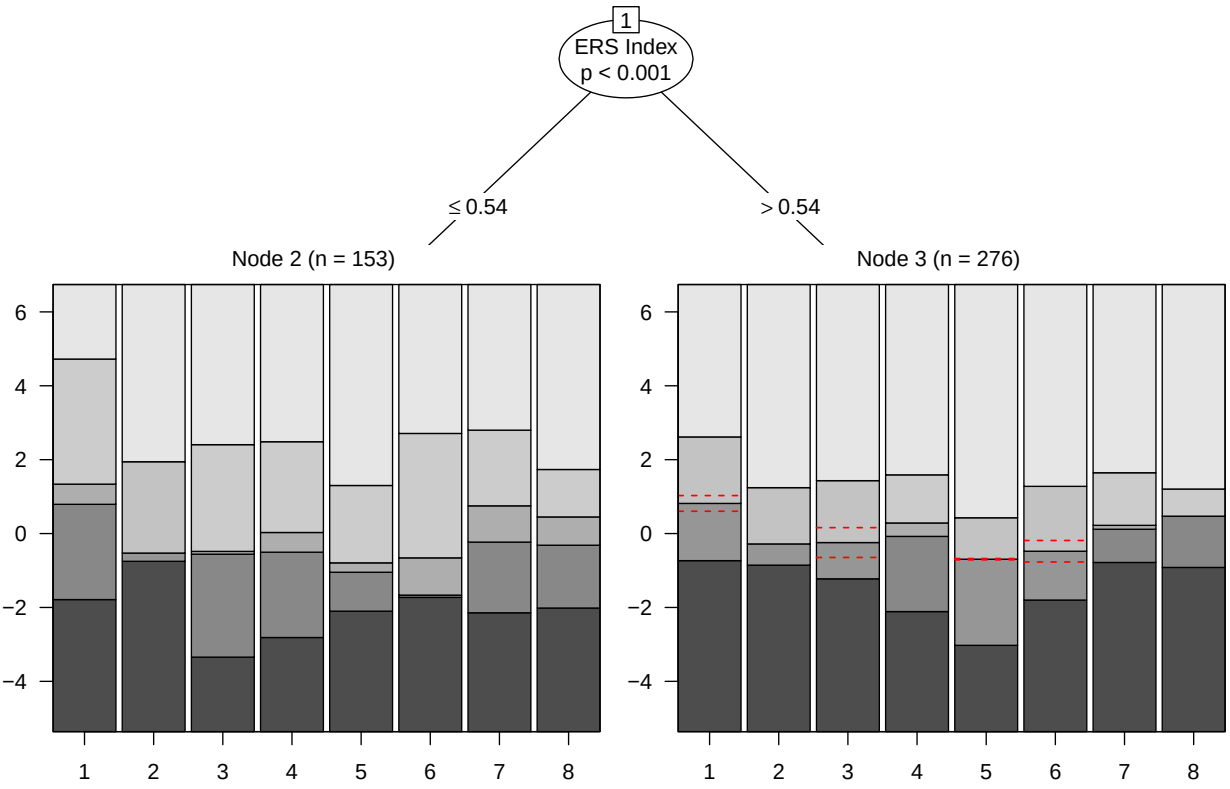

Figure 3.1: PC tree of the Order (ORD) scale with the ERS index from heterogeneous items, sex, age, and SRERS as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. SRERS = self-reported extreme response style.

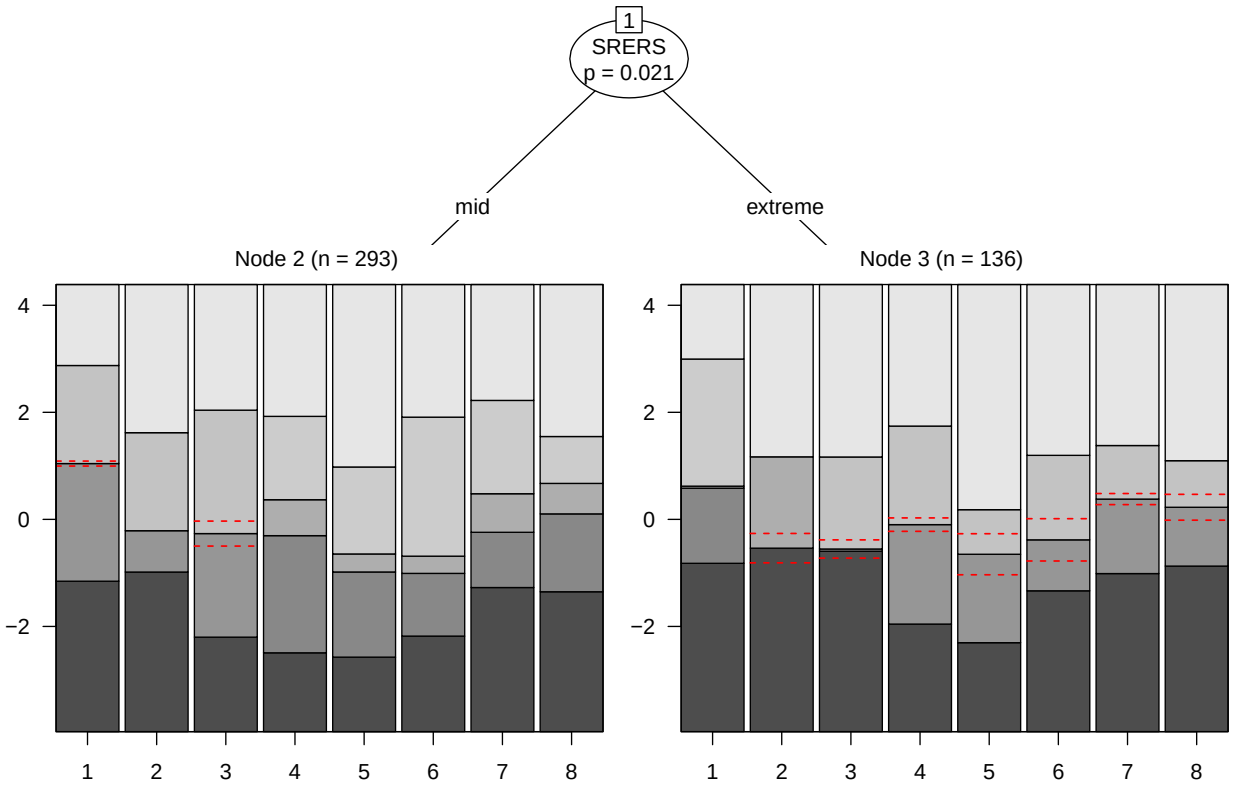

Figure 3.2: PC tree of the Order (ORD) scale with SRERS as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. SRERS = self-reported extreme response style.

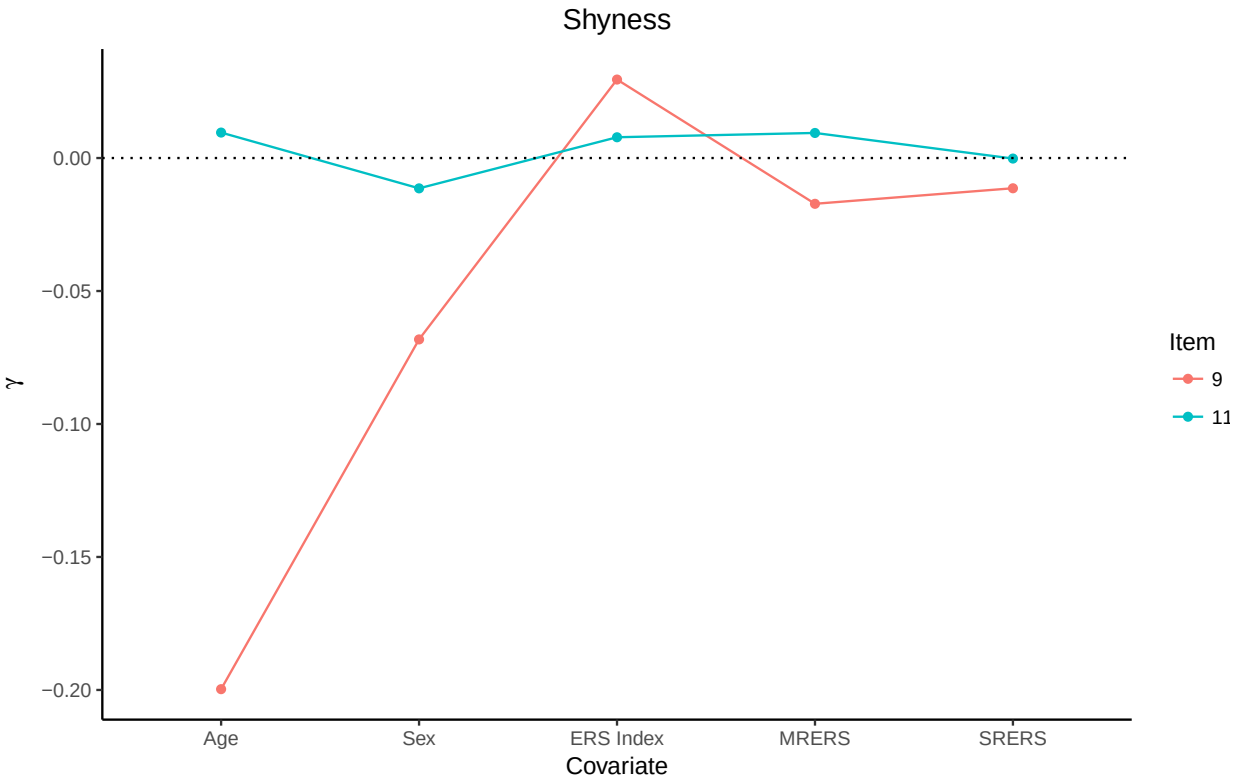

Figure 3.3: Summary of the DIF Lasso model for the Shyness (SHY) scale. The ERS index, sex, age, SRERS, and MRERS were included as covariates. Females are coded as one, males are coded as zero. Self-reported extreme responding is coded as one, self-reported mid responding as zero. Extreme responders classified by the mixed Rasch model are coded as one, midpoint responders as zero. Degrees of freedom for the BIC were computed based on the L2-norm method. The fitting procedure excluded 27 subjects who endorsed none or all SHY items. SRERS = self-reported extreme response style. MRERS = extreme response style classified by a constraint mixed Rasch model of the Order scale.

themselves as slightly less shy. Cronbach's Alpha was in line with the previous literature (95% CI: $[0.80; 0.85]$).

Results of the DIF Lasso model for the SHY scale with the ERS index, sex, age, SRERS, and MRERS as covariates are presented in Figure 3.3. As illustrated, two DIF items were detected by the BIC when the degrees of freedom were computed based on the L2-norm method. However, when the method by Yuan and Lin (2006) was used, zero DIF items were detected. Of the two potential DIF items, the covariates seemed to have a considerable effect only on item nine: "At school, I found it very difficult to speak in front of the class."[10] This item was easier for older, female subjects. With similar levels of shyness, speaking in front of the class was generally harder for these participants. All three response style measures had negligible effects on the difficulties of both potential DIF

[10] German wording: "Ich fand es in der Schule sehr schwer, vor der Klasse zu sprechen."

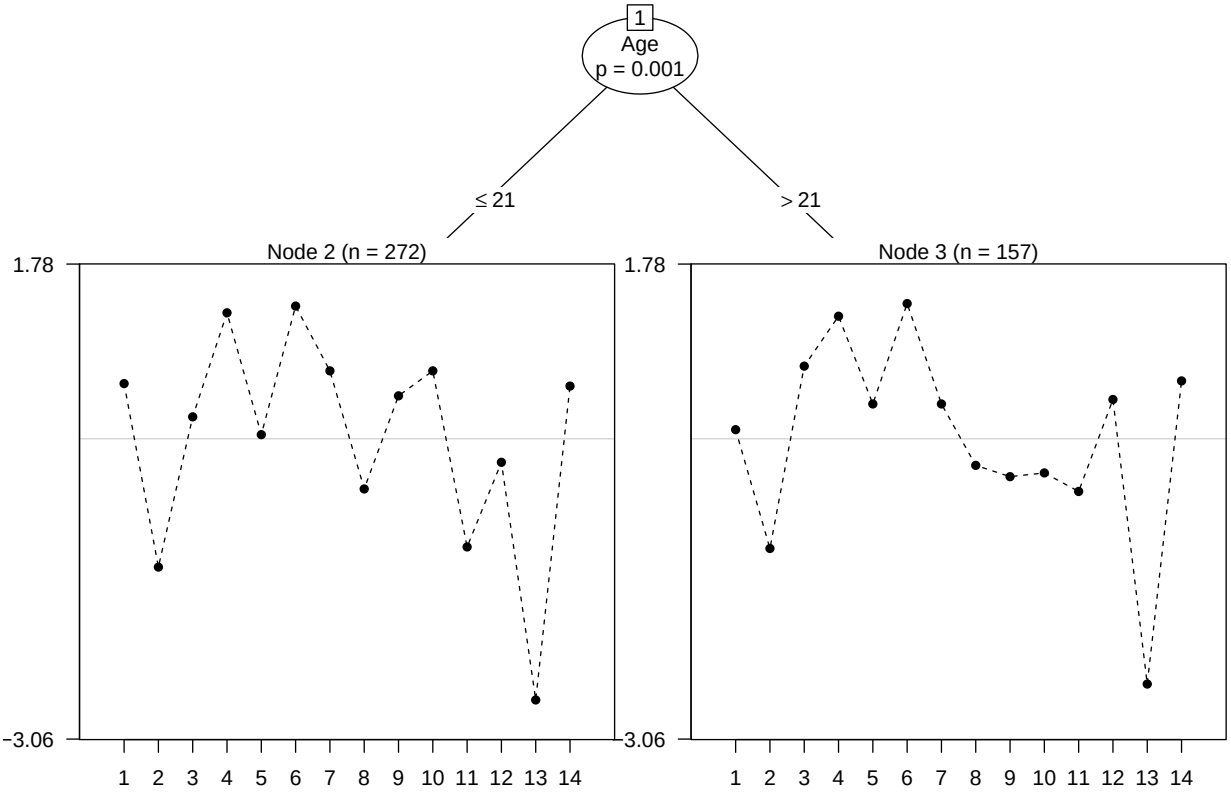

Figure 3.4: Rasch tree of the Shyness (SHY) scale with the ERS index from heterogeneous items, sex, age, SRERS, and MRERS as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. SRERS = self-reported extreme response style. MRERS = extreme response style classified by a constraint mixed Rasch model of the Order scale.

items. Moreover, when only the three response style variables were included in the DIF lasso, zero DIF items were detected, irrespective of which method was used to compute the degrees of freedom. Lasso paths for this model can be found in Appendix C.4. The optimal value for the regularization parameter was so high, that a noteworthy effect for any of the three response style measures seems very unlikely.

Figure 3.4 shows the Rasch tree of the SHY scale with the ERS index, sex, age, SRERS, and MRERS as covariates. Age was selected to split the sample once at an age of 21. None of the three response style measures was selected as splitting variable. When the ERS measures were the only covariates included in the Rasch tree, no DIF was detected and a single Rasch model was estimated for the whole sample. When examining the difficulty estimates of items nine and eleven, the same age pattern emerged as in the DIF Lasso model. While item nine was easier in the older group, item eleven was more difficult. Item ten was also easier in the older group, with an even bigger effect than for item nine.

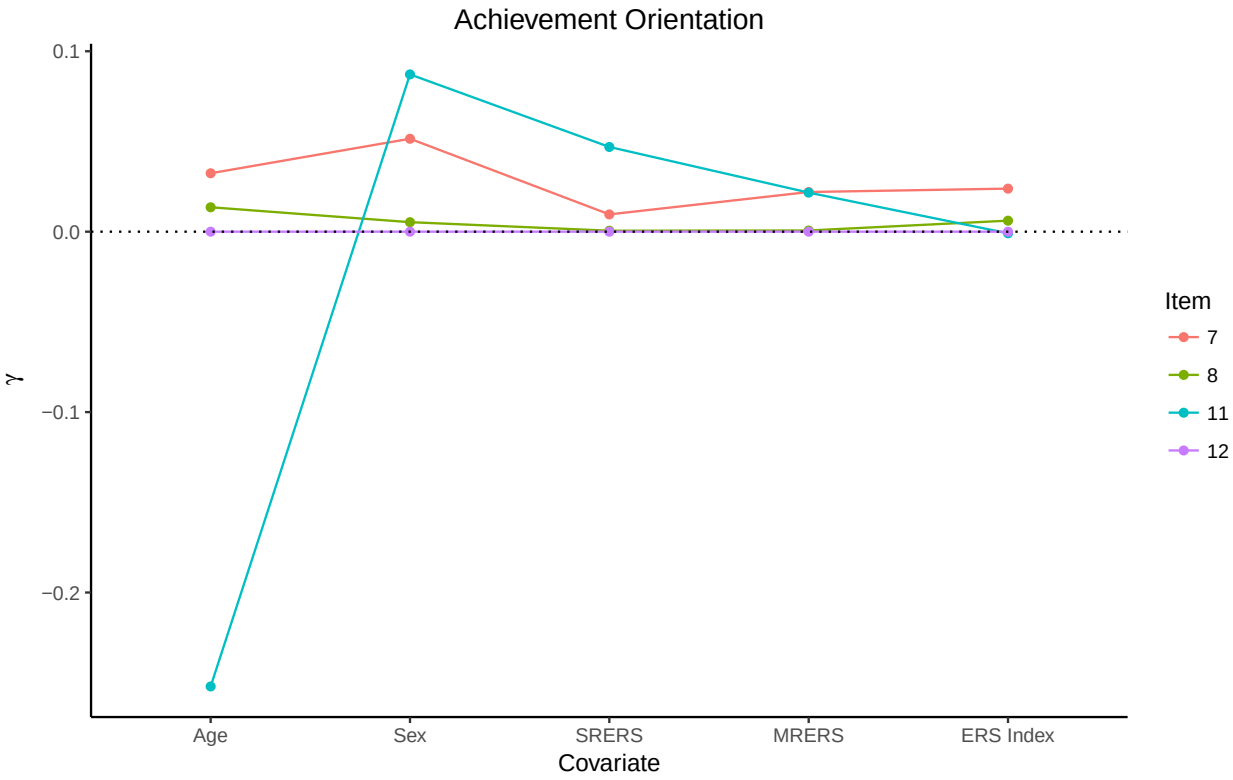

Figure 3.5: Summary of the DIF Lasso model for the Achievement Orientation (ACHI) scale. The ERS index, sex, age, SRERS, and MRERS were included as covariates. Females are coded as one, males are coded as zero. Self-reported extreme responding is coded as one, self-reported mid responding as zero. Extreme responders classified by the mixed Rasch model are coded as one, midpoint responders as zero. Degrees of freedom for the BIC were computed based on the L2-norm method. The fitting procedure excluded 7 subjects who endorsed none or all ACHI items. SRERS = self-reported extreme response style. MRERS = extreme response style classified by a constraint mixed Rasch model of the Order scale.

3.3.3 Analyses of the Achievement Orientation Scale

Relative response frequencies of the ACHI items are presented in Appendix C.2. Response rates were higher compared to the SHY scale, with a median relative response frequency of 0.62. Mean sum scores were 7.67 for male and 7.31 for female subjects. Compared to the German normative sample with mean sum scores of 7.63 and 6.53 our female participants described themselves as more achievement oriented.[11] Cronbach's Alpha was lower than in the normative sample (95% CI: [0.61; 0.70]).

Results of the DIF Lasso model for the ACHI scale with the ERS index, sex, age,

[11] In contrast to the MMPI-2, normative data for different age groups are available for the FPI-R. The mean sum score of females between the age of 20 and 29 in the normative sample is 7.04. This is much closer to our sample in which 55% of females were of that age.

SRERS, and MRERS as covariates are presented in Figure 3.5. As illustrated, four DIF items were detected when the degrees of freedom for the BIC were computed based on the L2-norm method. However, when the method by Yuan and Lin (2006) was used, only item 11 was selected. Of the four potential DIF items, the covariates seemed to have a considerable effect only on item 11: "I prefer the action to the making of plans."[12] This item was easier for older, male subjects. With similar levels of achievement orientation, taking action was generally more preferred by these participants. Similar to the SHY scale, all three response style measures had negligible effects on the difficulties of the four potential DIF items. When only the three response style variables were included in the DIF Lasso, zero DIF items were detected, irrespective of which method was used to compute the degrees of freedom. Lasso paths for this model can be found in Appendix C.4. While a noteworthy effect of any of the three response style measures seems unlikely, a special case might be item 12: "At my work, I am usually faster than others."[13] For decreasing values of the regularization parameter close to the optimal one, this item gets easier for participants that reported to frequently choose extreme responses in the SRERS. A weaker trend can also be observed for the ERS index.

Figure 3.6 shows the Rasch tree of the ACHI scale with the ERS index, sex, age, SRERS, and MRERS as covariates. It has three leaves: males up to an age of 25, females up to the same age and a combined sex group of subjects older than 25. Similar to the SHY scale, none of the three response style measures was selected as splitting variable. This did not change when the response style measures were the only covariates included in the Rasch tree. In that case, no DIF was detected and a single Rasch model was estimated for the whole sample. When examining the difficulty estimates of the four DIF items selected by the Lasso, the effects of age and sex were highly comparable. For example, item 11 had the highest difficulty estimate in the group of young females, followed by young males and older subjects. Furthermore, the ranking of the four items in terms of the amount of DIF was similar in the Rasch tree. The highest differences between leaves emerged for item 11 and the smallest differences were found for item 12.

3.3.4 Supplemental 2PL Analysis

Figure 3.7 shows the ICCs for the 2PL models of both scales which were fitted to the subsamples above and below the median ERS value. With few exceptions, ICCs were highly similar between the two subsamples. No pattern emerged with respect to the slope of the ICCs for any of the two scales.

[12] German wording: "Ich ziehe das Handeln dem Pläneschmieden vor."
[13] German wording: "Bei meiner Arbeit bin ich meist schneller als andere."

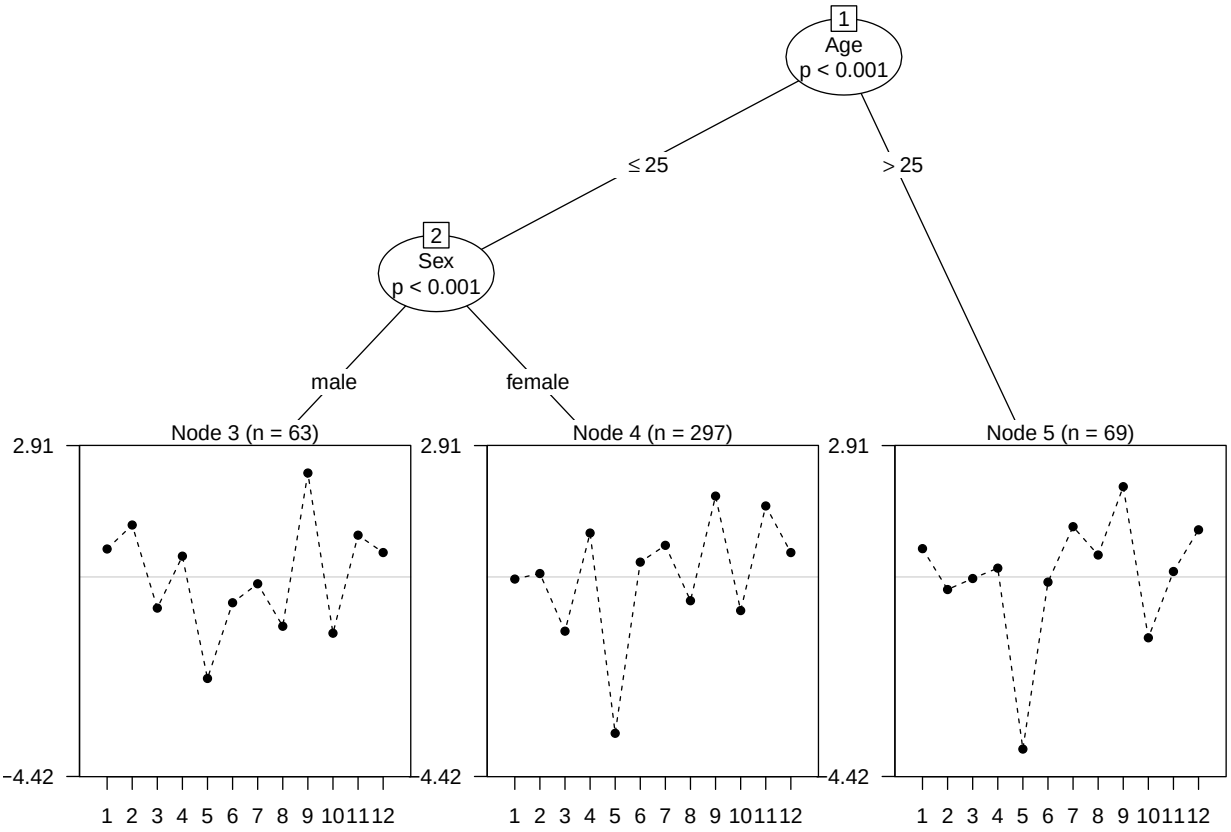

Figure 3.6: Rasch tree of the Achievement Orientation (ACHI) scale with the ERS index from heterogeneous items, sex, age, SRERS, and MRERS as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. SRERS = self-reported extreme response style. MRERS = extreme response style classified by a constraint mixed Rasch model of the Order scale.

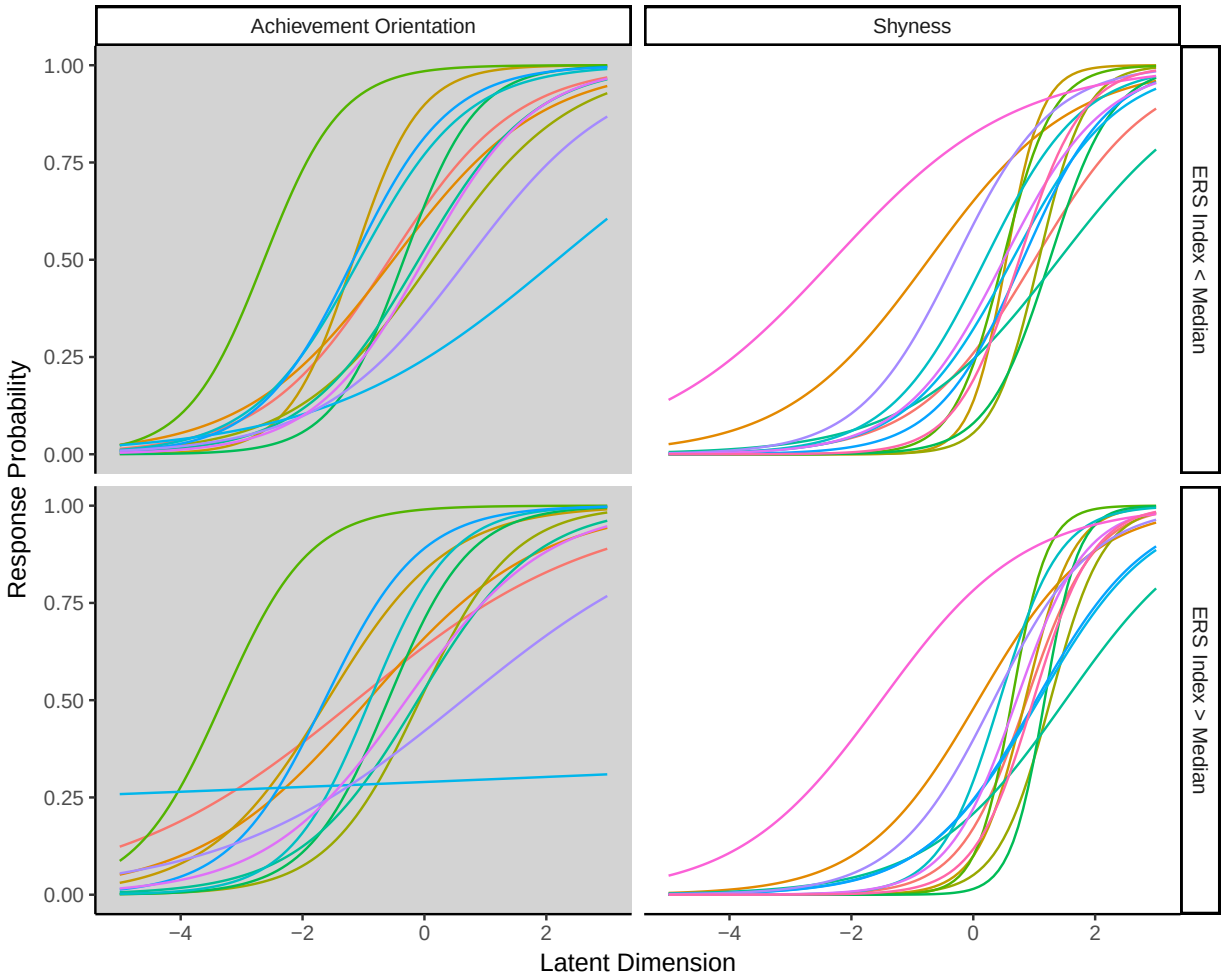

Figure 3.7: Item Characteristic Curves for 2PL Models of Achievement Orientation (ACHI) and Shyness (SHY) scales. Separate 2PL Models were estimated for participants with an ERS index below and above the median. Graphs of the same scale are shaded with the same color. Within each scale, similar items are marked by the same color in both ERS groups.

3.4 Discussion

3.4.1 Summary of Results

The inter item correlations between the randomly sampled ERS items turned out to be small. The resulting ERS index was unrelated to sex and age but showed medium correlations with SRERS and MRERS. In the PC tree analyses of the ORD scale, all three response style measures contained parameter instability. The resulting two leaves showed threshold patterns which suggest that extreme responding was detected by all three measures. In contrast, no measure of ERS had an effect on response patterns in the dichotomous SHY and ACHI scales. For items selected by the Lasso, DIF was primarily caused by the demographic variables sex and age. When only response style measures were included in the Lasso, no DIF items were detected. Similarly, no response style measure was selected as split variable in the Rasch tree analyses. Rasch trees also detected DIF that could be contributed to sex and age. The effects were comparable to the DIF items detected by the Lasso. Exploratory 2PL models of the dichotomous scales which were estimated for subsamples resulting from a median split of the ERS index, did not suggest any effect of ERS on item discriminations.

3.4.2 Comparison of ERS Measures

The applied sampling procedure resulted in a set of uncorrelated ERS items. The mean absolute inter-item correlation was comparable to previous research (e.g. Plieninger & Meiser, 2014). Not surprisingly, single correlations between ERS items were a bit higher than in study 1, where we used an algorithm to find the most uncorrelated items based on already collected data. On average, ERS values were slightly higher compared to study 1. This might be related to the concrete set of ERS items or our sample of psychology students. Another possible explanation could be the mixed presentation of dichotomous and polytomous items in our questionnaire. Without a midpoint category, the dichotomous items always required a clear decision. This might have had the effect that subjects also chose more extreme and less midpoint responses on the ERS items compared to a questionnaire which completely lacked dichotomous items.

PC tree analyses of the ORD scale suggest that all three ERS measures carried information about extreme responding. They were all selected as splitting variable, if used as the only covariate. As expected, a strong threshold pattern reflecting extreme responding was found for the MRERS measure, which was computed based on the same ORD scale. In the combined analyses with the ERS index, sex, age, and SRERS as covariates, only the ERS index was selected as splitting variable. In line with study 1, parameter instability contained by the ERS index was larger than for the demographic variables. In contrast to study 1, the ERS index was only selected to split the sample once, at least when used together with the other covariates. Sex and age were not selected as splitting variables in the ORD analyses at all. This is probably due to the much smaller sample size compared to study 1, which used the normative sample with several thousand participants.

Descriptive analyses showed that subjects with high ERS values were more likely to report tendencies of extreme responding in the SRERS question, and to be classified as an extreme responder by the constraint mixed Rasch model of the ORD scale. Still, associations were only moderate, suggesting that the different ERS variables are not exchangeable in practice, and might measure slightly different aspects of extreme responding. Especially the relationship between SRERS and MRERS turned out to be rather weak.

When SRERS was used as the only covariate in the PC tree of the ORD scale, thresholds revealed the familiar ERS pattern although it seemed to be weaker than for the other two ERS measures. The finding implies that people have at least some conscious knowledge about their own extreme response tendencies, which has not been investigated in earlier studies of ERS. This knowledge might be used to find more procedural remedies of ERS. For example, respondents could be assisted in producing more valid questionnaire data, by helping them recognize their own response tendencies and guiding their responses to more appropriate categories. These ideas had a big influence on the construction of the third study which will be presented later.

3.4.3 No Effect of ERS in the Dichotomous Scales

None of the three ERS measures revealed any effect of ERS on the item responses of the two analyzed dichotomous scales. No DIF items were detected when ERS measures were the only covariates in the DIF Lasso and none of the ERS measures was selected as splitting variable in the Rasch trees. When demographic variables were included, both approaches detected DIF based on sex and age. When sex and age were included in the model, more DIF items were detected by the DIF Lasso when using the less conservative L2-norm method to compute the degrees of freedom for the BIC. However, the method to compute the degrees of freedom did not matter if the ERS measures were the only covariates. DIF results were highly similar between the DIF Lasso models and the Rasch trees. This was also true with respect to DIF based on sex and age. DIF items detected by the Lasso also showed the biggest differences in threshold parameters in the corresponding Rasch tree. Also, the effect of the covariate was always in the same direction. Small differences can possibly be attributed to the relatively small sample size. As separate models are fitted within each leaf of a Rasch tree, the precision of parameter estimates for small leaves is very low. In these cases, small differences in threshold parameters can be expected even in the absence of a true DIF effect. On first glance, the Rasch tree of the ACHI scale might suggest an interaction effect for sex and age. This could not be modeled by the DIF Lasso in which interaction terms were not included. However, the high age group was so small that another split based on sex was not even attempted by the algorithm. Similar reasons might also explain why sex was not used to split the resulting age groups in the Rasch tree of the SHY scale. As explained in the introduction, the DIF Lasso results might be more trustworthy if the focus is not on which variables contain DIF but which items are most affected. The main DIF item based on the DIF Lasso of the SHY scale suggested that for equally shy subjects, speaking in front of the class back in school was generally harder for older women. It is the only item in the SHY scale about a performance situation. DIF

Lasso results of the ACHI scale indicated that for equally achievement oriented subjects, older men more preferred taking action to making plans. While several items of the ACHI scale focus on taking fast actions, the DIF item is the only one mentioning the planning aspect. We refrain from further post hoc interpretations of these findings. Nevertheless, they illustrate that the applied methods were generally capable of detecting DIF in in dichotomous items.

In the supplemental analyses with the 2PL model, we investigated the possibility that ERS might have an effect on item discrimination instead of item difficulty. More specifically, the theoretical understanding of ERS might suggest that item discrimination should be higher for people with high extreme response tendencies. However, the graphical investigation of ICCs did not indicate an effect in either direction. Note that we did not use a rigorous statistical test and that the power to detect any effects in the more complex 2PL model was further reduced by the applied median split. Strobl et al. (2013) mention in their article on Rasch trees that they would like to extend their methodology to the 2PL model, which should also be possible for the DIF Lasso approach. If any of these methods are developed in the future, investigating the effect of ERS on item discrimination based on a bigger sample should be worthwhile both for dichotomous and ordinal response formats.

3.4.4 Advantages and Disadvantages of Dichotomous Item Formats

In this study, we presented the first empirical evidence in favor of the widely held assumption that dichotomous items are unaffected by ERS. As suggested by a number of researchers (Bühner, 2011; Cabooter et al., 2016; Wetzel, Carstensen, & Böhnke, 2013), adopting dichotomous response formats for psychological measurements might be a reasonable strategy to deal with ERS.

In light of the mixed opinions in the community about the impact of ERS on psychological measurements with ordinal response formats, avoiding ERS per design seems like a good advice. While many authors claim devastating effects of ERS on the precision of psychological measurement (Baumgartner & Steenkamp, 2001; Mõttus et al., 2012), two recent simulation studies argue that ERS can be savely ignored as long as extreme responding and the primary trait in question are not strongly correlated (Plieninger, 2016; Wetzel, Böhnke, & Rose, 2016). In chapter 1.4, we argued that both simulation studies have a series of limitations. Until more convincing empirical research suggests that ERS has a low impact in practical applications, preventing ERS seems to be preferable.

In addition to the illustrated insensitivity against extreme responding, further advantages of dichotomous items are discussed by proponents of the response format: It has been found that completing a questionnaire with dichotomous items takes less time compared to a version with ordinal response format (Dolnicar, Grün, & Leisch, 2011). The authors argue that efficiency is crucial in marketing research, where completion times are negatively related to survey costs and response rates. Recently, shortened personality measures are becoming more popular in psychological research (Rammstedt & John, 2007). Two

reasons are an increasing number of large scale panel studies and a higher focus on more complex research questions which require to assess many psychological constructs at the same time (Ziegler, Kemper, & Kruyen, 2014). Dichotomous items might be a useful strategy to lower completion times in these settings, while still retaining a high number of items in a shortened scale. Thereby, constructs could be assessed from different angles, which increases the content validity of short scales without sacrificing too much efficiency.

Dichotomous items might also have their disadvantages: The prevalent reason why ordinal response formats are favored by most researchers is the assumption that more fine grained psychological measurements yield higher reliability (Preston & Colman, 2000; Weng, 2004). In contrast, Cronbach (1950) already noted that higher reliability estimates are precisely what should be expected in light of increased response style variance. Note that the reliability of the sum score of a psychological scale is only a useful quantity if the essentially tau-equivalent model of CTT holds. Even for the tau-congeneric model, the sum score is a deficient measure for the latent variable and would not be used by anyone who takes the measurement model seriously. If factor scores are used, the reliability of the sum score is not of any interest at all. The pattern found in the PC trees of study 1 suggests that ERS is highly prevalent in item response data with ordinal response format. In this case, no unidimensional model from CTT is appropriate and more complex IRT models should be investigated (see the discussion in chapter 2.4.4). As the ERS pattern is too complex to be described by a simple pattern of correlated errors in classical CTT models, the fact that reliability could still be estimated under these circumstances (Gu, Little, & Kingston, 2013; Raykov, 2001) is irrelevant. Of course, the dichotomous response format only provide a real advantage with regard to the above argument if measurement models hold for the resulting scales. Although we did not encounter any signs of ERS in the dichotomous SHY and ACHI scales, the Rasch model had to be rejected in both cases, as Rasch trees and the DIF Lasso detected violations of the model based on age and gender. As long as psychological test construction fails to provide dichotomous scales that show acceptable fit under meaningful measurement models, arguments in favor of dichotomous items that are mainly based on psychometric properties remain on weak ground. At least, our results suggest that when developing a psychological measure based on dichotomous items, ERS is one less thing to worry about.

A more convincing point in favor of ordinal response formats could be made, if studies would suggest higher criterion validity compared to dichotomous items. However, most studies on this topic report small increases in validity at best (Preston & Colman, 2000; Finn, Ben-Porath, & Tellegen, 2015). Unfortunately, all available studies used psychological scales as criterion measures instead of genuine external variables that are recorded without self-report items. We briefly mention the study by Finn et al. (2015), as it used the MMPI-2 with the original dichotomous and a new four-category format. Similar to the majority of the remaining MMPI-2 scales, correlations between the SHY sum score and a series of questionnaires measuring personality or clinical symptoms were indistinguishable for both versions. In contrast to other studies, some criterion measures consisted of dichotomous items. If both the scale of interest and the criterion measure use items with ordinal response format, response styles and other forms of method bias can trivially

increase the correlation between the sum scores (MacKenzie & Podsakoff, 2012; Podsakoff et al., 2011). This makes observed increases in validity hard to interpret.

From a different angle, the study by Plieninger and Meiser (2014) also suggests that polytomous items do not necessarily contain more information as dichotomous ones. Two datasets were analyzed with a sequential multiprocess model (Böckenholt, 2012), in which item responses are determined by a MRS process (choosing the midpoint category or not), followed by a content trait process (agreement or disagreement) and an ERS process (choosing the extreme reponse option of not).[14] In both datasets, the content process was correlated with relevant external criteria to the same extend as the ordinary sum score or the trait estimates from a unidimensional IRT model. A similar result was found in a later study (Böckenholt & Meiser, 2017), in which two content processes in a two dimensional multiprocess model were correlated to the same extend as the corresponding trait variables of a two dimensional polytomous mixed Rasch model. Note that based on the multiprocess model, the content process should only contain the binary information if the respondent agrees with the item or not, and is therefore conceptually comparable to a dichotomous item response on the item. These findings provide empirical evidence for some early work by Peabody (1962). Through analytical decompositions of the sum score variance of a series of psychological scales, he anticipated that ordinal item responses might primarily contain information about direction (agreement vs. disagreement). Moreover, he recognized that the small part of variance which could be attributed to the intensity of item responses showed clear indications of ERS.

Another common argument in favor of ordinal items is that the dichotomous response format is less preferred by respondents. In the study by Preston and Colman (2000), the dichotomous format was rated by participants as less adequate to express their feelings. At the same time, it was considered the least easy to use and the quickest to use. This is in contrast to another study (Dolnicar et al., 2011), where the dichotomous format was perceived as less complex than the ordinal one. In our study, missing values were 1.52 times more likely for the dichotomous SHY and ACHI items than for the ordinal ORD and ERS items (95% CI of odds ratio from Fisher's exact test: $[1.20; 1.94]$).[15] It is possible that our participants also felt that they could not give adequate responses to some dichotomous items. As a consequence, they might have refused to respond to dichotomous items more often compared to items with ordinal response format.[16] Finn et al. (2015) mention an unpublished doctoral thesis by Cox (2011), in which the MMPI-2 has been compared between the traditional dichotomous and a four-category response format. In that study, the dichotomous version was completed faster and also considered easier to complete. However, respondents indicated that they could describe themselves better with the four-category format, which is in line with the findings of Preston and Colman (2000).

Even if ERS seems to be absent in dichotomous item response data, it has been noted

[14] The MRS process is missing if the item does not contain a midpoint category. The ERS process is repeated if the item has more than five response categories.

[15] As missing values were rare, note that the relative risk and the odds ratio are virtually the same.

[16] Of course, it is also possible that missing values were related to the content of the SHY and ACHI scale.

that other response styles like ARS might still be important (Wetzel, Lüdtke, et al., 2016; Rorer, 1965). When Cronbach (1950) demanded researchers not to use item forms infested by response styles, he also discouraged the use of dichotomous agreement/disagreement response scales. It has been found that ERS might have a bigger impact than ARS in items with ordinal response formats (Wetzel & Carstensen, 2015; Kieruj & Moors, 2013). However, this can be hardly generalized to dichotomous items. Recent empirical studies about ARS in binary items are scarce, although model-based approaches to investigate this issue are available (Ferrando & Condon, 2006). By changing the coding of the items, response style indices from heterogeneous items can be used to investigate the occurrence of ARS. Applying an ARS index in addition to the ERS measures in this study could have clarified whether ARS was a problem in our data. Before issuing a general recommendation to rely on dichotomous item formats, the here presented work should be extended to the investigation of ARS. In the case that further studies find a high prevalence of ARS in dichotomous items, this response style has to be addressed in scale construction, too. Multiple studies suggest that ARS can be easily avoided by balancing the number of positively and negatively keyed items in a particular scale (Plieninger, 2016). Although it has been shown that reversely keyed items often pose serious measurement problems, it has been suggested that detrimental effects on psychometric properties should be rather small (Weijters & Baumgartner, 2012) when avoiding negations in the item wording. With careful wording, negatively keyed items might be worth the costs if their use leads to psychological scales which are mostly unaffected by the two most important response styles. We briefly note that multidimensional forced-choice response formats could be an alternate strategy to avoid response styles. Methodological advances in analyzing forced-choice item response data (Brown & Maydeu-Olivares, 2011) have reignited the interest in this item type. While forced-choice formats had been endorsed by Cronbach (1950) in the context of response styles many years ago, they have long been neglected due to methodological problems (Brown & Maydeu-Olivares, 2013). However, a detailed discussion of Thurstonian IRT is beyond the scope of this thesis. General implications for scale construction and psychometric practice which can be derived from our work on response styles will be further discussed in chapter 5.2.

Chapter 4

Study 3: Compensating Extreme Response Style with Item Instructions

4.1 Introduction

4.1.1 Instructions for an Experimental Manipulation of Extreme Response Style

Study 2 suggests that using binary item response formats might be an effective procedure to avoid ERS. Yet, we also discussed some disadvantages of dichotomous items that discourage researchers to opt for this response format. Procedural approaches to avoid ERS in items with ordinal response format would be more appealing to the majority of applied researchers. Cronbach (1950) has brought up the idea to reduce the impact of response styles by altering the instructions which precede most psychological measurements. While he gives an example of how to reduce acquiescence by informing test takers about the ratio of true questions in an ability test, it is not obvious how a similar strategy might be applicable to ERS. To design instructions which could reduce the impact of ERS, a theoretical framework is necessary that tries to explain how subjects respond to questionnaire items. This is important, as some critics stress that conceptual ideas about how item responses are chosen in questionnaires are lacking in most approaches to study response styles (Zettler et al., 2015).

According to a popular cognitive process model for answering questionnaire items (Tourangeau et al., 2000), respondents of self-report questions proceed through four stages. In the comprehension stage, subjects encounter an item and try to interpret its content. Then, relevant beliefs to answer the question are retrieved from memory in the retrieval stage. In the judgment stage, beliefs are combined into an internal judgment of how the subject ranks on the dimension described by the item. Finally, in the response stage, subjects map their internal judgment onto the response categories provided by the item and select their answer. Several authors have noted that response tendencies can be discussed within this framework (Podsakoff et al., 2011; Weijters, Cabooter, & Schillewaert, 2010;

Krosnick, 1991).

Considering how ERS might affect item responses during the answering process, the response stage might play the most important role. Per definition, ERS describes differences in the use of the response format. Thus, two respondents which arrive at the same internal judgment might not select the same response option due to different levels of ERS. This idea is corroborated by Weijters, Cabooter, and Schillewaert (2010) who found an effect of the response format on the impact of ERS and other response styles. However, patterns in item responses that can be detected by analyses of ERS as demonstrated in study 1 and 2, might also depend on earlier stages of the response process. For example, ERS could assert some influence in the judgment phase. Subjects with different levels of ERS might come to different internal judgments, based on an identical memory recall. Direct testing of this hypothesis is only possible through extensive cognitive interviews which aim to separate the internal judgment from the item responses. This has not been done thus far. However, the body of literature on the relationship between ERS and personality characteristics can be related to decision making in general (Austin et al., 2006; Naemi et al., 2009; Zettler et al., 2015). This suggests that ERS might already operate in the judgment stage. Cronbach (1950) noted that response styles seem to be more influential for ambiguous items. If subjects are uncertain how to respond to an item based on its content, response styles should have a big impact. If responses are fully determined by the item content due to clear wording, subjects' individual response tendencies cannot manifest themselves. Thus, ambiguous item wording might have a moderating effect on the impact of ERS on item responses, which should take place in the judgment stage.

The last argument might suggest that when designing an instruction to compensate ERS, the focus should be on item responses for which a subject perceives a high level of uncertainty. For those responses, ERS should have the largest impact and subjects should be most inclined to adjust their answers according to some external instruction. In contrast, the impact of response tendencies should be lower, if a subject is highly certain about an answer. Even if a small effect of ERS is still present, subjects are probably unwilling to reorient their decision in those cases. In the following study, we propose an ERS instruction which urges subjects to adjust their item responses if they are uncertain about an item. An important question is how respondents should adjust their responses. By using a self-report measure of ERS in study 2, we showed that respondents have some conscious knowledge about their own response tendencies. Thus, subjects could be asked whether they tend to choose mid or extreme categories in a first step. When confronted with questionnaire items, respondents could then be instructed to adjust their responses to compensate their individual response tendencies. Although self-reported ERS seems to be a valid indicator of extreme response tendencies, studies 1 and 2 also suggest that more information is contained in an objective measure like the ERS index from heterogeneous items. Thus, relying on subjects' self-reported ERS is probably not the most effective way to compensate ERS. Moreover, subjects might be overwhelmed if they are not given a concrete procedure how to adjust their responses. An effective instruction should be concrete to give sufficient guidance but flexible enough to ensure that subjects can use knowledge about their own response tendencies to control whether an item response should

be adjusted. In this study, we use an ERS instruction with the following rationale: If subjects cannot decide between two neighboring response categories, they should rather choose either the more extreme/the less extreme of the two options. For this instruction, it is necessary to specify if subjects should give more or less extreme responses. In a practical application, it would be necessary to measure each participant's level of ERS, either by self-report or with an index from heterogeneous items, and select either the extreme or mid-responding instruction, based on a binary ERS classification for each participant. If using self-reported ERS similar to study 2, such a design could be easily achieved in an online survey which can be programmed to present a certain instruction dependent on an earlier response on the SRERS item. The procedure would be more difficult when using an ERS index, because a binary cutoff has to be specified beforehand. To compare the effectiveness of both ERS measures, we use a within subject design with an additional control group. All participants first fill out two psychological scales under standard instructions together with a series of heterogeneous items which can be used to compute an ERS index. Subjects in the experimental group answer the scales again under both the extreme, and the mid-responding instruction. In contrast, the control group fills out the same scales two more times under neutral instructions to rule out order effects or other confounding factor of the within subject design. Finally, all subjects answer a self-report question of ERS. During data analysis, subjects in the experimental group can be classified into extreme and mid responders based on both the ERS index and self-reported ERS. Different classifications can then be used to create artificial datasets. Item responses are used from those trials in which the ERS instruction leads to a compensation of individual ERS. To better understand the effectiveness of our design ERS manipulation, item responses can also be analyzed under a matching condition which should lead to an aggravated impact of ERS. If our analyses suggest that compensating ERS with our instructions is successful, the within subject design can be omitted in future studies and an a priori classification based on the superior measure can be used to select the respective ERS manipulation.

In the previous studies, we exclusively used psychometric models to detect patterns of ERS in item response data. To investigate whether the impact of ERS can be reduced by specific instructions that are targeted at respondents' ERS values, a different approach is chosen in the following study. As described in chapter 1.4 ERS is assumed to have a detrimental effect on criterion validity of psychological measures (e.g. Baumgartner & Steenkamp, 2001). Consequently, when item responses with ordinal response format are used to predict external criteria, including an appropriate measure of ERS in the predictive model should lead to increased predictive performance. Similarly, the criterion validity of polytomous item responses should be higher if respondents faced an instruction that was targeted to compensate their own response style, in contrast to standard instructions or an instruction which should further aggravate the impact of ERS. To evaluate the predictive potential of psychological scales, we rely on statistical methods from the field of predictive modeling. In the following chapters, we contrast predictive modeling with common practices in mainstream psychology and introduce some basic principles, as well as a powerful off-the-shelf prediction algorithm that can be used to investigate the effectiveness of the proposed ERS manipulations.

4.1.2 An Introduction to Predictive Modeling

The American Psychological Association defines psychology as the science of human experiences and behavior. It claims that "the understanding of behavior is the enterprise of psychologists" (see for example the FAQs on http://www.apa.org/support/about-apa.aspx). As a consequence, statistical inference in psychology aims at explaining psychological phenomena. The most common approach is to fit a statistical model to a sample of data and interpret the estimated coefficients with regard to some psychological theory. A prevalent assumption is that when a model is found which sufficiently describes the relationship between psychological variables in the sample, this model does help to provide insights into true psychological processes, but can also be used to predict human experiences and behavior. Although the predictive aspect of psychological science is mostly treated as a secondary goal in the explanatory approach, this has not hindered scholars to make bold claims about the success of psychological theories. Supposedly, a variety of concepts with high societal importance like job performance (Schmidt & Hunter, 1998; Ziegler, Dietl, Danay, Vogel, & Bühner, 2011) or health issues (Alarcon, Eschleman, & Bowling, 2009; Bogg & Roberts, 2004; McEachan, Conner, Taylor, & Lawton, 2011; Faragher, Cass, & Cooper, 2005) can be accurately predicted. However, it has been noted that these claims of predictive performance are seldom evaluated by suitable statistical methods (Yarkoni & Westfall, 2017). Worse, if predictive performance is evaluated in a rigorous way, it often turns out that the criterion validity is much lower than initially claimed. The authors convincingly argue that psychology's focus on "understanding" human behavior might actually be in conflict with achieving most accurate predictions of it. On the one hand, psychologists postulate large structural models which incorporate a high number of predictor variables. Due to generally small samples, parameter estimates are often unstable and show high variability between samples. On the other hand, researches suggest theoretically elegant models with a small number of linear relations, which are too simplistic to appropriately account for complex psychological phenomena. In both scenarios, traditional goodness of fit indices can easily mislead scholars to assume that they were successful in revealing some invariant psychological law. Research on the replicability of psychological science shows that many effects are not recovered if the same model is estimated in new samples (Open Science Collaboration, 2015). When original model estimates would be used to make predictions for new observations, this can be expected to fail even more spectacularly. Nevertheless, individual predictions are central goals in many divisions of psychology like personnel selection in the industrial-organizational setting, or relapse forecasting in clinical applications.

We agree with Yarkoni and Westfall (2017) that psychology could heavily profit from prioritizing prediction over explanation in a large number of research settings. Unknown to many psychologists, a similar claim has been made in statistics several years ago (Breiman et al., 2001). Predictive modeling emerged as a new field in the statistics and machine learning community which puts a primary focus on maximizing the predictive performance of statistical models. In contrast to the majority of psychological science, understanding how a set of predictors is truly related to some target variable is of minor concern here. This blind spot justifies the use of "black box" model classes that yield powerful predictions but

do not offer human comprehensible model equations which can be easily used to understand some underlying relationship. Predictive models are primarily evaluated with regard to their predictive performance, as it is highly difficult to assess their meaningfulness in any contextual sense. Predictive modeling has developed sophisticated methods to estimate the predictive performance of statistical models, as this is essential to achieve the main goal of the field.

Estimating Predictive Performance

Yarkoni and Westfall (2017) note that any statistical model is in some sense also a predictive model. In any statistical analysis, a frequently asked question is how useful the estimated model is in predicting new observations. In psychology, the coefficient of determination (R^2) is typically reported to assess the goodness of fit for the predominant linear models in the field. Yarkoni and Westfall (2017) rightly point out that most authors interpret R^2 as the amount of variance, which could be explained if the currently estimated model is used to predict new observations. However, it has been known for a long time that the standard R^2 and commonly used correction methods are overly optimistic (Yin & Fan, 2001) for this purpose. This is not a fault of the measure per se, but stems from the fact that R^2 is traditionally computed based on model predictions for the same observations which were already used to estimate the model coefficients. Naturally, any model better fits some data on which it was "trained" than on new, previously unseen observations. In predictive modeling where predictive performance is the ultimate goal, this bias is taken more seriously. Intuitively, an appropriate procedure would be to assess the predictive performance of a model on different observations than used for model estimation. After all, this is exactly how predictions of unknown observations work in practice. However, in many applications only one dataset is available and collecting new data from the same population to evaluate the model would be very costly or even impossible.

Instead of collecting new data, resampling techniques can be applied to use the dataset most efficiently. The simplest resampling scheme is to randomly split the dataset into a training set, which is used to estimate the model, and a test or holdout set for which predictions are computed based on the estimated model. It makes intuitive sense that the performance of the estimated model to predict observations in the test set is a reasonable estimate for the performance of the model fitted to the complete dataset to predict additional unobserved data. While the holdout technique might be sufficient for very large samples, it has several limitations if the sample is rather small: The crucial decision for holdout is the proportional size of the training and test set. In general, predictive performance of a model increases with the size of the training set up to a certain point, as parameter estimates improve if more information is available about the true relationship in the population. If the training set is chosen too small, performance estimation is downwardly biased. Predictions of new observations would be of higher quality as they are based on the whole sample which consists of more training observations. Thus, one goal might be to make the training set as big as possible, in order to approximate the performance of the model based on the whole sample. At the same time, performance estimates based

on small test sets have high variance, as predictive accuracy would greatly vary between several test sets with few observations. A general principle to reduce variance is to average across multiple estimates. For example, if equal sample sizes are chosen for the trainings and test set in the holdout procedure, the variance of the performance estimation can be reduced by switching the training and test sets, and averaging the two resulting estimates to gain an aggregated estimate of predictive performance. This approach is called two fold cross-validation (CV). In the general version of k-fold CV, the complete dataset is randomly divided in k equally sized parts. Each subsample serves as a test set which is predicted based on a model estimated on the combined sample of the remaining $k - 1$ parts. While each observation of the complete sample is used in $k - 1$ model estimations, each target value is predicted exactly once based on a model for which this observation was not included in the estimation process. Finally, the average of the k resulting performance estimates is used as an estimate for the full model based on the whole sample. In two fold CV, models are estimated on trainings samples only half the size of the whole dataset, which can result in downwardly biased performance estimates in small samples. Increasing k leads to bigger trainings sets but at the same time reduces the size of the corresponding test sets. The disadvantage of smaller test sets is somewhat reduced by the fact that performance is averaged across a greater number of estimates. However, as k increases, training sets highly overlap. This leads to increasingly correlated predictions of the k estimated models, which reduces the effectiveness of variance reduction achieved by averaging the performance estimates. Empirical results suggest that in most settings, cross-validation with five or ten folds provides a good trade-off between negative bias introduced by small training sets, and high variance introduced by small test sets as well as strongly correlated predictions of the estimated models (Hastie et al., 2009). In addition to k-fold CV, a variety of different resampling strategies exist, which might be preferable in some situations. A detailed description can be found in Bischl, Mersmann, Trautmann, and Weihs (2012). With the presented resampling techniques, it is possible to fairly assess the predictive performance of any algorithm, including the linear models dominantly used in psychological science. When doing this, one can expect that specialized algorithms from predictive modeling would often yield better results. The next paragraph focuses on the question why some predictive models might generally perform better than others.

Balancing Model Complexity

Although regularized linear models are highly predictive in many applied settings with very high numbers of predictors compared to the number of observations (for a detailed discussion see Hastie, Tibshirani, & Wainwright, 2015), the rise of predictive modeling is closely related to the development of highly non-linear model classes. Most powerful prediction algorithms are theoretically capable of reproducing almost arbitrarily complex relations between predictor and target variables. Flexible algorithms produce statistical models with very low bias in the sense that given enough data with little noise, they are able to mimic the true data generating process. However, high flexibility is not a desirable property without any control mechanism, as each dataset contains spurious peculiarities

which are not present in the population. When given a single dataset, many algorithms are flexible enough to perfectly interpolate all data points within the training set, thereby "predicting" them without error. Obviously, such an overfitted model has bad predictive performance on new observations, as it does not only incorporates the real target relationship but also the complete noise in the training set. A model which naturally tends to overfit in the above sense is said to have high variance. Imagine fitting the model to a number of training sets and comparing their predictions on a single test set. As every fitted model incorporates some specific noise, model predictions on the single test set will strongly differ as they are highly dependent on the respective trainings sets.

Some model classes naturally have low bias and high variance, while for others it is the other way round. One model class which is very flexible but tends to overfit are classification and regression trees (CART; Breiman, Friedman, Stone, & Olshen, 1984) which will be introduced in the next chapter. In the above terminology, the CART algorithm has low bias but high variance. In contrast, ordinary linear models are on the other side of the bias-variance spectrum. They are very inflexible and are mostly used to model linear functions while ignoring all interactions. If the real relationships is highly non-linear or contains complex interactions, predictions must be biased as, on average, they tend to be be off the point. At the same time, the inflexibility of the linear model makes it robust against overfitting, as long as the number of predictors is not too high. Thus, linear models are said to have low variance, as predictions based on different training sets tend to be be highly comparable. It can be shown that when predicting previously unobserved test data, the expected error of some statistical model is positively related to the bias and variance of its predictions, as well as the amount of random noise in the data. This bias-variance trade-off helps to explain why some predictive models show different performance under certain data situations: If there are strong complex relations between predictor and target variables in combination with high sample size and a high signal to noise ratio, flexible non-linear models with higher variance should be successful. In contrast, if the true relationship is weak or linear, data is sparse and the signal to noise ratio is low, inflexible models with low variance might yield good predictive performance. However, the bias-variance trade-off also comes into play within a single model class. This helps to understand why some algorithms are more powerful than others across a wider range of problems. The generally most successful prediction algorithms combine a high flexibility to reflect non-linear relationships and interactions, with a series of powerful regularization mechanisms that prevent the algorithm from overfitting. This makes it possible to efficiently estimate a predictive model with suitable complexity based on a limited amount of training data.

While many algorithms have some built-in mechanisms to automatically reduce overfitting, regularization is often achieved via some hyper-parameters which control the flexibility of the estimated model. As the right amount of model flexibility is strongly dependent on the properties of the data generating process, the degree of regularization needs to be adjusted to the nature of the problem at hand. Suitable values for the hyper-parameters have to be found before the final predictive model can be estimated. This tuning process is typically guided by some resampling strategy like k-fold CV. The algorithm is fitted with different combinations of hyper-parameter values in the training sets and predictive

performance is measured on the test sets. For the final model, the parameter combination is selected which achieved the best average performance on the k test sets.

The presence of hyper-parameters makes it more complicated to correctly estimate the predictive performance of the final model. If hyper-parameters are determined once for the whole dataset as described above and then applied in all CV iterations, predictive performance is typically overestimated, as all observations in the test sets were already used to find the optimal hyper-parameters. When the final model based on the whole sample is used to predict new unobserved cases, it can be expected to perform worse than estimated by the resampling scheme, in which the hyper-parameters were overly adapted to the respective test observations. This bias which can pose severe problems is still encountered in applied published research (Simon, 2007). Fortunately, it can be avoided by repeating the tuning of hyper-parameters in each iteration of the resampling process. This procedure is sometimes termed "nested resampling" (Bischl et al., 2012). In an outer loop, predictive performance is evaluated similar to the simple case described earlier. However, another resampling loop is created within each of the k model estimations to find the most optimal hyper-parameters in the fold. For example, if 10-fold CV is used for the performance estimation loop, one test set is predicted by nine tenth of the whole sample. If 5-fold CV is used in the inner tuning loop, each training set is again divided into five parts. For different combinations of hyper-parameters, the five inner test sets are predicted based on the corresponding inner training sets. The combination with the highest performance averaged over the five inner test sets, is then used for model estimation in the outer loop. This results in one of the ten outer performance estimates. The ten estimates in the outer loop are again averaged for an aggregated performance estimate of the final model, which would be fitted to the whole sample.

In this chapter, we could only give a brief, informal introduction into basic principles of predictive modeling like the bias-variance trade-off and how to evaluate the performance of predictive algorithms with or without tuning hyper-parameters. A more careful discussion can be found in the modern classic by Hastie et al. (2009) and its beginner friendly companion (James, Witten, Hastie, & Tibshirani, 2013).

4.1.3 Predictive Modeling with Random Forests

A predictive algorithm which has been found to be highly successful in many applications (Biau & Scornet, 2016) is the random forest (RF) invented by Breiman (2001). The RF is well suited for an introduction to predictive modeling, as its basic mechanisms are easy to understand and do not require strong mathematical or statistical background. Moreover, the RF requires little tuning of hyper-parameters and does not overfit, which makes it more accessible to unexperienced users.

A RF is an ensemble method which combines a large number of decision trees into an aggregated estimator that yields better predictive performance than its single components. Each decision tree is constructed with the CART algorithm (Breiman et al., 1984). CART is a supervised learning algorithm: Some target variable Y is predicted based on a set of p predictor variables, using training data with known values on both target and predictor

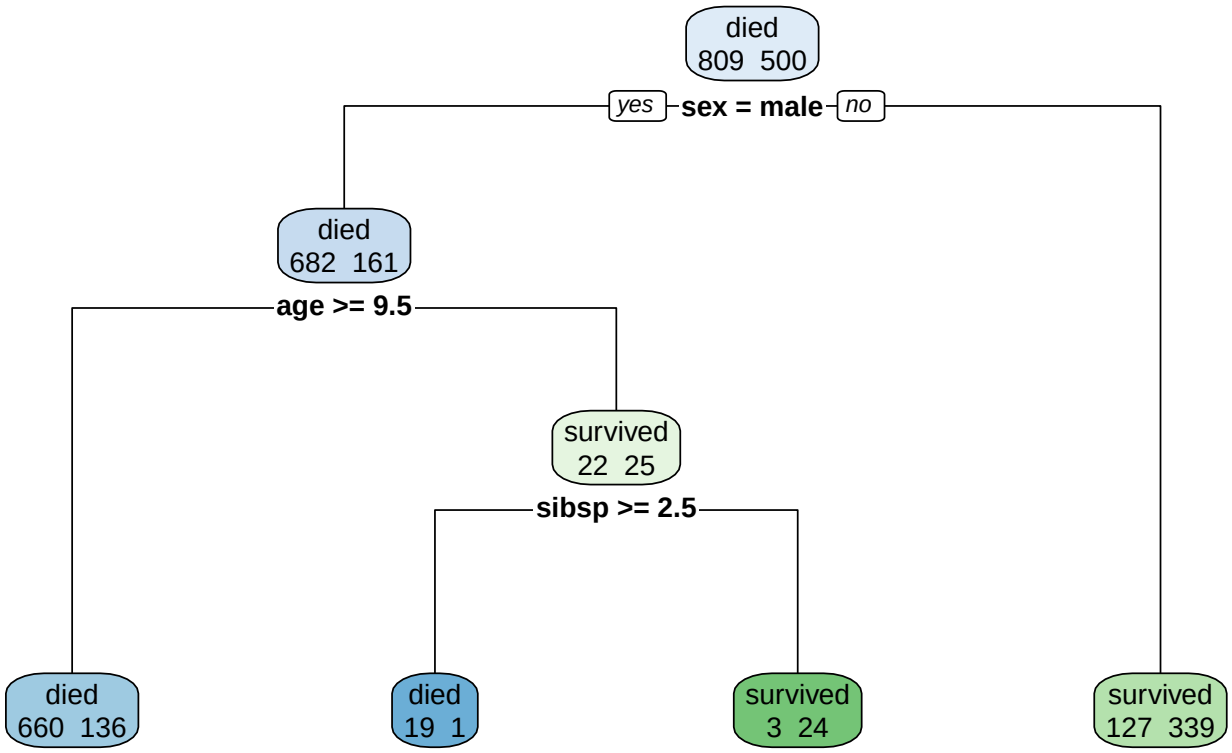

Figure 4.1: Classification tree of the Titanic dataset in R. Death is predicted by the covariates age, sex, passenger class (pclass), number of siblings or spouses aboard (sibsp), and number of parents or children aboard (parch).

variables. In classification trees, Y is a categorical variable with m categories. In regression trees, Y is metric. The main idea is to start with some training data, sequentially split it into smaller subsamples based on rules derived from the available covariates, and make constant predictions within each of the final subsamples.

Figure 4.1 shows a classification tree for the Titanic dataset, which is available in R.[1] The fitted model predicts if a passenger of the Titanic survived the disaster, based on the passenger variables age, sex, passenger class, number of siblings or spouses aboard, and number of parents or children aboard. Each box in the figure is called a tree node and represents a subsample of the data, with the whole training set displayed on top of the tree. For each node, we see the number of deceased passengers on the left and the number of survivors on the right. The label of the most frequent target class is also shown. In our example, the first node shows that the training set included 1309 passenger of which 809 did not survive. Each non-terminal node further displays the name of a variable in the dataset which is used to split the current sample. The exact split-point value within this variable is also shown. Nodes on the bottom of the tree are called leaves. The most frequent

[1] The model was estimated with the `rpart` package. The dataset was taken from the `rpart.plot` package, which was also used to create the Figure 4.1. Default settings of `rpart` were used except for the complexity parameter which was set to 0.02.

class label in a leaf can be used to make predictions about other unobserved passengers. For example, the tree would predict that a seven year old boy traveling together with at least three siblings died, while it would predict that any woman survived regardless of her values on the remaining covariates.

The structure of a decision tree is determined by the following algorithm, starting with the whole training set:

The CART Algorithm

1. Term the current sample the father node F. For all observations in F, compute an impurity measure $I(F)$. It quantifies some average error per observation that results from a constant prediction for all observations in F. In regression trees, the impurity function is typically the mean squared error (MSE)[2], which in this case is identical to the variance of the target variable in F:

$$I_R(F) = \frac{1}{|F|} \sum\nolimits_{(x,y)\in F} (y - \bar{y}_F)^2 \tag{4.1}$$

   $|F|$ is the total number of observations in F. In classification trees, the simplest impurity function is the mean misclassification error (MMCE)[3], which is just the ratio of observations not belonging to the biggest class in F:

$$I_C(F) = 1 - \underset{m}{max} \frac{n_m}{|F|} \tag{4.2}$$

2. Iterate over all covariates and for each covariate over all possible split-points, defined by all different covariate values of the observations in F. For each split-point, compute the impurity reduction (IR):

$$IR = I(F) - \frac{|L|}{|F|} I(L) - \frac{|R|}{|F|} I(R) \tag{4.3}$$

   L and R reflect the left and right child nodes if the split is performed.

3. Split the sample at the point which yields the largest IR. For metric features, the actual split-point is ambiguous and is typically chosen in the middle between the two closest observations.

---

[2] In general, the MSE is defined as $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and describes the average squared deviation between some model predictions and the true target values.

[3] In general, the MMCE is defined as $MMCE = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$ and describes the proportion of classification errors made by a prediction models. Here, I is the indicator function, returning 1 if the statement in brackets is true and 0 if it is not. Although the MMCE is well suited to illustrate the general principle for classification, other impurity functions typically yield better predictive performance in practice. In the classification analyses presented later, the Gini index was used as impurity function which is also the default in the original CART algorithm. For further details, we refer the interested reader to Hastie et al. (2009).

4. Recursively repeat steps 1-3 for both child nodes until one of several stopping criteria is satisfied: the minimal number of observations in a node to try further splitting, the minimal number of observations in each leaf, the minimal impurity reduction to try further splitting, and the maximum number of tree levels.

5. Make an optimal constant prediction in each of the final leaves. In regression trees, this is the mean of Y for all observations in the leaf. In classification trees, the constant prediction is the label of the majority class in the leaf. Predicted values for new observations are obtained by selecting the constant prediction of the leaf to which they belong, based on their values on all covariates.

CART has several favorable properties: The Titanic example already showed that decision trees are easily interpretable due to their graphical representation. At the same time, they can capture complex interactions between variables and model non-linear functions between the predictors and the target variable. If stopping criteria are chosen in a way which allow deep trees with a high number of leaves, CART has very little bias. CART is invariant to monotone transformations of the covariates and can easily handle categorical variables, outliers in the covariates, as well as missing values. Last but not least, CART has a built-in selection process to deal with a high number of covariates, as the algorithm should simply ignore irrelevant variables in the splitting process.

One disadvantage of CART is that constant predictions are made within each leaf of the tree. Linear or smooth relationships between the predictors and the target variable can only be approximated by a large number of splits, which typically leads to suboptimal predictive performance. However, the main problem with CART is its high tendency for overfitting. Note that in theory, the tree growing algorithm can continue until each leaf contains only a single observation. At least in this case, strong overfitting occurs, as the tree captures the complete noise contained in the training data. To avoid this behavior, the bias-variance trade-off of CART can be controlled by adapting the stopping criteria.[4] Increasing the number of leaves and reducing the number of observations in each leaf decreases the bias of CART. At the same time this leads to overfitting and an increase in variance. Even with careful tuning of the model complexity, minimal changes to the training set can lead to highly different structures of the resulting trees (Hastie et al., 2009). As a consequence, the proclaimed high comprehensibility of CART has to be taken with a grain of salt, as a changing tree structure may imply a completely different interpretation of the estimated model. Moreover, the instability of the tree structure also implies highly variable predictions for new observations. This explains why CART and other tree growing algorithms usually yield only moderate predictive performance in practice.

Decision trees are conceptually related to the Rasch and PC trees introduced in chapter 2.1.1, which are built on the model-based recursive partitioning framework (Zeileis et al., 2008). In model-based recursive partitioning, a statistical model is fitted within each leaf of

[4] Instead of tuning the stopping criteria, pruning is a popular approach which scales down a fully grown tree in an attempt to reduce overfitting. Pruning is not described here, as it is unnecessary for the RF.

a tree, while only constant predictions are computed in simple decision trees. Consequently, the splitting criterion in model-based recursive partitioning relies on hypothesis tests, which assess the parameter stability of the statistical model in the current node.

The Random Forest Algorithm

Breiman (2001) found a simple, but powerful procedure to increase the predictive performance of the CART algorithm he invented earlier. The main idea of the RF is to capitalize on the low bias of deep trees, while reducing their variance. Variance reduction is achieved by fitting a large number of slightly modified decision trees to bootstrap samples of the original training set, and averaging the predictions from all individual trees.

In detail, the RF algorithm works as follows:

1. Draw B full bootstrap samples with replacement from the training set.
2. To each bootstrap sample, fit a tree with the CART algorithm described earlier, except for two modifications:
   - Trees are grown as deep as possible with minimal stopping criteria.[5]
   - For each split in the tree, only a sample of $mtry$ randomly drawn covariates are available to find the optimal split-point.
3. Aggregate the predictions of the B resulting trees. For regression, the RF estimator is the mean of all predictions for a certain observation, resulting from all trees. For classification, the estimator is the class which is predicted most frequently for the observation by the trees. This is often called the "majority vote".

In the RF, the CART algorithm for growing trees is modified for two reasons: The first modification aims at reducing the bias as far as possible, in order to guarantee a high flexibility to capture complex dependencies in the data generating process. The second modification is a clever mechanism to "decorrelate" the trees, thereby increasing the effectiveness of the variance reduction achieved by the final averaging. Intuitively, using lower values of $mtry$ than the total number of predictors p can be thought of as "giving weaker predictors a better chance". For simplicity, imagine a variable with a much stronger effect on the target than all the others. Even if all trees in the RF are estimated based on slightly different bootstrap samples, the rough structure of all trees might look highly similar with $mtry = p$. The important variable is usually used in the early splits, which have the highest impact on the general tree structure. However, if $mtry < p$, the strong predictor might not be available and a weaker variable is chosen instead. The individual trees, resulting from this shakeup of the general tree structure, produce more diverse predictions. This leads to a reduced variance of the aggregated RF estimator without inducing a strong bias. Typically, $mtry$ is tuned by CV or fixed to values around

[5] In the `randomForest` package in R, the minimum node size is 1 for classification and 5 for regression. No further stopping criteria are applied per default.

$\sqrt{p}$. It can be shown that low values of *mtry* are especially useful if the correlation between the covariates is high (Hastie et al., 2009). It makes intuitive sense that with many highly correlated predictors, the tree structure markedly changes only if none of those variables are available for splitting. This is much more likely if *mtry* is small.

Variable Importance

The RF typically shows far better predictive performance than single trees, sacrificing only one of the important advantages of CART. The interpretability of single decision trees gets lost with the aggregation of the RF. A graphical inspection of several hundred trees with deep, highly variable structure is not useful. In order to evaluate the impact of single features in the predictions, several variable importance measures can be computed. A common one is the unstandardized out-of-bag (OOB) permutation measure[6], although some improved versions have been suggested (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). Each tree in the RF is fitted to a full bootstrap sample drawn with replacement. It can be shown that for the estimation of each tree, about one third of the observations from the complete training set is not used. These are called the OOB observations. To assess the importance of one covariate, subject identifiers for this variable are randomly shuffled within each of the OOB samples. OOB samples are then predicted by their respective trees. The resulting prediction error is compared with the prediction error for the un-shuffled OOB samples. Prediction error is measured by MSE for regression and MMCE for classification. To compute the final variable importance measure, the differences in prediction error are averaged across all trees.

Reshuffling synthetically destroys all associations of the covariate with the target variable. However, as this is done after the model is fit, the decrease in accuracy does not measure the best possible performance if the covariate of interest was not available. If the RF were built again without the covariate, other variables might step in (Hastie et al., 2009). It has been shown that the OOB variable importance measure is upwardly biased for covariates with a high number of different observed values. This has to be taken into account when comparing variables with varying granularity (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). Further note that the OOB permutation variable importance contains not only main effects but also all modeled interactions with the remaining covariates. One consequence is that for a variable with high importance, it is unclear how much of its effect is additive (would be unchanged if other variables were missing) or depends on interactions (would be lower if other important variables were missing).

Partial Dependence Plots

Individual partial dependence plots (Goldstein et al., 2015) are a graphical method to visualize the average additive effect of a covariate and to assess how strong the covariate

[6] The OOB permutation measure follows a general principle. If slightly adapted, it can be applied to any predictive modeling algorithm.

interacts with the remaining ones.[7] The RF does not easily reveal how a covariate is used in its predictions. The main idea of individual partial dependence plots is that the relationship between one covariate of interest and the target in the estimated model can be inferred by creating predictions for a series of different values of the covariate. First, a grid of covariate values is chosen between the observed minimum and maximum in the dataset. Then, predictions are computed for all observations, with each covariate value on the grid. Importantly, original values are used for all the remaining covariates. While the predicted target values are used in regression, predicted class probabilities are used in classification. In an individual partial dependence plot, the grid of the covariate is plotted against the target predictions, while each observation in the dataset is represented by a single line. Individual partial dependence plots help to identify interactions with the covariate of interest. If the effect of the covariate is exclusively additive, all lines should have approximately the same form. As the covariate of interest is fixed by the grid, varying trends can only be attributed to the remaining covariates, on which the observations naturally differ.[8] If interactions are detected and one is particularly interested in how the covariate interacts with another categorical variable, lines for observations with the same value on the categorical variable can be labeled by the same color. Color patterns can reveal if the categorical variable does interact with the covariate of interest and might help to interpret the interaction. If one is only interested in the average additive effect of a covariate, individual predictions can be averaged for each value on the grid. This yields the original dependence plots as introduced by Friedman (2001).

The methods presented in this chapter are powerful tools which can be used by researchers in psychological science to investigate many important questions (see the discussion in Yarkoni & Westfall, 2017). This introduction should help interested psychologists to delve into the fascinating field of predictive modeling or proceed to other disciplines of machine learning.

4.1.4 Aim of Study and Research Questions

The main goal of this study is to examine whether an instruction, which is targeted at compensating each subject's extreme response tendencies, can counterbalance the impact of ERS and lead to more valid item responses. The effectiveness of the ERS manipulation is investigated by comparing the criterion validity for related self-reported behavior. Different conditions of matching the ERS instruction to subjects' estimates of ERS are investigated. Criterion validity is determined by methods from predictive modeling, using all questionnaire items as separate predictors in RF models. A within-subjects design is used to administer two different ERS instructions to participants in the experimental con-

[7] Similar to the out-of-bag variable importance, partial dependence plots are a useful tool that can be used in combination with any predictive modeling algorithm.

[8] To increase the interpretability of the individual partial prediction plots, individual lines are sometimes centered so that the prediction on the first grid point is zero for each observation. In some cases, interactions with the covariate of interest can be seen even better in the centered version, as all lines should strongly overlap here in the case of a perfectly additive effect.

dition, while a control group with neutral instructions is also included. Previous to the ERS manipulation, all participants once answered all items under standard instructions. Extreme responding is assessed by an ERS index from additional heterogeneous items and self-report measure of ERS. For the main analyses, both ERS estimates are used to choose those item responses for each subject which were given under the instruction that should either compensate or aggravate the effect of ERS. The compensation and aggravation settings are compared with item responses from the control group. Building up to the primary analysis, a series of preliminary analyses are used to check necessary prerequisites. We test the validity of the ERS instructions, rule out several confounding factors, and investigate whether measures of ERS contribute to the predictive performance of questionnaire items.

Preliminary Analyses

Presence of ERS Our main research questions require that ERS is present in our data. To ensure that the scales used in the current study are affected by ERS, we compute PC trees (see chapter 2.1.1) under standard instructions with an ERS index from heterogeneous items as covariate. We expect several splits based on the ERS index, revealing the same threshold pattern as in study 1 and 2. A split should also emerge when a binary self-report measure of ERS is used as covariate in the PC trees.

ERS Manipulation Checks To make sure that the developed ERS manipulation has an effect on item responses, we compute PC trees with the type of instruction (extreme-responding, mid-responding, or neutral) as covariate. If the instructions have an effect on the subjects' item responses, PC trees should split the sample in three parts based on the different manipulation conditions. Moreover, we expect that the threshold patterns in the resulting leaves do reflect the corresponding instructions. Specifically, we expect narrower thresholds and larger regions for the highest and lowest response categories under the extreme-responding instruction than under the mid-responding instruction. The threshold pattern under the neutral instruction should lie in between. We will also investigate this graphically, by comparing histograms of item responses under the different instructions.

Potential Order Effects In the experimental group, the order of the within-subjects extreme and mid-responding instructions is randomized. However, randomization does not hold when instructions are matched to subject's ERS estimates for the main analyses. We investigate potential order effects, by graphically comparing histograms of item responses for the control subjects who complete the same scales three times under neutral instructions. Without order effects, histograms from the three parts of the questionnaire should have approximately the same shape. To rule out that repeatedly answering the same items has an effect on criterion validity which could confound the main analyses, we also compare the correlation of scale sum scores with the target variables between the three consecutive questionnaire rounds of the control group. If order effects do not affect criterion validity, questionnaire parts should not be systematically related to the size of the correlations.

Matching ERS Instructions and the Impact of ERS To investigate whether matching ERS instructions to participants' extreme response tendencies reduces the impact of ERS, we compute PC trees under different matching conditions with the ERS index as covariate: In the compensation setting, participants' item responses will be taken from the part of the questionnaire in which the given ERS instruction should compensate their individual ERS tendencies. In the aggravation setting, item responses will be chosen so that the ERS instruction should further aggravate the response style. In the control setting, we analyze item responses of the control group which always respond under neutral instructions. Matching is either based on a median split of the ERS index from heterogeneous items or based on the binary self-report measure of ERS. If matching the ERS manipulations to respondents' own response style reduces the impact of ERS, we would expect the following pattern: In the control group, the sample should be split based on the ERS index, similar to the analyses of the whole sample under the standard instruction. In contrast, we expect no splits in the compensation setting and a higher number of splits in the aggravation setting. If no clear order in the number of splits can be observed, we at least expect the respective threshold pattern to suggest a lower impact of ERS in the compensation setting than in the aggravation setting.

Contribution of ERS Measures to the Prediction of External Criteria The idea that ERS instructions might compensate effects of ERS, thereby increasing the predictive performance of questionnaire items, has an important prerequisite: If measures of ERS are included in a predictive model together with questionnaire items, they should contribute something to the predictive performance. To investigate this, we fit two RF models to predict a series of self-reported behaviors. In the first model, item responses from psychological scales under standard instructions, sex, and age are used as predictor variables. In the second model, we further include the ERS index from heterogeneous items and the self-report measure of ERS. For both models, we assess predictive performance using resampling techniques. If the ERS measures contribute to the predictions, we expect better performance of the model in which ERS variables are included. The impact of ERS on predictions will be further investigated by estimating variable importance of both ERS measures and visualizing the effect of the ERS index in individual partial prediction plots.

Primary Analyses

Matching ERS Instructions and Predictive Performance In the final analysis, we investigate whether item responses under instructions targeted at compensating participants' individual ERS tendencies leads to better predictive performance, compared to instructions that are supposed to aggravate the impact of ERS. In these analyses, we again apply the RF algorithm but use only item responses from the psychological scales as predictors. We match the ERS instructions to subjects' response style either based on the median split of the ERS index or based on the self-reported ERS measure. The analysis is then performed under both matching regimes. For all models, predictive performance is again evaluated with resampling. If matching ERS instructions to subjects' ERS estimates

leads to more valid item responses, we expect the highest predictive performance under the compensation setting and the lowest performance under the aggravation setting. The performance in the control setting should lie between the other two conditions. If both ERS measures capture extreme responding, we expect similar results regardless whether the ERS index from heterogeneous items or the self-report measure of ERS is used in the matching process.

4.2 Methods

4.2.1 Scales and Variables

Impulsivity and Order

We again used two of the NEO-PI-R facets (see chapter 2.2.1) for which Wetzel, Carstensen, and Böhnke (2013) showed superior fit of the constraint mixed Rasch model, suggesting that they are good candidates to reveal effects of ERS. This notion was further supported by the PC tree analyses of those scales in study 1.[9] The eight item scale Impulsivity (IMP) is a facet of the Big Five factor neuroticism. It captures impulsive behavior like the inability to control feelings or to suppress the desire for food and other consumer goods. In contrast to other concepts of impulsivity, it does not entail fast decision making and risk taking. Similar to study 2, the Order (ORD) scale was also used. For a description of the scale see chapter 3.2.1. All IMP and ORD items were rated on the original fully labeled five-point Likert scale of the German NEO-PI-R (Ostendorf & Angleitner, 2004).

ERS Index

Our questionnaire contained 30 items of heterogeneous content, which were used to compute an ERS index similar to the previous studies. ERS items were sampled by the same procedure and from the same item data banks as in study 2. For a description of the procedure see chapter 3.2.1. The wording of the final set of ERS items and their original source can be found in Appendix D.1. We sampled 40 questions until enough suitable items were found. The original response format of the German NEO-PI-R was also used for the ERS items. We mostly retained the original item wording. Some slight modifications were made to increase comprehensibility when encountered out of context, and to ensure that item wordings were compatible with the response format of the NEO-PI-R.

Self-Reported ERS

Self-reported ERS (SRERS) was measured by the same dichotomous item also used in study 2 (see chapter 3.2.1). While the question wording was exactly the same as in study 2,

[9] As a reminder, check out the presented PC trees in Appendix B.3

response categories were slightly altered to improve comprehensibility.[10]

Target Variables

We constructed a series of questions which would be used as target variables in the main predictive modeling analyses. These are listed in Table 4.1 while the German wording of the original questions can be found in Appendix D.2. A dichotomous response format with categories "Yes" and "No" was used for the variables Clothes, Argument, Lying, Electronics, Desk, Dishes, Document, and Bed. For the Snacking question, a check-box was presented next to each snack to mark if it was consumed, as well as another box to enter the consumed amount. As a result of pretesting the Series question, a third category labeled with "I do not watch any series." was presented in addition to "Yes" and "No". Height was reported in centimeters while Weight was reported in kilograms. Both variables were used to compute each participant's body mass index (BMI). Height and Weight were not used as separate target variables in predictive modeling. For each target variable, either IMP or ORD is considered the corresponding primary scale in the sense that the target variable can be expected to be successfully predicted by the items of this scale. The IMP items should be well suited to predict the variables Argument, Clothes, Electronics, Lying, Series, Snacking, and BMI. ORD items should be related to Bed, Document, Dishes, Desk, and Vacuum.

As target variables for the IMP scale, we considered impulsive buying and unhealthy snacking, two behaviors which seem to be positively related (Verplanken, Herabadi, Perry, & Silvera, 2005). Impulsive buying has been shown to be associated with personality traits (Verplanken & Herabadi, 2001). Likewise, a general measure of impulsivity which does not focus on specific behaviors, has been reported to predict self-reported snacking (Churchill, Jessop, & Sparks, 2008). The IMP scale might be an even better predictor of snacking, as it contains two items which explicitly target the tendency of eating too much. If impulsivity is related to snacking, it may have a mediated effect on body weight. Indeed, the IMP facet of the NEO-PI-R has been found to be a positive predictor of BMI and other obesity measures like hip-to-waist ratio in two studies with large samples (Terracciano et al., 2009; Sutin, Ferrucci, Zonderman, & Terracciano, 2011). The relationship even holds prospectively for BMI measured several years after the personality assessment. Interestingly, both studies also found a negative relationship between BMI and the ORD facet, although the effect was weaker than for IMP. In an influential taxonomy of impulsivity, the facet urgency is described as "the tendency to commit rash or regrettable actions as a result of intense negative affect" (Whiteside & Lynam, 2001). As urgency seems to be closely related to the IMP scale of the NEO-PI-R, the target variables Argument and Lying were constructed

[10] German wording: "Manche Personen tendieren in Fragebögen häufig dazu, eine der äußeren Antwortmöglichkeiten anzukreuzen, während andere Personen häufig dazu tendieren, eine Antwortmöglichkeit nahe der Mitte anzukreuzen. Wie würden Sie sich spontan am ehesten einschätzen?"

"Ich tendiere häufiger dazu Antworten anzukreuzen, die in der Mitte liegen."

"Ich tendiere häufiger dazu Antworten anzukreuzen, die am Rand liegen."

to reflect this aspect of impulsivity. Apart from eating, several IMP items assess a general inability to resist pleasurable stimuli. The Series variable was considered a contemporary impulsive behavior which could also be affected by this unspecific aspect of impulsivity.

ORD targets have high face validity, as they are all related to cleaning or other orderly behaviors of everyday life. All questions were adapted from the Behavioral Indicators of Conscientiousness (Jackson et al., 2010), which have been shown to be positively related to conscientiousness in pure self-report studies. Notably, the Behavioral Indicators of Conscientiousness also include more impulsive behaviors like shouting at strangers, playing sick to avoid something, spontaneous shopping in contrast to buying stuff from a shopping list, telling lies, and eating until feeling sick. These behaviors seem to be negatively correlated with conscientiousness.

As we used an online questionnaire in order to reach an appropriate sample size, we had to rely on participants' self-reports. However, when constructing the target questions, we tried to maximize the objectivity of our results, by taking the following aspects into account: Associations between two sets of self-report items can be inflated by ERS and other forms of method bias (Podsakoff et al., 2011). Therefore, we did not use an ordinal response format for any of the target variables. A binary response format was used for the majority of items. According to our results from study 2, Yes/No items are probably unaffected by ERS. For the remaining questions, subjects could freely report numbers and check whether they consumed something or not. These formats should also be free of ERS. Furthermore, we put our focus on concrete behaviors or observations and used a recent time frame to increase the likelihood that subjects give accurate answers based on active memory retrieval. We also tried to use precise wording to ensure that subjects interpret our questions as similar as possible. Target items were validated in a pretest: All variables showed promising correlations with the sum score of their primary scale. Moreover, we asked participants whether they were able to remember the respective behavior. This seemed to be the case.

Table 4.1: Target Variables

	Impulsivity
Argument	Last week, did you say something during an argument which you would have liked to take back?
Clothes	Last week, did you spontaneously buy clothes online or in store, without planning it beforehand?
Electronics	In the last two weeks, did you spontaneously buy electronic devices or consumer electronics, without planning this beforehand?
Lying	In the last two weeks, did you lie in order to avoid an appointment or date?
Series	Remember the last time you watched your favorite series. Did you watch more episodes at a stretch than you had intended?
Snacking	Think about yesterday. Which of the following listed snacks/sweets did you consume yesterday and how much of them? One chocolate bar/cereal bar One handful of chips/nuts/saltsticks One share of chocolate (three pieces)/one chocolate truffle One large cookie or two small biscuits One piece of cake/one piece of pastry One handful of gummy bears/fruit gums/candies One scoop of ice cream/one Popsicle One cup of yogurt/pudding
Height	How tall are you?
Weight	What is your current weight?
	Order
Bed	Did you make your bed this morning?
Document	Remember the last time you received an important document (e.g. insurance papers, school certificate, contract). Did you immediately file that document?
Dishes	At this moment, are there any unwashed dishes you used outside of the kitchen area?
Desk	At this moment, is your desk or workstation tidy?
Vacuum	Think about the last two weeks. In that period, how many times did you vacuum your bedroom?

Note. Target variables are listed with respect to their primary scale (Impulsivity or Order). Height and Weight were combined into body mass index (BMI) for all analyses.

4.2.2 Questionnaire Design and ERS Manipulation

We designed and hosted an online questionnaire on the SoSci Survey framework.[11] The complete questionnaire had 29 pages and was presented in German. A progress bar, the logo of the university and a link to the author's university email address was shown on top of each page. Browsing through the questionnaire was achieved by a forward button. If subjects did not respond to all questions on a page, proceeding was not possible and a reminder message appeared.

At the beginning of the questionnaire, subjects were thanked for their participation in a study about the relationship between personality and attributes which would take about 15-20 minutes. They would get reimbursed with course credit and have the ability to participate in a lottery. We instructed them to respond honestly to all questions, no matter if those appeared unusual or similar to earlier questions. They should carefully read all questions and instructions. If respondents felt like they did not know an answer on a question, they should choose the most applicable response options instead of skipping the question. Finally, subjects were informed that their data would be treated anonymously and only be used for instructional and scientific purposes.

After the introduction, we retrieved the following demographic variables: sex, age, highest educational qualification[12], and current profession.[13] If the student category was selected in the question about profession, an additional box appeared where participants had to specify if they are enrolled in psychology or not.

The main part of the questionnaire is visualized in Figure 4.2. It can be structured into three parts – A, B, and C – followed by the target variables and the SRERS item. Part A contained the ERS items mixed with the IMP and ORD items. The item order was the same for all participants. Each page contained a maximum of five items and was always introduced with the question: "How much do you agree with the following statements?"

Preceding both part B and part C, all participants were informed on a separate page that they would now complete some questions again, and that they should reconsider which response is most applicable to them. Moreover, they were ensured that the reason for the repetition is not how good they can remember what they had responded earlier. This was followed by one of three different ERS manipulations, which are presented in Table 4.2.

The extreme and mid-responding instructions were followed by two examples: the decision process was explained for one case in which a subjects was indecisive between the third and fourth category, as well as another case in which a subject was indecisive between the first and second category. Examples were supposed to clarify that participants should only alter their response if they were undecided. Responses should be shifted only by one category, instead of automatically selecting the midpoint or one of the most extreme

[11] For more information, visit https://www.soscisurvey.de/.

[12] German education categories: Kein Schulabschluss, Hauptschulabschluss/Volksschulabschluss, Realschulabschluss/Mittlere Reife, Allgemeine Hochschulreife/Fachhochschulreife, Hochschul/Fachhochschulabschluss, Promotion.

[13] German profession categories: Schüler/in, in Ausbildung/Lehre, Student/in, Angestellte/r, Selbstständig/Freiberuflich, Arbeitslos/Arbeit suchend, Hausmann/Hausfrau.

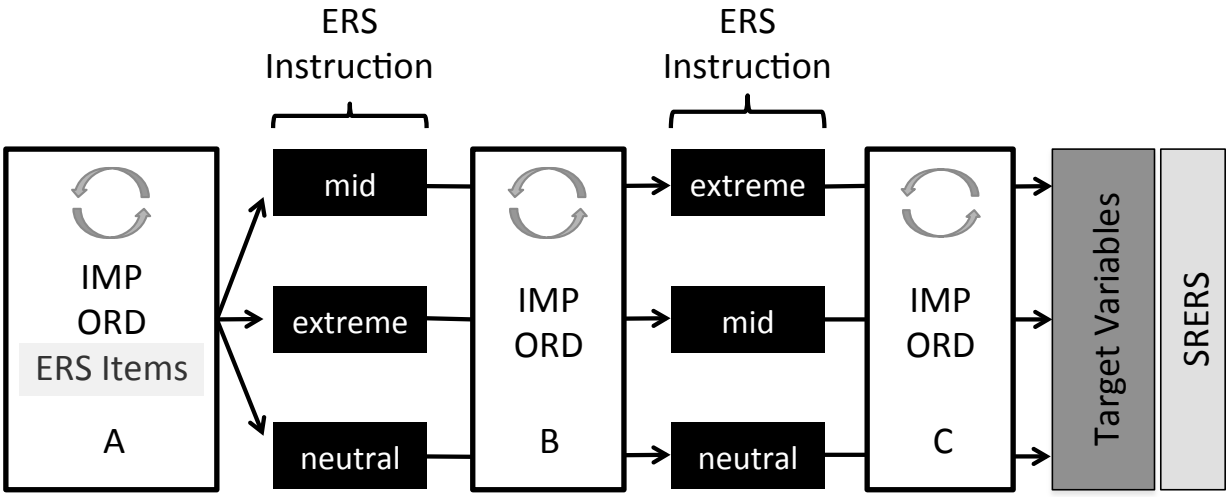

Figure 4.2: Flowchart describing the structure of the online questionnaire. Impulsivity (IMP) and Order (ORD) items where encountered in part A, B, and C of the questionnaire. Within each part, items were presented in mixed order. Participants in the mid-extreme responding condition answered part B under the mid-responding instruction and part C under the extreme-responding instruction. For subjects in the extreme-mid responding condition, the order was reversed. The control group answered part B and part C under neutral instructions. SRERS = self-reported extreme response style.

Table 4.2: Wording of ERS Instructions

Extreme	IMPORTANT: Should you be unable to decide between two response options, now rather choose that response option which is closer to the side.
Mid	IMPORTANT: Should you be unable to decide between two response options, now rather choose that response option which is closer to the center.
Neutral	If you cannot decide between two response options, try to find that response option which is more applicable to you.

Note. Instructions for the ERS manipulations were presented in a green box. The box was repeated on each of the following pages together with the instruction already used in part A.

categories. Examples consisted of explanatory text in combination with pictures showing which category to choose. The same item was used for all examples. Screenshots of questionnaire pages illustrating the three ERS manipulations can be found in Appendix D.3. On pages displaying the ERS manipulation, the next button stayed invisible for a few

seconds to avoid that participants accidentally skipped the instructions.

Before facing the ERS instruction in part B of the questionnaire, participants were randomly assigned to one of three experimental conditions. Participants in the mid-extreme condition encountered the mid-responding instruction in part B, followed by the extreme responding manipulation in part C. In the extreme-mid condition, participants encountered the extreme-responding manipulation in part B, followed by the mid-responding manipulation in part C. We also had a control group whose subjects encountered the neutral manipulation in both part B and part C of the questionnaire. Urn randomization was used to ensure that participants were evenly assigned to all three conditions. In part B and C, IMP and ORD items were presented in the same alternating order, starting with the first IMP item followed by the first ORD item. Four items were shown on each page.

When part C was completed, the target questions were presented in the following order: Snacking, Clothes, Vacuum, Argument, Desk, Dishes, Lying, Document, Electronics, Series, Bed, Height, and Weight. Except for Snacking, which was presented on a separate page, always four target questions were shown at the same time. Finally, the SRERS item was displayed on a separate page.[14]

[14] Target questions were followed by one page, on which participants could enter their email address if they wanted to take part in the lottery. By design of the survey software, addresses could not be associated with the subjects identifiers to assure privacy. On a final page, subjects were thanked again to take part in the study. A link was provided to download a certificate for course credit.

4.2.3 Participants

From March until July 2015, 934 participants started our online questionnaire and 789 finished it. As participants could not proceed through the questionnaire without answering all questions, there were no further missing data. However, we excluded one female participant with a BMI score of 9.47 from all analyses.[15] Consequently, the analyzed sample consisted of 788 subjects (222 men and 566 women) between the age of 18 and 74 ($M = 25.35,\ SD = 8.02$). Of 586 included students, 152 were enrolled in psychology. Only 48 subjects did not have at least a subject-related entrance qualification for university.

Participants had the possibility to win one of four Amazon vouchers worth of 50 euro and all subjects could get a certificate for course credit. An invitation to the online study was sent to the study mailing list of LMU Munich. Additionally, the study was promoted in Facebook groups and popular science outlets like www.psychologie-heute.de.

In the analyzed sample, 263 participants were part of the mid-extreme condition, while 262 participants were part of the extreme-mid condition. The control group consisted of 263 subjects.

4.2.4 Statistical Analyses

All 30 ERS items were used for the computation of the ERS index. Similar to the previous studies, ERS items were first recoded to a response style scale. As all ERS items were rated on a five-point Likert scale, the recoded values on the response style scale were: 1, 0.5, 0, 0.5, 1. A subject's ERS index was computed as the mean response on the recoded ERS items. In the SRERS measure, subjects were coded with one, if they expressed a tendency to choose extreme categories and zero if they expressed a tendency to choose the midpoint category. To simplify the predictive modeling analyses, we decided to ignore the "I do not watch any series." response option of Series and treated it as another binary target variable. This increased the accessibility and clarity of our results by avoiding the peculiarities of classification with more than two classes. Also, we were able to conveniently report all classification results side by side. Note that as a consequence, all correlations and predictive modeling analyses which include the Series variable are based on a reduced sample size, as only subjects who reported "Yes" or "No" were analyzed.

To evaluate whether matching the ERS instruction to the subjects' own response tendencies might reduce the impact of ERS, we used the within-subjects design to construct a series of artificially matched datasets. These were used for the main predictive modeling analysis, and for one PC tree analysis. Matching was done once based on a median split of the ERS index, and once based on the SRERS variable. The compensation and aggravation settings only included subjects from the experimental conditions mid-extreme and extreme-mid. The remaining subjects were used in the control setting.

[15] The excessively low weight of 30kg with a corresponding height of 1.78m was most likely a responding error.

Compensation setting: For each participant with an ERS index above the median, we included the IMP and ORD item responses from the part of the questionnaire with the mid-responding instruction. For participants with an ERS index below or equal to the median, item responses under the extreme-responding instruction were used. When matching was based on SRERS, responses under the mid-responding instruction were used for participants in the group of extreme responders and vice versa.

Aggravation setting: IMP and ORD item responses under the extreme-responding instruction were used for each participant with an ERS index above the median and responses under the mid-responding instruction were chosen for subjects with an ERS index below the median. When matching was based on SRERS, responses under the extreme-responding instruction were used for participants in the group of extreme responders and vice versa.

Control setting: For each participant in the control group, we randomly selected part B or C of the questionnaire and took all IMP and ORD item responses from the chosen part.

A series of descriptive statistics were computed: Similar to the previous studies, we report the mean and standard deviation of the ERS index, the mean absolute correlation, as well as the maximum absolute correlation between ERS items, and 95% confidence intervals for the correlations of the ERS index with sex, age, and SRERS. For binary target variables, the frequency of each response category and the ratio of "Yes" responses are reported. For metric target variables, we show the minimum, maximum, mean, and median values, as well as the standard deviation. We further provide the correlations of the target variables with the demographic variables, the ERS measures, and the sum scores of the IMP and ORD scale from part A of the questionnaire. To ease the interpretation of our results, target variables were coded so that a positive correlation could be expected with the sum score of the respective primary NEO-PI-R scale. All reported correlations are Pearson product-moment coefficients.

Several PC trees were computed as part of the preliminary analyses: To ensure the presence of ERS, we fitted PC trees to the IMP and ORD scales from part A, with the ERS index as the only covariate. As a manipulation check for the ERS instructions, separate PC trees were computed for the IMP and ORD scale from part B and C with the type of instruction (mid-responding, extreme-responding, neutral) as single covariate.

As a graphical manipulation check of the administered ERS instructions, we compare histograms of item responses under the extreme and mid-responding instructions to the standard instruction in part A. Note that these plots only contain subjects from both experimental conditions, and ignore which instructions were encountered in part B or C. To investigate possible order effects, another series of histograms compare item responses of the control group for part A, B, and C. If item responses differ between the three parts of the questionnaire in the control group, repeatedly answering the same items could have an effect on the criterion validity of the scale. This might be a confounding factor for the main analyses because in the process of matching ERS instructions to subjects' own

response style tendencies, it cannot be controlled if item responses stem from part B or C. To rule out this possibility, we provide another graphical analysis: In the control group, correlations of the target variables with the sum score of the primary NEO-PI-R scale were contrasted between the three parts. If repeated responses to the same items under neutral instructions leads to better answers, as subjects are more invested to find the most appropriate answer, we expect a clear order in the correlations of the three parts.

In another analysis, PC trees were fitted to both scales under the compensation, aggravation, and control settings. Again, the ERS index was used as the only covariate. For all PC trees, a significance level of 0.05 was used. The minimum number of observations in a tree node was set to the default value, which is equal to ten times the number of estimated parameters. Similar to the previous studies, the results of PC tree analyses are presented graphically.

Finally, we performed two predictive modeling analyses based on the RF algorithm: Negatively keyed IMP and ORD items were recoded. As a monotone transformation, recoding does not have any effect on the fitted RFs. However, it makes the interpretation of the following partial dependence plots more intuitive. For binary target variables, the positive class was chosen so that the predicted probability for the positive class should be higher for high values on the corresponding NEO-PI-R scale.[16]

In each application of the RF algorithm, we tuned the *mtry* parameter on a grid ranging from 1 to 10 with 10-fold CV. The number of trees was fixed to 1500 in all analyses. When estimating the predictive performance of a model on new data, nested resampling was performed. We used 10-fold CV in the outer loop and repeated the whole procedure 3 times, in order to further reduce the variance of the performance estimation. This is sometimes called repeated CV (Bischl et al., 2012). For binary target variables, resampling in the inner and outer loop was stratified, which ensures that the class distribution in each fold is the same as in the whole sample.

For binary target variables, Cohen's κ (Cohen, 1960) was used as the primary performance measure. This means that it was used to choose the optimal values of *mtry* in the hyper-parameter tuning, and to assess the predictive performance estimated by the outer loop of nested resampling. Compared to the commonly used accuracy ($ACC = 1 - MMCE$), κ is better suited to capture the performance of a classification model with a single number: Imagine a binary target variable for which 80% of all observations belong to the positive class. In such an unbalanced case, useful predictions are typically harder to achieve, due to the lack of data from the smaller class. Under these circumstances, algorithms tend to predict the positive class with extremely high probability, rendering the predictive model almost useless. Unfortunately, this behavior is not reflected by the ACC measure. In the current example, a model which exclusively predicts the positive class yields an ACC of 0.8. This shows that interpreting ACC is nearly impossible without additional information. In contrast, κ would be zero for this example, as it takes into account the true ratio of positive observations as well as the ratio of positive predictions made by the predictive model. In this sense, positive κ can only be achieved if the ACC of a predic-

[16] "Yes" was the positive class for all binary target variables except for "Dishes".

tive model is higher than can be expected "by chance". As κ is still a single number, it does not carry the complete information about predictive performance. To further increase the interpretability of our results, we also present MMCE, sensitivity (SENS), and specificity (SPEC) as secondary measures in a table. For metric targets, R^2 was used as the primary performance measure. The MSE and the root mean squared error ($RMSE = \sqrt{MSE}$) are also reported in tables. R^2 was chosen over MSE, as it might be more intuitive to psychologists who are familiar with its interpretation from linear models. To aggregate the 10 times 3 performance estimates resulting from the repeated CV scheme of each predictive modeling exercise, the mean was used for both classification and regression models. Estimated predictive performance of RF analyses will be presented graphically in the form of boxplots. For each predictive model, an observation in the boxplot corresponds to one of the 10 times 3 performance estimates from the outer loops of nested resampling. Binary and metric target variables will be presented in separate plots due to the different primary performance measures.

The first RF analysis contrasts the ability of two different models to predict each target variable. In the first model, all IMP and ORD items from part A of the questionnaire, sex, and age were used as predictors. The second model additionally included the ERS index and the SRERS measure. To evaluate predictive performance, we repeated the described nested resampling scheme for each target variable. For the prediction models which included the two ERS measures, we further estimated a final RF based on the whole dataset and computed the OOB variable importance. Similar to the earlier performance estimation, *mtry* was tuned with 10-fold CV. Variable importance is presented in a separate line plot for each target variable.[17] We also present individual partial dependence plots for this analysis. Partial predictions were computed for the ERS index on an evenly spaced grid of length 30, ranging from the lowest to the highest ERS value observed in the complete dataset. Similar to the variable importance, partial predictions were based on the final models which were fitted to the complete sample. Partial dependence plots show the grid values for the ERS index on the x-axis. Additionally, individual predictions are color labeled based on the responses to the NEO-PI-R item with the highest OOB variable importance for the respective target variable. In chapter 1.4, we noted that ERS might heavily influence trait estimates for subjects with extremely high or low levels of ERS, but exert only moderate bias on average. Comparing predictive performance between models with and without ERS measures highlights the general impact of ERS. In contrast, the individual partial dependence plots can illustrate the impact of ERS for different levels of the ERS index.

The second RF analysis compared the predictive performance for each target variable between the compensation, aggravation, and control settings. For these analyses, matching was performed by the procedure described earlier. We used the same nested resampling scheme as in the previous analysis. However, only the IMP and ORD items were used

[17] Remember that the OOB variable importance does not provide any information about the direction of the covariate effects. Negative importance values result if predictive performance is improved by permuting the respective variable. This can be a sign of overfitting.

as predictors. Demographic variables and ERS measures were omitted. By design, they are unaffected by our experimental conditions and the corresponding matching process. Moreover, ERS measures were already used in the matching process.

In the present study, we did not use ERS measures based on polytomous mixed Rasch models of the IMP and ORD scales. As discussed in chapter 1.3, the mixed Rasch model is a less objective method to detect ERS compared to the ERS index from heterogeneous items and SRERS. Latent classes from mixed Rasch models can capture various factors in addition to ERS. Classes can differ with respect to the latent trait which is measured by the scale of interest or reflect subgroups based on an unobserved covariate. This confounding by content is problematic when using an ERS measure derived from the class membership of a mixed Rasch model as predictor in our predictive models. The RF algorithm is highly flexible to capture complex relations between the predictors and the target variable. In contrast to the ERS index from heterogeneous items and SRERS, the classification based on the mixed Rasch model does not provide strictly new information to the predictive model. The classification is just a complex transformation of the item responses which are already included into the model. Thus, if adding the mixed Rasch ERS measure leads to an increase in predictive performance, the improvement cannot be reliably attributed to ERS.

All statistical analyses were conducted in `R` (R Core Team, 2016b). PC trees together with the corresponding tree plots were computed with the `psychotree` package (Zeileis et al., 2016). All remaining plots were created with the `ggplot2` package (Wickham & Chang, 2016). The predictive modeling analyses were conducted within the infrastructure provided by the `mlr` package (Bischl et al., 2017). The `mlr` package provides a convenient standardized interface to build predictive models in R, which includes all related steps like hyper-parameter tuning, model fitting, assessing predictive performance, and even model visualization (Bischl et al., 2016). To fit the RF models, the algorithm in the `randomForest` package (Breiman, Cutler, Liaw, & Wiener, 2015) was chosen. The unstandardized OOB variable importance was computed with the same package. In addition to those already cited in study 1 and 2, the packages `Hmisc` (Harrell, 2016), `dplyr` (Wickham & Francois, 2016), `parallelMap` (Bischl & Lang, 2015), `rpart` (Therneau, Atkinson, & Ripley, 2015), and `rpart.plot` (Milborrow, 2017) were used.

4.3 Results

4.3.1 Descriptive Statistics

Descriptive statistics of the ERS index from heterogeneous items are presented in Table 4.3. The ERS index was unrelated to sex and age. In the SRERS question, 21% of participants reported a tendency to choose extreme categories. On average, subjects from the extreme SRERS group had higher values on the ERS index. They also were slightly older ($r_{SRERS,Age}$ 95% CI : $[0.03; 0.17]$).

Tables 4.4 and 4.5 show descriptive statistics for the binary and metric target variables.

Table 4.3: Descriptive Statistics of the ERS Index

M	SD	MeanAC	MaxAC	r_{Sex} CI	r_{Age} CI	r_{SRERS} CI
0.50	0.09	0.07	0.33	[-0.13, 0.01]	[-0.08, 0.06]	[0.23, 0.36]

Note. 95% confidence intervals are presented for the correlations of the ERS index with sex, age, and SRERS. Females are coded as one, males are coded as zero. Self-reported extreme responding is coded as one, self-reported mid responding as zero. MeanAC = Mean of absolute correlations between all items within the ERS index. MaxAC = Highest absolute correlation between items within the ERS index. SRERS = self-reported extreme response style.

Table 4.4: Descriptive Statistics of Binary Target Variables

Target	Yes	No	Ratio
Argument	137.00	651.00	0.83
Clothes	193.00	595.00	0.76
Electronics	54.00	734.00	0.93
Lying	263.00	525.00	0.67
Series*	391.00	299.00	0.57
Bed	315.00	473.00	0.60
Document	280.00	508.00	0.64
Dishes	283.00	505.00	0.64
Desk	418.00	370.00	0.53

Note. The Ratio column contains the relative frequency of the majority class.
* In the Series variable, 98 subjects responded that they did not watch any series. This option was ignored for computing the ratio.

Table 4.5: Descriptive Statistics of Metric Target Variables

Target	M	Med	SD	Min	Max
Snacking	3.17	2.00	3.42	0.00	34.00
BMI	22.69	21.89	4.31	15.02	59.52
Vacuum	1.69	1.00	1.84	0.00	14.00

Note. Med = median. Min = minimum, Max = maximum.

The binary targets Argument, Clothes, and Electronics had highly unbalanced distributions. Similarly, the dispersion of the metric variable Vacuum was small as the majority of participants reported zero or one round of vacuum cleaning.

Table 4.6: Target Correlations

	Sex	Age	ERS Index	SRERS	IMP	ORD
Argument	0.10*	-0.10*	0.05	0.03	0.19*	-0.08*
Clothes	0.18*	-0.04	-0.02	-0.02	0.11*	-0.01
Electronics	-0.09*	-0.02	0.02	-0.04	0.08*	-0.10*
Lying	0.01	-0.06	0.04	0.00	0.15*	-0.17*
Series	-0.06	-0.17*	0.03	-0.04	0.17*	-0.12*
Snacking	0.05	0.07	0.00	-0.01	0.17*	-0.06
BMI	-0.17*	0.28*	0.00	-0.02	0.15*	-0.12*
Bed	0.03	0.11*	-0.01	0.06	-0.14*	0.25*
Document	0.06	-0.10*	-0.01	0.08*	-0.15*	0.31*
Dishes	0.00	0.05	0.01	0.01	-0.08*	0.18*
Desk	0.07*	0.10*	-0.07	0.04	-0.11*	0.37*
Vacuum	0.09*	-0.02	0.04	0.11*	-0.10*	0.21*

Note. Correlations of target variables with demographic variables, ERS measures and the sum scores of the Impulsivity (IMP) and Order (ORD) scales. IMP is considered the primary scale for variables Argument to BMI, while Bed to Vacuum are target variables for ORD. SRERS = self-reported extreme response style.
* $p < 0.05$.

We present correlations of target variables with demographic variables, ERS measures and the sum scores of the IMP and ORD scales in Table 4.6. All target variables correlated with their primary scale in the expected direction, although all correlations were of small or medium size. The highest correlation of 0.37 was observed for the relationship between Desk and the sum score of the ORD scale. In most cases, we also observed correlations with the secondary scale. As sum scores of the IMP and ORD scales were negatively correlated ($r_{IMP,ORD}$ 95% CI: $[-0.35; -0.22]$), correlations with the secondary scale were always in the opposite direction. For Lying and Series, correlations with the secondary scale were slightly higher in absolute terms than for the primary scale. The ERS index did not correlate with any of the target variables. SRERS was slightly correlated with Document and Vacuum. Some correlations emerged for the demographic variables. In particular, BMI was related to participants' sex and age.

4.3.2 Presence of ERS

Extreme responding clearly affected item responses of the IMP and ORD scale under standard instructions. Appendix D.4 shows PC trees fitted to item responses of part A of the questionnaire, with the ERS index from heterogeneous items as single covariate. Multiple splits emerged for both scales. The emerging threshold pattern was similar to the previous studies: In leaves which contain subjects with high values on the ERS index, distances between thresholds were smaller and regions of highest probability for the most

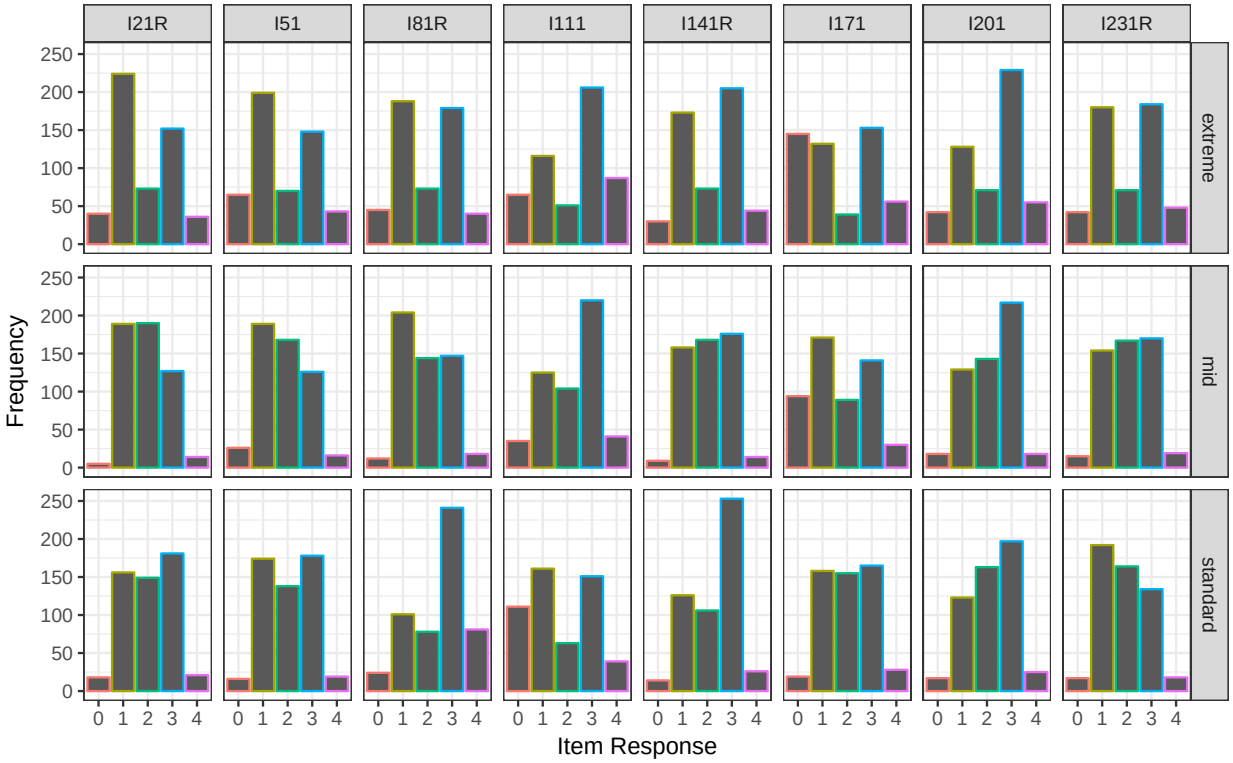

Figure 4.3: Histograms for item responses of the Impulsivity (IMP) scale given under the extreme-responding, mid-responding, and standard instruction. Only subjects in the two experimental conditions were included. Item responses under standard instructions were taken from part A of the questionnaire. Items are labeled with I for Impulsivity, the position of the item in the German NEOI-PI-R, and the letter R if the item has been recoded.

extreme response categories were larger. As illustrated in Appendix D.5, SRERS was also selected as splitting variable when it was used as the single covariate in the PC trees.

4.3.3 ERS Manipulation Checks

The effect of the extreme-responding and mid-responding instructions on participants' responses on the IMP and ORD items is illustrated in Figures 4.3 and 4.4. Under the extreme-responding instruction, the highest and lowest response categories were chosen slightly more often, compared to the mid-responding and standard instructions. At the same time, the midpoint category was chosen less frequently under extreme-responding. For some items, midpoint responses seemed to be shifted to the neighboring categories, however this effect was inconsistent. No clear difference could be observed between the mid-responding and standard instructions. Importantly, under the mid-responding manipulation, the frequency of the midpoint category was not consistently higher compared to the standard instruction. Frequencies of the most extreme categories were also not

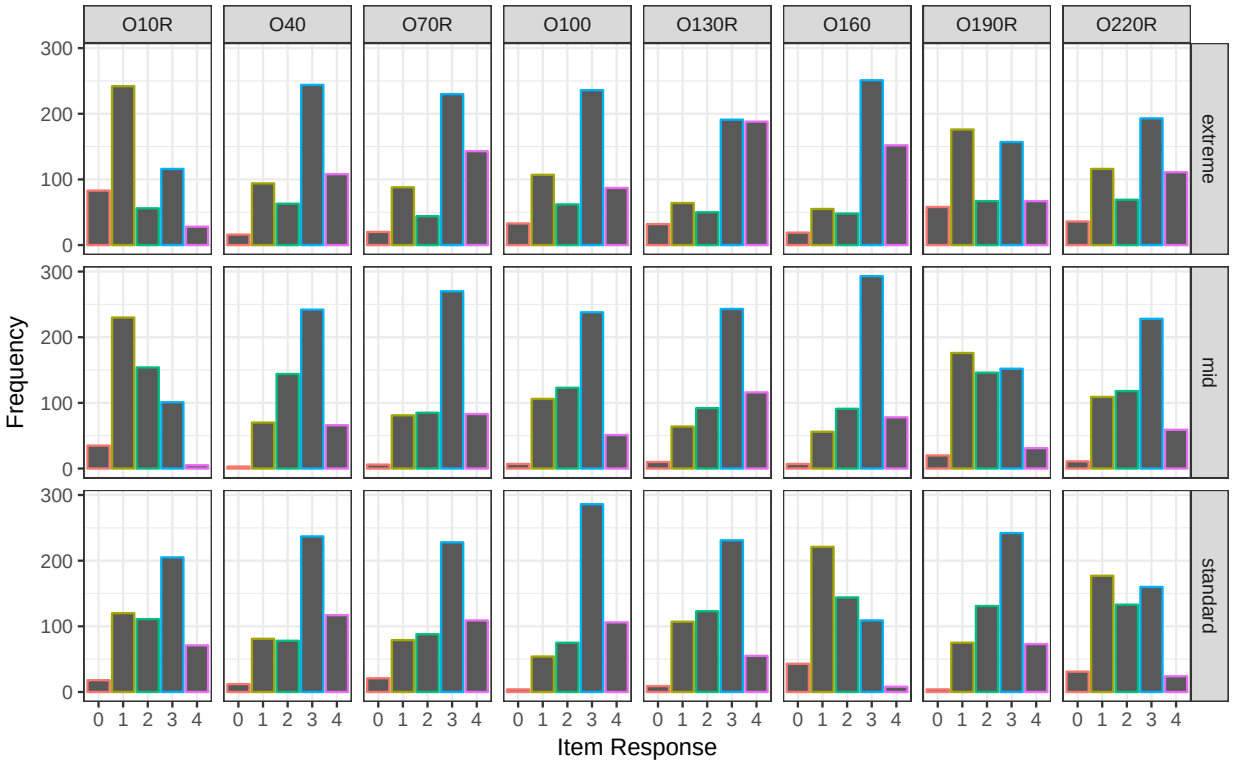

Figure 4.4: Histograms for item responses of the Order (ORD) scale given under the extreme-responding, mid-responding, and standard instruction. Only subjects in the two experimental conditions were included. Item responses under standard instructions were taken from part A of the questionnaire. Items are labeled with O for Order, the position of the item in the German NEOI-PI-R, and the letter R if the item has been recoded.

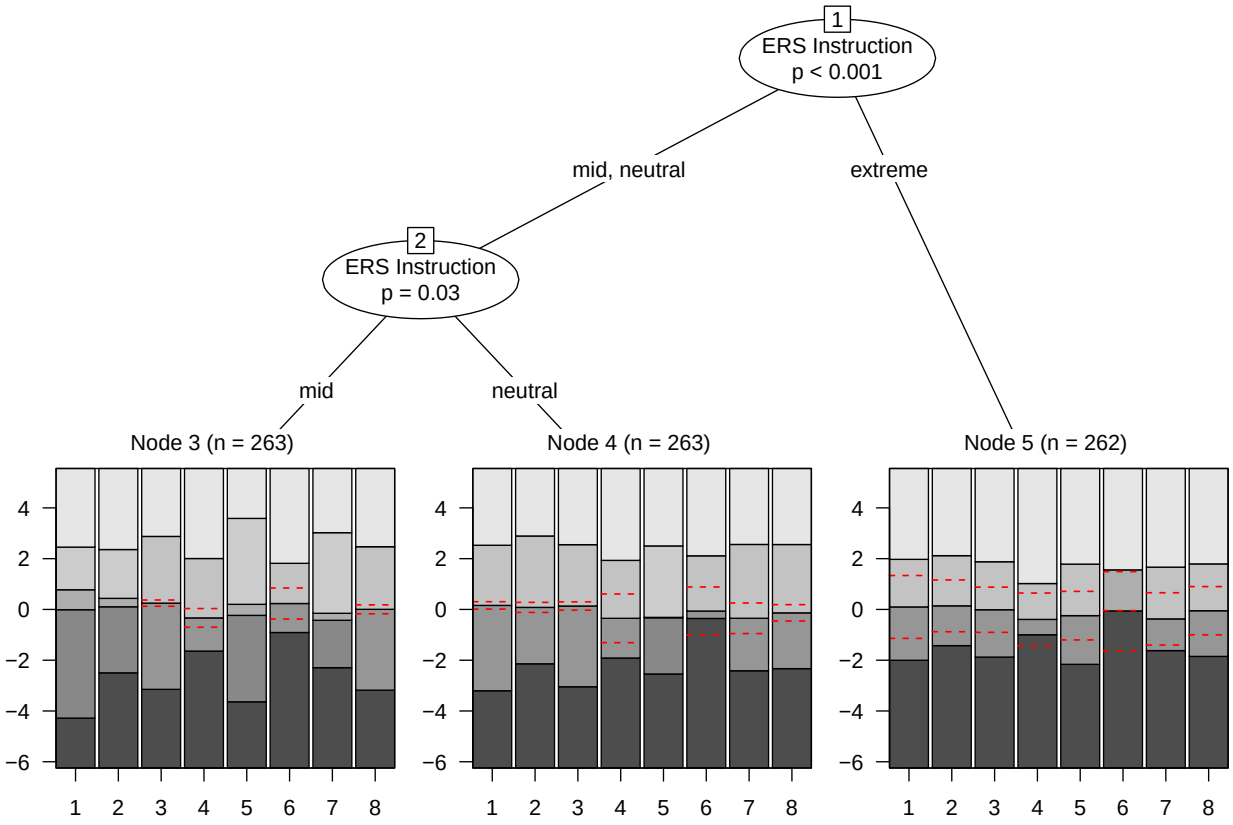

Figure 4.5: PC tree of the Impulsivity (IMP) scale from part B of the questionnaire, with the type of ERS instruction as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

consistently lower under the mid-responding instruction.

PC trees of IMP and ORD item responses from part B of the questionnaire with the ERS instruction as covariate are presented in Figures 4.5 and 4.6. While the IMP scale was split into three leaves defined by all separate ERS instructions, a combined leaf for the mid and neutral-responding condition emerged in the tree of the ORD scale. Appendix D.6 shows the same analyses for part C of the questionnaire. Three splits emerged in both scales from part C. Otherwise, trees of part B and C were highly similar. In leaves defined by the extreme-responding condition, narrow thresholds and large regions for the most extreme categories indicated a more extreme response pattern. Differences in thresholds between the mid-responding and neutral conditions were rather small. Under the mid-responding instruction, the region of highest response probability was more pronounced for the midpoint category.

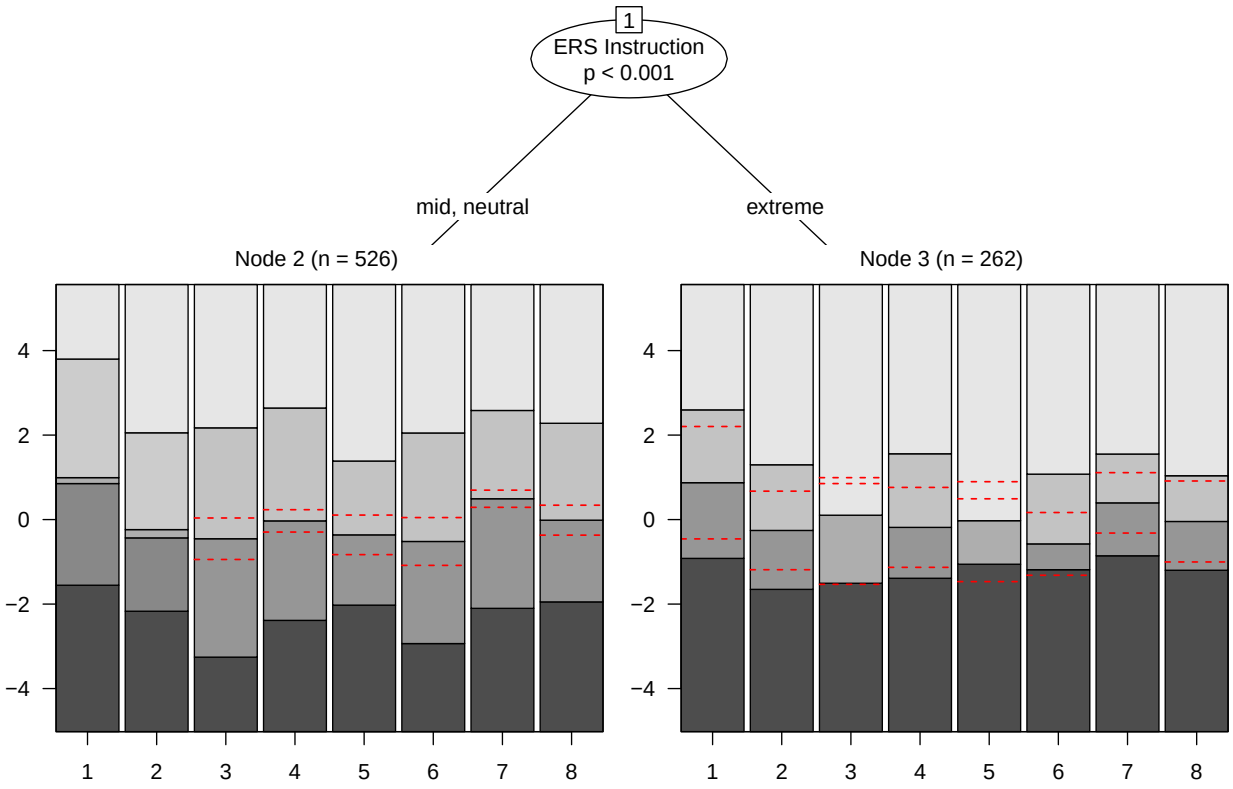

Figure 4.6: PC tree of the Order (ORD) scale from part B of the questionnaire, with the type of ERS instruction as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

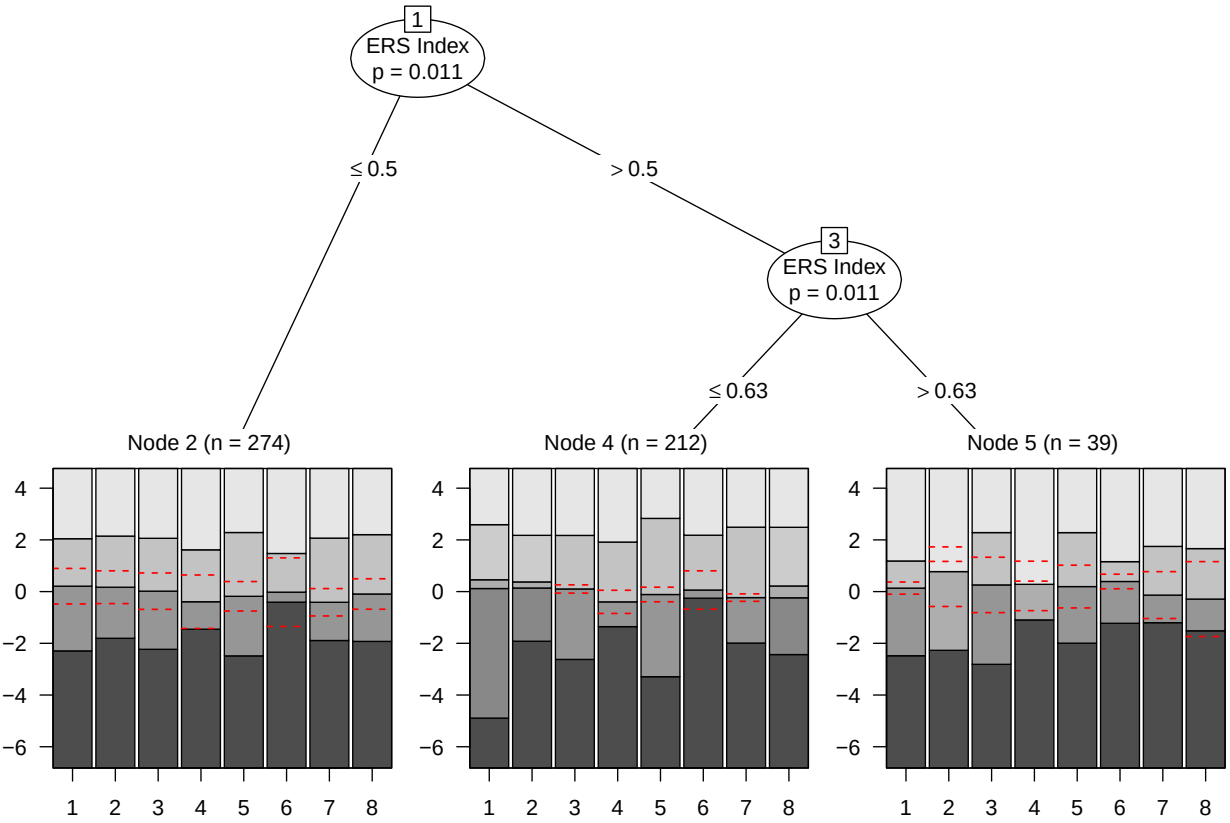

Figure 4.7: PC tree of the Impulsivity (IMP) scale in the compensation setting, with the ERS index from heterogeneous items as single covariate. Matching the ERS instruction to individual response style was based on a median split of the ERS index. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

4.3.4 Potential Order Effects

Comparing item responses in the control group between the three parts of the questionnaires suggested that a small order effect might have affected item response patterns under different ERS instructions (see Appendix D.7). With repeated presentation of the same items, responses became slightly more pronounced, in the sense that midpoint responses were shifted to one of the neighboring categories. Frequencies of the most extreme response categories seemed to be unaffected by the shifting process. However, possible order effects did not influence criterion validity as illustrated in Appendix D.8. Correlations between each target variable and the sum score of the primary scale, computed based on the different parts of the questionnaire in the control group, did not show a consistent ranking.

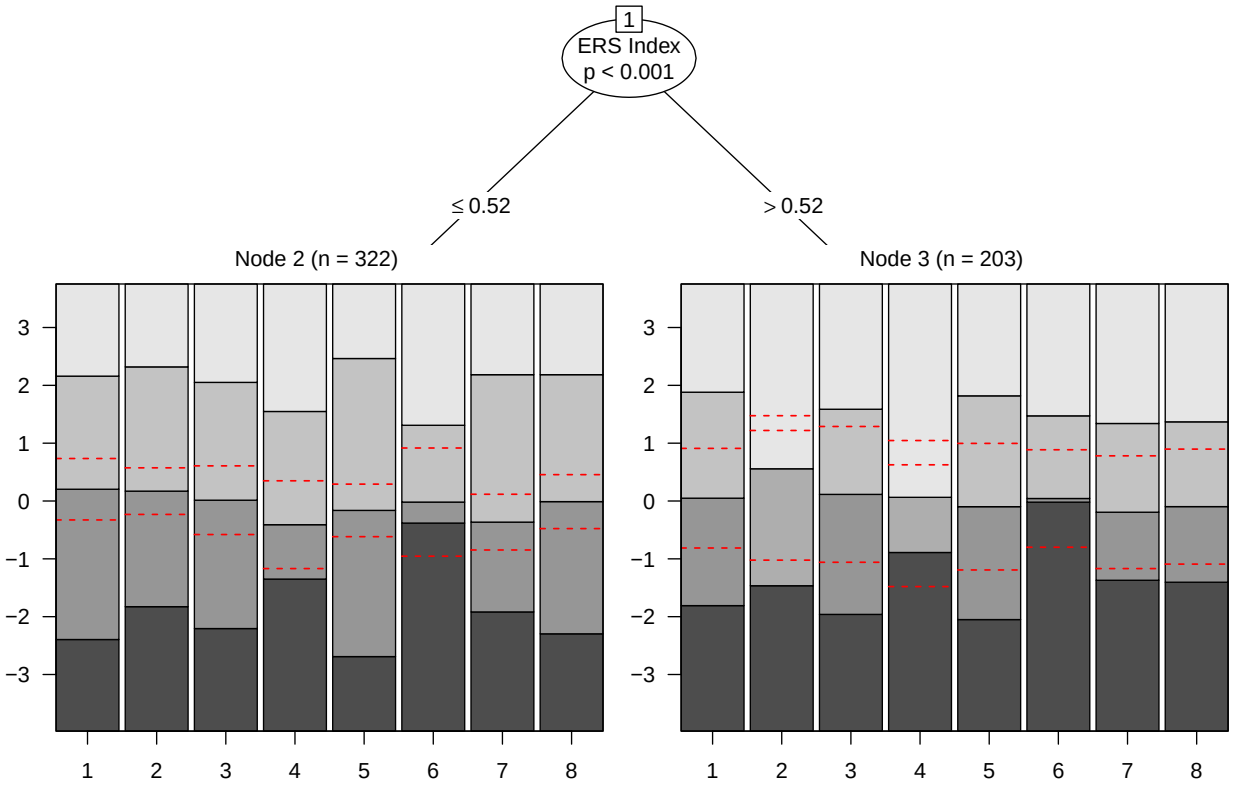

Figure 4.8: PC tree of the Impulsivity (IMP) scale in the compensation setting, with the ERS index from heterogeneous items as single covariate. Matching the ERS instruction to individual response style was based on SRERS. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. SRERS = self-reported extreme response style.

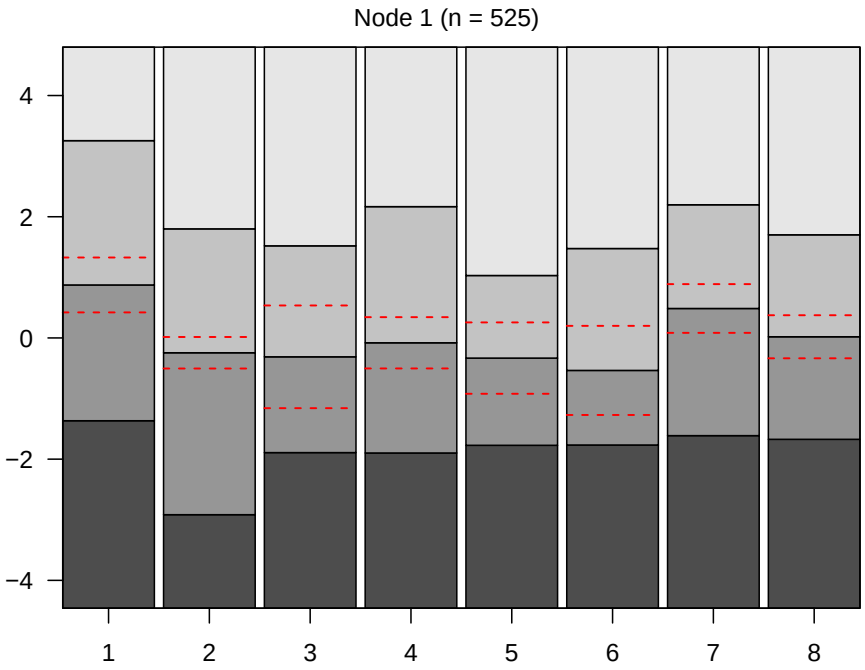

Figure 4.9: PC tree of the Order (ORD) scale in the compensation setting, with the ERS index from heterogeneous items as single covariate. Matching the ERS instruction to individual response style was based on a median split of the ERS index. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

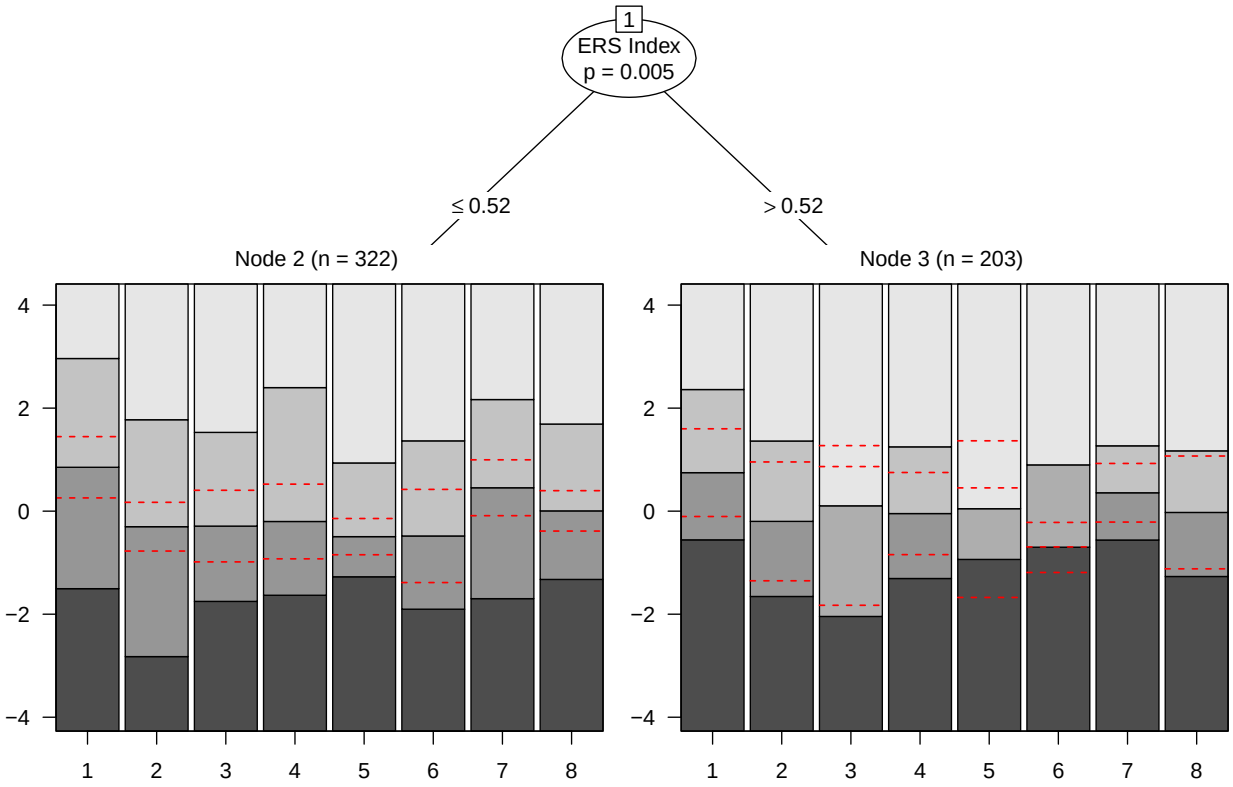

Figure 4.10: PC tree of the Order (ORD) scale in the compensation setting, with the ERS index from heterogeneous items as single covariate. Matching the ERS instruction to individual response style was based on SRERS. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. SRERS = self-reported extreme response style.

4.3.5 Matching ERS Instructions and the Impact of ERS

Figures 4.7 and 4.8 show the PC trees for the IMP scale in the compensation setting. For the first plot, instructions were matched to individual response tendencies based on a median split of the ERS index, while matching was based on SRERS in the second plot. The same analyses for the ORD scale are presented in Figures 4.9 and 4.10. For all analyses in this section, PC trees included only the ERS index from heterogeneous items as covariate.

In spite of the compensatory matching of the ERS instructions, the ERS index was selected as splitting variable in three of the four models. The PC tree of the ORD scale with the median ERS index as matching variable was the only condition in which the ERS index was not selected. Three leaves emerged for the IMP scale, when matching was based on the median split of the ERS index. Compared to the left leaf, the probability for midpoint responses was higher in the mid leaf and the probability for the most extreme categories was lower. Note that the left leaf contained subjects with lower values of the ERS index. If SRERS was used in the matching process, two leaves emerged for both scales. The right leaves, which consisted of subjects with higher values on the ERS index, clearly showed a pattern of more extreme item responses. PC trees in the aggravation setting for both scales and matching variables are presented in Appendix D.9. In the PC tree of the IMP scale with SRERS as matching variable, three splits emerged. In each of the remaining models, the ERS index was selected as splitting variable once. In comparison to the PC trees in the compensation setting, differences in the threshold pattern between leaves were slightly more pronounced in the aggravation setting. The analyses of the control setting are presented in Appendix D.10. For both scales, the ERS index was not selected as splitting variable and a single PC model was estimated for all subjects in the control group.

4.3.6 Contribution of ERS to the Prediction of External Criteria

Figure 4.11 visualizes the estimated performance when predicting the binary target variables with item responses of the IMP and ORD scales from part A of the questionnaire.[18] Predictive performance for metric target variables is presented in Figure 4.12. For both variable types, a table with all used performance measures can be found in Appendix D.11. Predictive performance did not depend on whether the ERS index from heterogeneous items and SRERS were included in the models. In general, estimated performance was higher for ORD targets. Robust above-chance predictions could not be achieved for the IMP target variables Argument, Clothes, Electronics, and Snacking. This was reflected in negative values of κ or R^2. Specificity values very close to one revealed that in the cases of Argument, Clothes, and Electronics, models almost exclusively predicted the majority class.

Figures 4.13 and 4.14 show the estimated variable importance if the predictive models including both ERS measures were fitted to the whole dataset. In line with the perfor-

[18] Only 690 subjects could be included in the predictive model of the Series variable. The remaining participants reported that they do not watch any TV series in general.

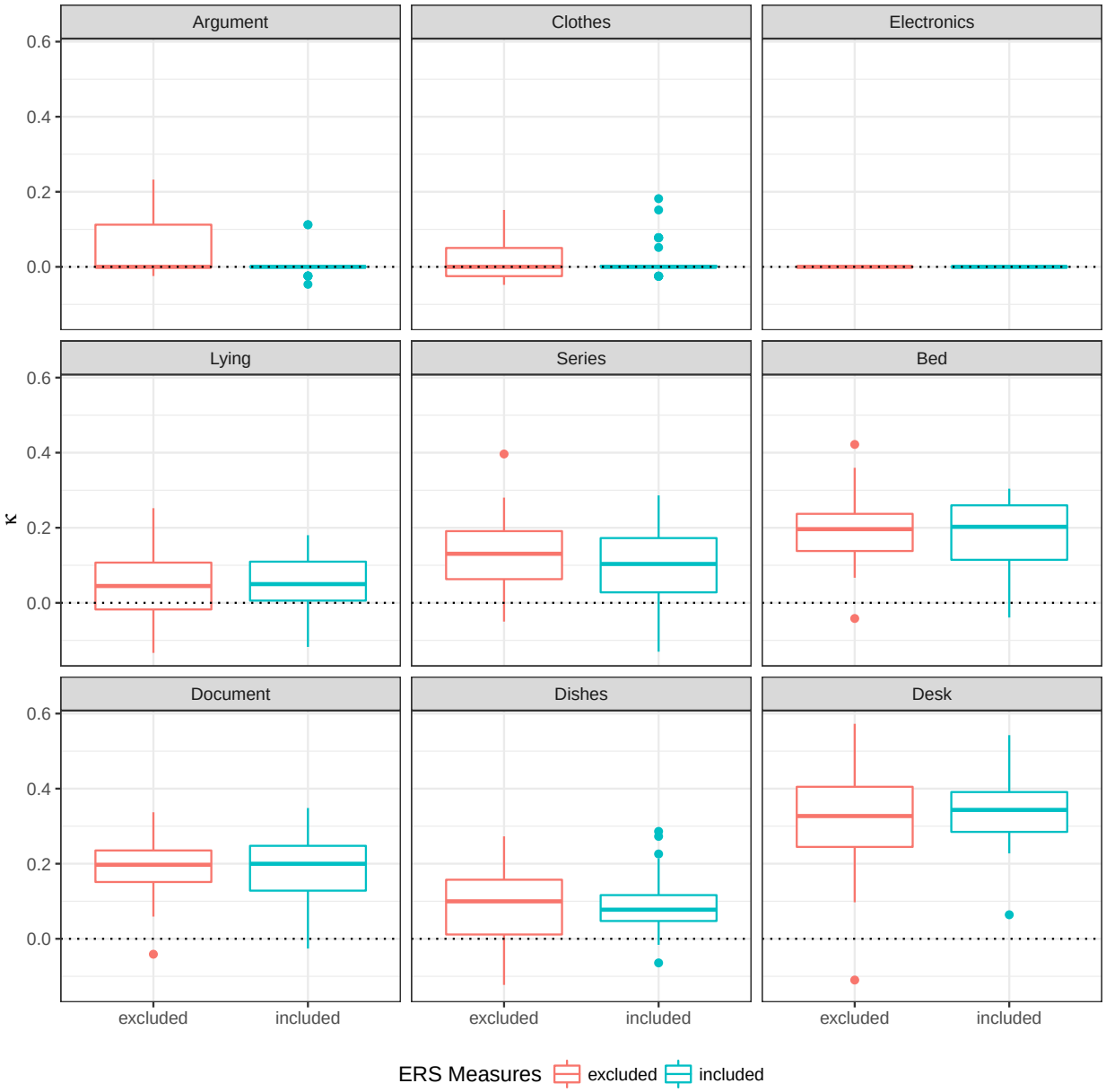

Figure 4.11: Estimated performance for predicting binary target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale from part A of the questionnaire, sex, age, the ERS index from heterogeneous items and SRERS. Predictive models were estimated once with and once without the ERS measures. Performance was measured by Cohen's κ. SRERS = self-reported extreme response style.

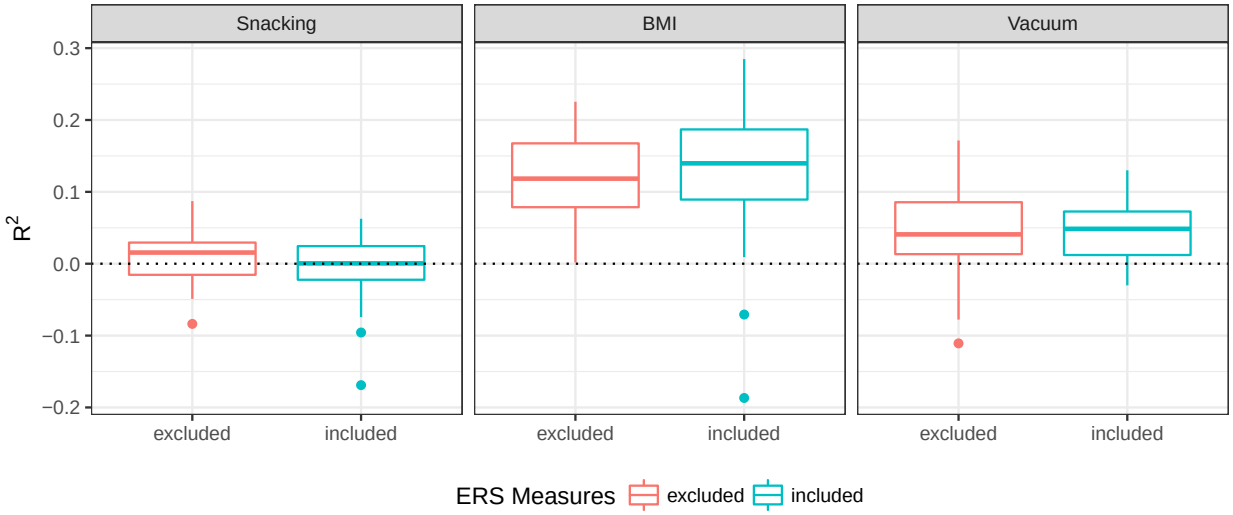

Figure 4.12: Estimated performance for predicting metric target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale from part A of the questionnaire, sex, age, the ERS index from heterogeneous items and SRERS. Predictive models were estimated once with and once without the ERS measures. Performance was measured by R^2. SRERS = self-reported extreme response style, R^2 = coefficient of determination.

mance estimates, both ERS measures mostly ended up in the bottom half of the variable importance ranking. The ERS index was the most important predictor for Electronics, and the sixth most important one for Desk. Sex and age seemed to have an effect on some target variables, especially BMI. In most cases, the NEO-PI-R item with the highest variable importance did belong to the primary scale of the respective target variable. The only exceptions were Snacking and BMI. The items with the highest variable importance differed between target variables. Worth mentioning is the item O40 ("I keep my belongings neat and clean."), which had the highest variable importance for Bed, Document, and Desk.

The effect of the ERS index from heterogeneous items was further investigated with the individual partial prediction plots shown in Figures 4.15 and 4.16. For most binary targets, two distinct clusters of lines emerged. Except for the variable Desk, individual lines rarely crossed the 0.5 probability threshold. Thus, for each subject, the models predict the same class, regardless of the value of the ERS index. Considering the color labeled responses to the NEO-PI-R item with the highest variable importance (see Figure 4.13), the clearest pattern in the ranking of item responses emerged for the variable Desk. Subjects who chose the highest category (violet lines) on the corresponding item O40 were predicted to have a tidy desk with high probability. In contrast, the model assigned small probabilities for subjects who chose the lowest or second lowest response category (orange and yellow lines). To a lesser extent, the same pattern could be observed for the variables Bed and Document. For Bed, Document, and Desk, the shape of individual partial dependence lines seemed to be related to the item responses on the item with the highest variable

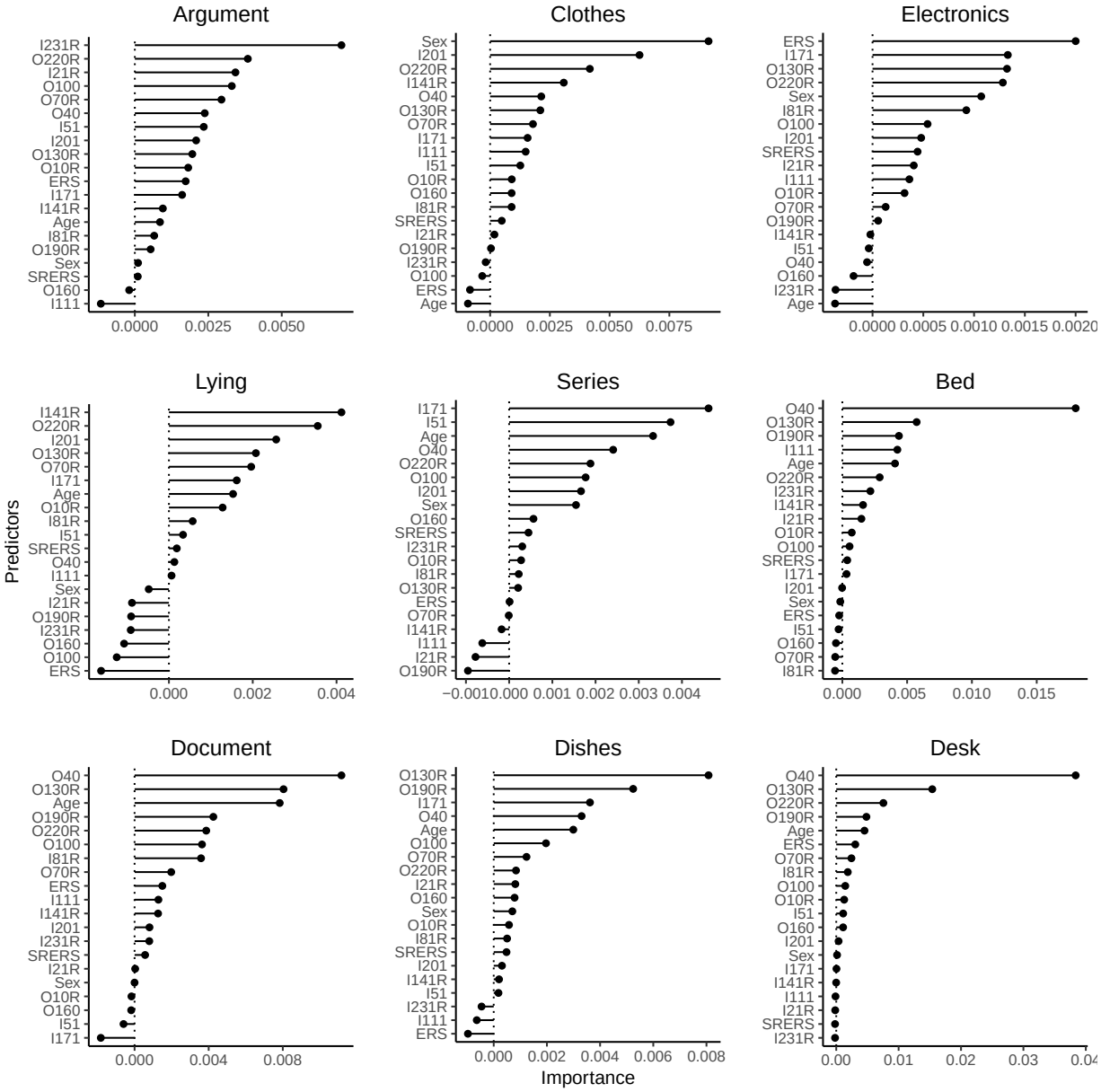

Figure 4.13: Variable importance of predictive models for binary target variables, with item responses of the Impulsivity (IMP) and Order (ORD) scale from part A of the questionnaire, sex, age, the ERS index from heterogeneous items, and SRERS as predictors. Items are labeled with O for Order or I for Impulsivity, the position of the item in the German NEOI-PI-R, and the letter R if the item has been recoded. SRERS = self-reported extreme response style.

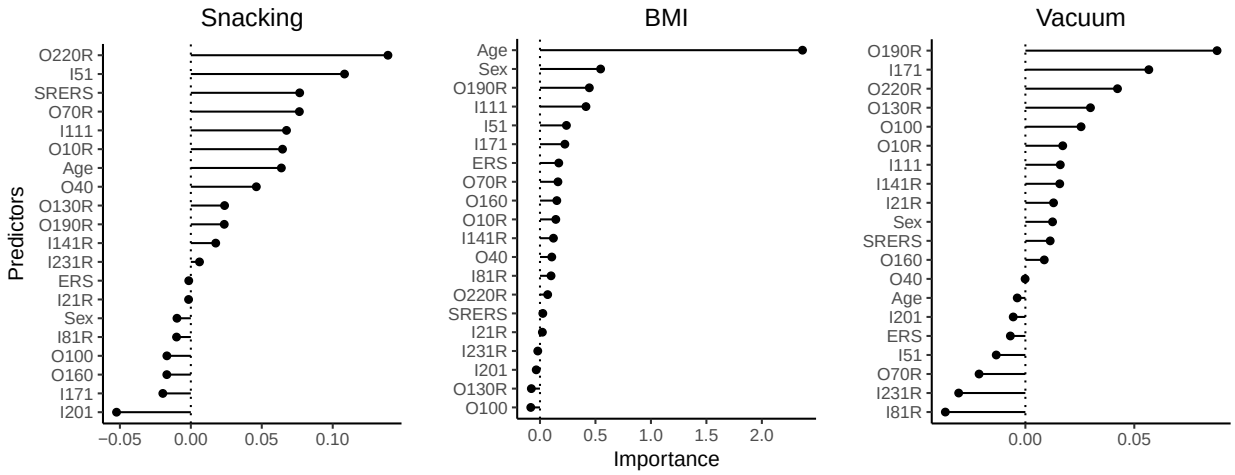

Figure 4.14: Variable importance of predictive models for metric target variables, with item responses of the Impulsivity (IMP) and Order (ORD) scale from part A of the questionnaire, sex, age, the ERS index from heterogeneous items, and SRERS as predictors. Items are labeled with O for Order or I for Impulsivity, the position of the item in the German NEOI-PI-R, and the letter R if the item has been recoded. SRERS = self-reported extreme response style.

importance, which was the item O40 in all three cases. The relationship was U-shaped for lower item responses and inversely U-shaped for higher item responses. For the variable Desk, a sharp decrease in target probability can be observed around an ERS index of about 0.57. However, this phenomenon mostly appeared for subjects who chose the highest or second highest category on item O40 (violet and blue lines). It could not be observed for subjects who chose the lowest of second lowest category (orange and yellow lines).

4.3.7 Matching ERS Instructions and Predictive Performance

We estimated the performance of predicting target variables with item responses of the IMP and ORD scales under the compensation, aggravation, and control settings. ERS instructions were matched to individual response tendencies based on a median split of the ERS index from heterogeneous items. Results for binary and metric target variables are visualized in Figures 4.17 and 4.18. Tables with additional performance measures are presented in Appendix D.12. We repeated the analysis under the condition that ERS instructions were matched to individual response tendencies based on SRERS. The corresponding performance estimates are presented graphically in Appendix D.13 and additional performance measures are reported in Appendix D.14. Both analyses are based on 525 subjects in the experimental settings, and 263 subjects in the control setting.[19]

[19] For the target Series, only 464 subjects could be analyzed in the compensation and aggravation settings, and 226 subjects were analyzed in the control setting. The remaining participants in each setting reported that they do not watch any TV series in general.

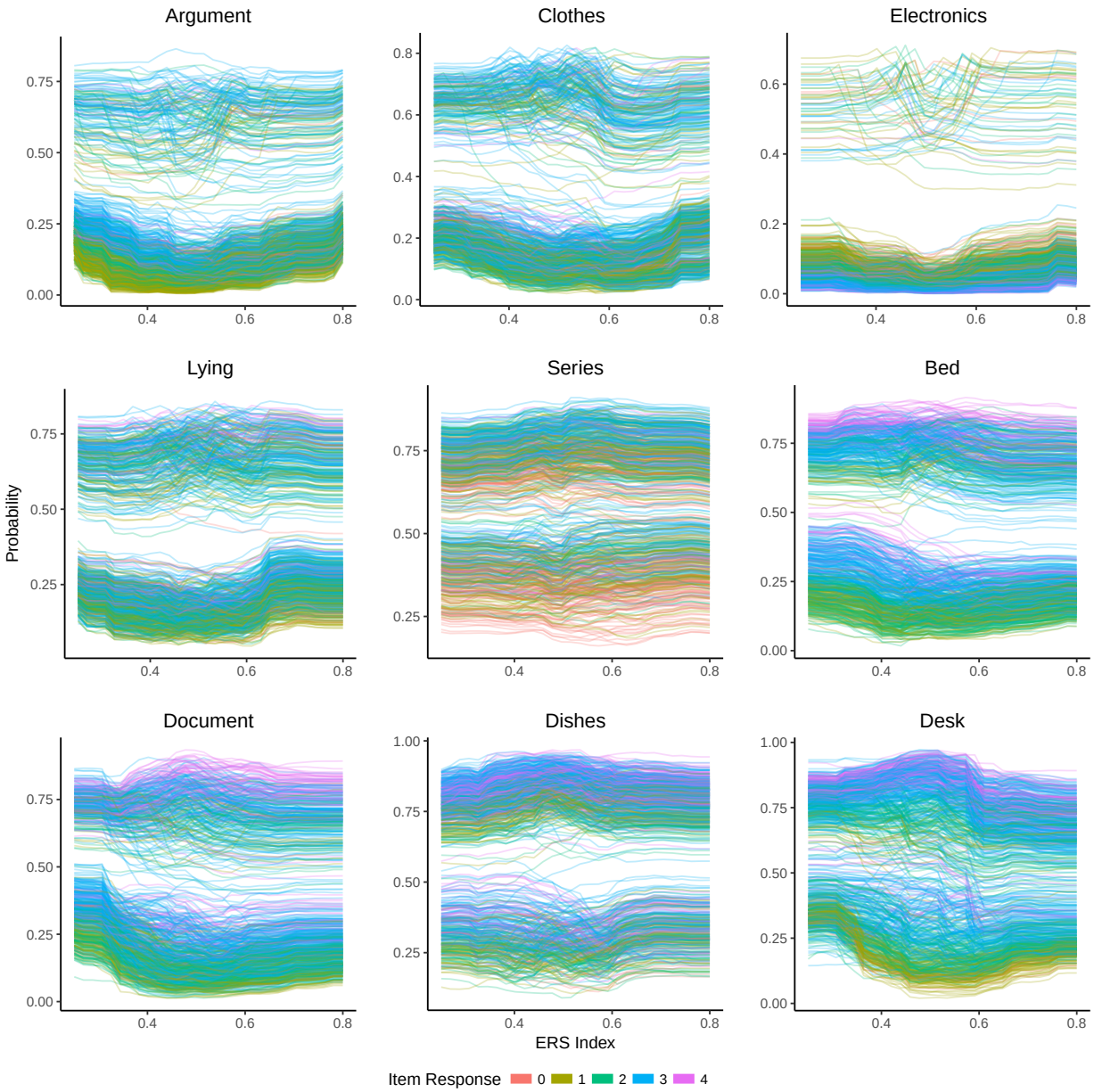

Figure 4.15: Individual partial dependence plots of predictive models for binary target variables, with item responses of the Impulsivity (IMP) and Order (ORD) scale from part A of the questionnaire, sex, age, the ERS index from heterogeneous items, and SRERS as predictors. The ERS index is plotted against the predicted probability to belong to the positive class. Individual predictions are color labeled based on the subjects' responses to the NEO-PI-R item with the highest variable importance (see Figure 4.13). Negatively keyed items were recoded. SRERS = self-reported extreme response style.

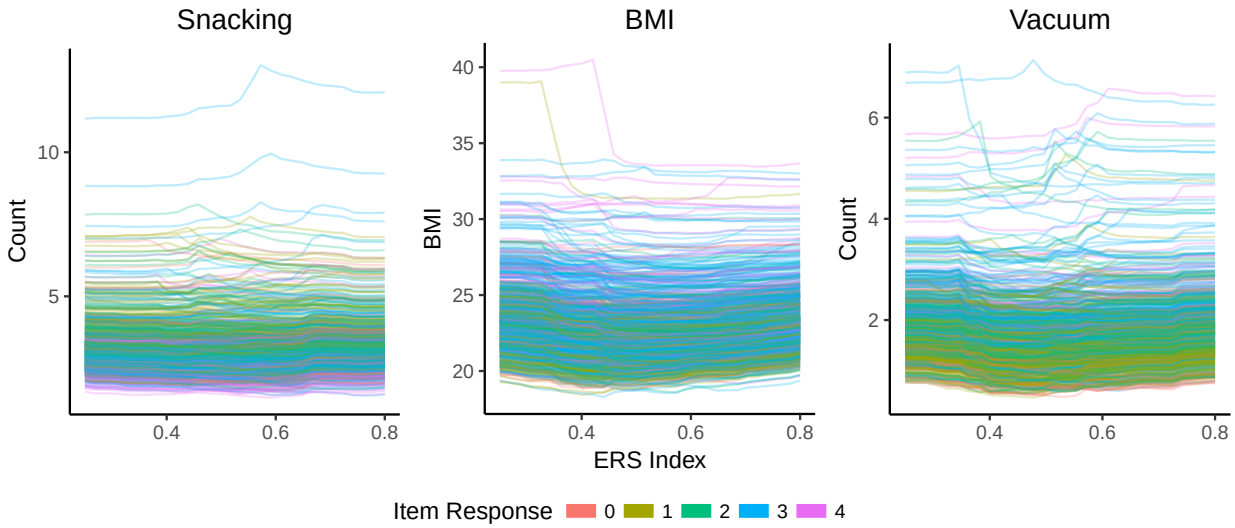

Figure 4.16: Individual partial dependence plots of predictive models for metric target variables, with item responses of the Impulsivity (IMP) and Order (ORD) scale from part A of the questionnaire, sex, age, the ERS index from heterogeneous items, and SRERS as predictors. The ERS index is plotted against the targets predictions. Individual predictions are color labeled based on the subjects' responses to the NEO-PI-R item with the highest variable importance (see Figure 4.14). Negatively keyed items were recoded. SRERS = self-reported extreme response style.

No robust differences in predictive performance emerged between the compensation and aggravation settings. If small differences could be observed in favor of the compensation setting, performance in the control setting was still comparable or even higher in spite of the smaller sample size. The variability of performance estimates was slightly higher in the control setting for most target variables.

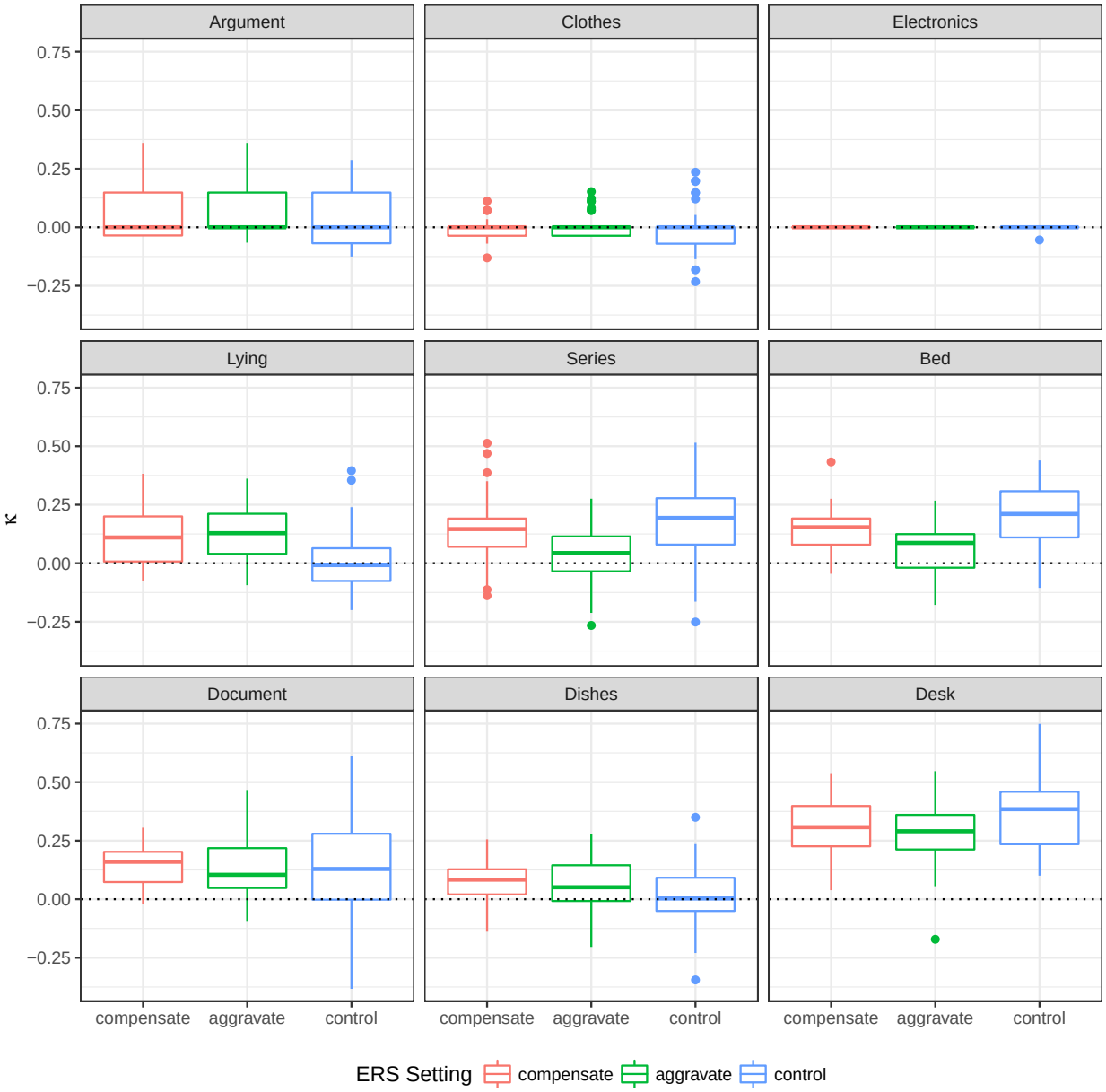

Figure 4.17: Estimated performance for predicting binary target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale under the compensation, aggravation, and control settings. ERS instructions were matched to individual response tendencies based on a median split of the ERS index from heterogeneous items. Performance was measured by Cohen's κ.

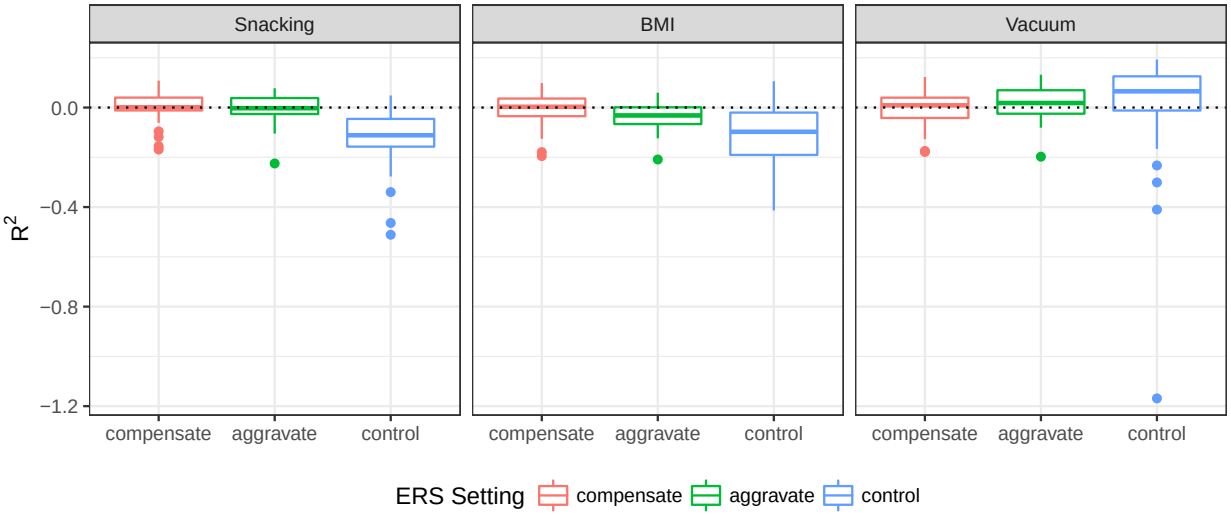

Figure 4.18: Estimated performance for predicting metric target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale under the compensation, aggravation, and control settings. ERS instructions were matched to individual response tendencies based on a median split of the ERS index from heterogeneous items. Performance was measured by R^2. R^2 = coefficient of determination.

4.4 Discussion

4.4.1 Summary of Results

Descriptive statistics revealed extremely unbalanced response distributions for three binary target variables, and low dispersion for one of the metric targets. We found small but consistent correlations of all target variables with their primary NEO-PI-R scale. The characteristics of the ERS index from heterogeneous items was highly similar to the previous studies. The ERS index was uncorrelated with all demographic and target variables. Under standard instructions, response style patterns were detected by PC trees with the ERS index or SRERS as covariates. An effect of the ERS instruction on the general extremeness of item responses could be identified with PC tree analyses with type of instruction as covariate. This was also shown by the histograms of item responses given under different instructions. A comparison of item responses in the control group between the three parts of the questionnaire revealed a small order effect, with less midpoint responses after repeated presentation of the same items. However, correlations between the target variables and their primary scale did not vary systematically. PC tree analyses of the compensation, aggravation, and control settings provided mixed results. Although PC trees suggested that ERS still had an effect on item responses when ERS instructions were chosen to compensate individual response tendencies, the impact seemed to be slightly smaller compared to the aggravation setting. None of the two ERS measures could detect any signs of ERS in the control setting. Target predictions based on demographic variables and item responses from part A of the questionnaire could not be improved by including ERS measures as predictors in the RF models. In general, predictive performance was higher for ORD targets and target variables with balanced response distributions. Variable importance estimates and individual partial dependence plots also suggested negligible contributions of both ERS measures. When ERS instructions were matched to individual response tendencies, predictive performance did not differ between the compensation and the aggravation setting. This result did not depend on whether a median split of the ERS index or SRERS was used in the matching process.

4.4.2 No Impact of Extreme Response Style on Criterion Validity

To evaluate the effectiveness of the developed ERS instructions in reducing the impact of ERS, a necessary prerequisite was that ERS measures can be used to improve the criterion validity of psychological scales. However, estimated performance of predictive models based on demographic variables and item responses under standard instructions did not increase when we added the ERS index from heterogeneous items and SRERS as predictors. Although item responses in the control group revealed a small order effect of repeatedly responding to the same items, the relationship between sum scores and target variables was unaffected.

Our findings are in line with the simulation studies by Plieninger (2016) and Wetzel, Böhnke, and Rose (2016), which suggest that the impact of ERS on criterion validity is

negligible under most circumstances. Both simulation studies assume that apart from the impact of ERS, item responses reflect a homogeneous trait. This assumption seems to be violated in our study. Estimated variable importance suggested that the relationship between NEO-PI-R items and target variables was not equally strong for all IMP and ORD items. In case of a homogeneous trait, the item with the highest reliability should have the highest variable importance, regardless of the criterion variable. The most important item was not always the same for all primary targets of the ORD and IMP scale. This suggests that variable importance was also determined by the item content. It makes intuitive sense that items I141R ("I seldom give in to my impulses.")[20] and I201 ("Sometimes I do things on impulse that I later regret.")[21] were more important for Lying, while the items I111 ("When I am having my favorite foods, I tend to eat to much.")[22] and I171 ("I sometimes eat myself sick")[23] had a higher rank for BMI. However, these findings should be interpreted carefully, as some important items had a high rank for the majority of target variables and not all variable importance estimates were equally intuitive: As already mentioned, item O40 ("I keep my belongings neat and clean.")[24] was the most important item for the ORD targets Bed, Document, and Desk. In contrast, for the target variables Dishes and Vacuum, the item O190R ("I'm not compulsive about cleaning.")[25] was more important. Less intuitive, the most important item for BMI from both scales was O190R, and the most important item for Series was I171.

All predictive models achieved unsatisfying performance for the unbalanced binary target variables Argument, Clothes, and Electronics. In these cases, sensitivity and specificity revealed that RF models almost exclusively predicted the majority class. A series of techniques like sampling methods, cost-sensitive classification, or threshold tuning are available to deal with this common problem. For an introduction of these methods, we refer the interested reader to chapter 16 in Kuhn and Johnson (2013). We did not use any of this methodology to keep our analyses accessible to our audience in psychology, which is generally not familiar with these methods. Yet, this decision probably did not affect any of our main conclusions. For the majority of target variables, distributions were highly balanced and κ indicated rather moderate predictive performance beyond chance-level. One could argue that predictive performance in general could be improved if no self-report items were used to measure target variables. However, target items were designed to be as objective as possible, which was confirmed by cognitive surveys in the pretest. Thus, we do not expect results to differ if behavioral observations are used instead of self-reports.

The most important limitation of our study is the relatively small sample size. It is a well known heuristic in predictive modeling that increasing the amount of data is often the most effective way to improve predictive performance (Yarkoni & Westfall, 2017). This is especially relevant for our research questions, which postulate a complex relationship

[20] German wording: "Ich gebe selten meinen spontanen Gefühlen nach."
[21] German wording: "Manchmal handle ich aus einem spontanem Gefühl heraus und bereue es später."
[22] German wording: "Ich esse meist zu viel von meinen Lieblingsspeisen."
[23] German wording: "Manchmal esse ich, bis mir schlecht wird."
[24] German wording: "Ich halte meine Sache ordentlich und sauber."
[25] German wording: "Ich bin beim Putzen nicht pingelig."

between predictors and target variables. In line with the theory of extreme responding, descriptive correlations suggested that ERS measures did not have a main effect on the target responses. In order to reflect the hypothesized effects of extreme responding, interactions between the ERS measures and item responses have to be modeled. As discussed in the introduction, the RF like many successful predictive modeling algorithms performs implicit regularization to avoid overfitting. Naturally, if the amount of data is small, higher order interactions cannot be reliably detected and complexity is reduced by mostly incorporating main effects into the predictive model. The expected effect of ERS on predictive performance might be detectable with bigger samples.

One reason for this assumption are the individual partial dependence plots for variables with higher predictive performance (Bed, Document, and Desk), which might suggest that the ERS index from heterogeneous items did play a small role in the predictive models. For the target variable Desk in particular, individual partial predictions did not have the same shape. This means that some interactions between the ERS index and other predictor variables were captured (Goldstein et al., 2015). For subjects with a high response on item O40, the probability to belong to the tidy desk class sharply decreased with an ERS index above 0.58. This observation would be in line with the theory on ERS: with a strong tendency to give extreme responses, a high response on item O40 might reflect only moderate ORD levels. Consequently, the probability to show tidy behavior should be lower compared to a subject with the same item response, but a lower value on the ERS index. The partial dependence plots also suggest that even if we captured some influence of the ERS index, the modeled effect was very weak. For all binary target variables, two separated clusters could be observed with respect to the target probability on the y-axis. For the majority of individuals, the same class was predicted by the model, regardless of any hypothetical ERS values. This explains why predictive performance estimates were indistinguishable between models with and without ERS measures, even if a small effect of ERS might have been captured.

However, the general U-shaped effect for subjects with low item responses on O40 and the inverse U-shape for subjects with high item responses is not in line with the theory of ERS. For low values of the ERS index, the probability to belong to the tidy desk class decreases for high O40 responders and increases for low O40 responders. The theory on ERS would have suggested the opposite effect, as extreme item responses should have a stronger impact for subjects, who tend to give extreme responses regardless of their true level on the latent construct. We can think about two possible explanations for this unexpected result: The U and inverse U-shapes might be a reflection of some implicit regularization of the RF algorithm to avoid overfitting. The ERS index was symmetrically distributed around 0.5. For regions of uncommonly low or high ERS values, model predictions might become more conservative and are shifted closer to a probability of 0.5 as data with extreme ERS values is sparse. However, the probability shift already sets in at ERS values close to 0.5. Even when taking the distribution of item responses on O40 into account, enough observations are still available for ERS values around 0.4 and 0.6. It is therefore rather unlikely that the observed effect is purely an artifact. Another explanation might be that instead of treating ERS as a single construct, MRS has to be modeled separately. In this view, values of the

ERS index above 0.5 should be mainly related to extreme responding and the shift in class probability at ERS values around 0.58 might reflect the theoretically expected relationship. In contrast, very low values of the ERS index would be related to a high tendency of MRS that might be qualitatively different than just the absence of extreme response tendencies. Midpoint responses can be a sign of indifference and are sometimes considered an indicator of bad response quality (MacKenzie & Podsakoff, 2012; Kulas & Stachowski, 2009). Thus, it would be more difficult to make accurate behavioral predictions for high MRS subjects if their item responses carry less information about the associated constructs of interest. This higher uncertainty might be reflected in a class probability closer to 0.5 for low values of the ERS index. Constructing a separate MRS index from heterogeneous items which only captures midpoint responses might be a useful endeavor in the future. Including this index into the predictive models and investigating the corresponding partial prediction plots could contribute to the discussion of the dimensionality of ERS which was mentioned in the general introduction (see chapter 1.1).

To argue for the importance of considering ERS in practical applications, it has to be shown that correcting for ERS does increase criterion validity (Wetzel, Böhnke, & Rose, 2016). This was not the case in our analyses. Nonetheless, investigating whether ERS measures can improve the criterion validity of psychological scales is a worthy topic for further research. Large datasets from online panels could be used for this purpose, like we did in study 1. In such datasets, a large number of heterogeneous items are often available. The algorithm used in study 1 and described in Appendix A seems to be a reasonable approach to find a set of uncorrelated items that can be used to compute an ERS index. Analyzing available panel data with the techniques from predictive modeling introduced here might give deeper insights into the impact of ERS on criterion validity in applied work. Thus far, convincing empirical evidence is still missing in the literature.

4.4.3 Instructions Were Ineffective in Reducing the Impact of Extreme Response Style

Our main analysis revealed that matching ERS instructions to compensate individual response tendencies did not lead to improved predictive performance compared to when ERS instructions were chosen to further aggravate the impact of ERS. Considering that our predictive models were unable to achieve higher predictive performance if ERS measures were included under standard instructions, it is not surprising that we did not find any effect on predictive performance between the compensation, aggravation, and control settings. PC tree analyses of the different settings with the ERS index from heterogeneous items as covariate also suggested that compensating ERS with varying instructions was not successful, as we repeatedly detected parameter instability with the ERS index. Although the comparison of PC trees between the compensation and aggravation settings suggested a slightly stronger impact of ERS in the aggravation setting, this difference was small and only reflected in the threshold patterns, while the number of splits was comparable between both settings. The PC tree of the ORD scale with the median ERS index as matching

variable was the only condition in which the ERS index was not chosen as splitting variable in the compensation setting. We refrain from placing any emphasis on this finding as it might be a result of the generally low stability of tree models (Strobl et al., 2013). Indeed, differences in extreme responding for the ORD scale are again observed if SRERS is used as matching variable. For both scales, the ERS index was not used as splitting variable in the control setting, which makes the interpretation of those PC trees even more complicated. Not detecting DIF in the control setting was probably an effect of the smaller sample size compared to the whole sample. In the complete sample, we clearly detected an effect of ERS under standard instructions.

It cannot be assumed that the different ERS instructions had no effect on item responses. This was clearly refuted by the PC tree analyses of item responses from part B and C of the questionnaire, with the type of instruction as covariate. Threshold patterns in these trees revealed that different ERS instructions influenced participants item responses in line with our expectations. PC trees and histograms suggested that the item responses under the extreme-responding instruction differed more strongly from the standard instruction than the mid-responding manipulation. As a consequence, the mid-responding instruction might have been too weak to compensate extremely high levels of ERS. At the same time, subjects with low levels of ERS might have overcompensated when confronted with the extreme-responding instruction. Both arguments might explain the curious finding from the PC tree of the IMP scale in the compensation setting, in which subjects with low ERS values showed a more extreme response pattern than participants with medium values on the ERS index.

Multiprocess models (Böckenholt, 2012; Böckenholt & Meiser, 2017) could be used to further investigate the effect of ERS instructions on item responses. If the manipulation worked as intended, instruction type should be related to the mid/indifference and intensity process, but not to the content/direction process. In theory, an interaction between instruction type and ERS index could be specified to investigate whether the effect of the manipulation on intensity varies for different levels of ERS.

In hindsight, using binary ERS instructions to compensate a continuous response style poses several difficulties. To mitigate the full impact of ERS, instructions have to be flexible so that participants can automatically adapt the magnitude of their compensatory response to their individual level of ERS. By instructing participants to sometimes adjust their responses if they could not decide between two response categories, we hoped to achieve higher compensation for subjects with extremely low and high values on the ERS index. PC tree results suggest that this was not successful. In particular, subjects with high ERS values did not adequately adjust their responses. Based on studies suggesting that ERS is related to intolerance of ambiguity and decisiveness (Naemi et al., 2009), it is possible that response certainty increases with ERS. In this case, low ERS subjects are often undecided, thus adjusting their responses more strongly than high ERS subjects who mostly feel certain and do not adjust their extreme responses. Also, the specific instruction to choose the neighboring category might prevent subjects to flexibly adjust their item responses.

In future research, different ERS instructions could be tested as procedural remedies

against ERS. The current results suggest that a less authoritative variant might be most promising. Cronbach (1950) already recommended to reduce the impact of response bias by increasing participants' testwiseness. It would be interesting to investigate whether an intensive survey training in which respondents' extreme response tendencies are repeatedly monitored, could lead to item responses which are not affected by ERS according to PC tree analyses. Although an extensive training of survey respondents is impractical for applied psychological research, a compact version might be feasible. Our analyses with the SRERS measure in both studies 2 and 3 showed that when asked explicitly, subjects without training already have some valid insight into their own response tendencies. This knowledge could be used to develop modified instructions which take extreme responding into account but do not explicitly instruct participants how to adjust their responses: In a first step, respondents could be informed about extreme responding and confronted with a self-report measure of ERS in order to raise awareness on their own response tendencies. Then, subjects could be simply requested to keep their own response style in mind, without giving specific instructions that depend on self-reported ERS. By giving no explicit advice into how and when participants should adjust their responses, overall compensation of ERS might be more effective. Yet, the success of this modified instruction still depends on the assumption that subjects' insight into their own response tendencies does not depend on their own level of ERS. Due to the low cost of such simple instructions, further investigating these issues might be worthwhile.

4.4.4 General Issues in Predictive Modeling Analyses

When psychological measures are used to predict external criteria, an important point not discussed so far is whether scale scores or single items should be used as variables in predictive models. Traditionally, psychologists use sum scores of scales as variables in their statistical models. While this approach is reasonable from a measurement perspective in which all items are supposed to measure a single construct, it might not be ideal if the main goal is predictive performance. Most psychological constructs are based on factor models and the theory of latent variables. In these models, all items are supposed to measure the same construct and can therefore be combined into an aggregated score which contains all available information about the latent variable. The most common aggregated measure is the sum score of all items in the scale, but factor scores from confirmatory factor analysis or person parameter estimates from IRT models are also used. All those statistics rely on a certain measurement model. Aggregated measures are only useful if the measurement model from which they are derived is an appropriate description of the item response data. Therefore, model fit is of utmost importance when the central goal is the measurement of a certain psychological concept. However, when the focus lies on prediction, measurement models can be ignored and aggregating all item responses into a single score might lead to lower predictive performance.

Psychological instruments like personality inventories often consist of several hundred items. With unregularized linear models, which are the standard approach in psychology, including all items as predictors would lead to overfitting and yield bad predictive perfor-

mance on unobserved data. However, modern predictive algorithms, like the RF (Breiman, 2001) are designed to effectively incorporate a large number of predictors without overfitting. There are at least two circumstances in which single items can be expected to yield better predictions. First, when unidimensional measurement models do not hold, items often contain information about multiple constructs which can be related to the predicted criterion in different ways. Those varying contributions can be implemented in the predictive model, if single items are used as predictors instead of an aggregated score. Second, if a construct based on a latent variable model is only weakly related to a certain criterion, the criterion can still be related to the item errors. This is possible even if errors within one construct are uncorrelated, which is demanded by most common measurement models (Cronbach, 1951). Information in the error terms of a latent variable model is completely lost if only the aggregated score is used as predictor.

Chapman et al. (2016) emphasize that a common latent construct is not necessary to make useful predictions. In fact, items in a criterion-oriented measure do not have to be correlated at all. From a theoretical perspective, predictions can be maximized with independent predictors which make unique contributions to the prediction of the criterion. In this study, single items were used in predictive models, in order to investigate the effect of ERS on item responses. Estimated variable importance was higher for items which shared some content with the respective criteria beyond the latent construct of the scale. Although items within each analyzed scale were highly correlated, as assumed by the underlying personality model, unique contributions of single items were possible. We did not compare our predictive model with single items to a model in which only the sum scores of the scales were included, as this was not part of our research questions. However, if predictive performance is the main goal, researchers should routinely use resampling techniques to compare a predictive model with single items against one in which aggregated scale scores are used as predictors.

Predictive performance is often affected by the model class used in the analysis. Due to overfitting, ordinary linear models are not appropriate for a large number of predictors. Regularized linear models (Hastie et al., 2015) are an obvious extension which should be added to the methodological toolkit of every psychological researcher. However, nonlinear relationships might also be present in psychological research. Possibly, some interesting psychological phenomena are too complicated to be explained by more interpretable models (Yarkoni & Westfall, 2017). This requires powerful predictive algorithms which can automatically learn complex nonlinear functions between the predictors and the criterion variable. Popular non-linear model classes are the RF we used here, gradient boosting, support vector machines, and neural networks. Introductions to all of these methods can be found in Hastie et al. (2009) or James et al. (2013). It is generally hard to predict whether nonlinear methods outperform linear models in a certain setting. Therefore, resampling techniques are routinely used to compare models from different classes. Psychological science would profit from experimenting with non-linear models to find out how they can be used to answer important research questions in the field. In this study, we used a non-linear model based on theoretical considerations. With our operationalization of ERS, the response style is supposed to have a non-linear effect on item responses. To investi-

gate the expected effect, a non-linear model was necessary. We did not find any effect of ERS on predictive performance with the RF algorithm but refrained from comparing different model classes. It is possible that other algorithms might have succeeded in using our ERS measures to increase predictive performance. Although some experimentation with a state-of-the-art variant of gradient boosting (Chen & Guestrin, 2016) did not yield a significant increase in predictive performance, this algorithm was also based on decision trees. In another unreported analysis, linear models yielded worse predictive performance. However, these issues were not investigated in a rigorous manner. If predictive performance for other popular algorithms are also insensitive to whether ERS measures are included as predictors, this will put further weight on our reported null effect.

Chapter 5

General Discussion

5.1 Summary of Empirical Studies

In this dissertation, we presented three empirical studies in which we investigated different aspects of ERS. In study 1, we introduced a new method to detect ERS. PC trees (Komboz et al., 2016) with an ERS index from heterogeneous items (Weijters et al., 2008) as covariate combine the elegant threshold interpretation of ERS known from ordinal mixed Rasch models (Rost et al., 1997), with the high objectivity of response style indices from heterogeneous items. The effectiveness of the new method was shown by two analyses of large datasets. First, we analyzed the nonclinical normative sample of the NEO-PI-R (Ostendorf & Angleitner, 2004) and detected ERS in facets which had already been found to be affected by ERS in previous studies. PC trees resulted in a large number of leaves with threshold patterns that could clearly be interpreted as different levels of extreme responding. The observed pattern suggested that ERS is best understood as a continuous trait. In a second analysis of data from the GESIS panel (GESIS, 2015), we found the same pattern of ERS in scales from the PANAS (Krohne et al., 1996) and the FRS (Münzer & Hölscher, 2011). In the GESIS analysis, items for the ERS index were taken from different panel waves, completed several months apart. Thus, our results can be taken as further evidence for the stability of ERS across traits and over time. We discussed recent developments of continuous IRT models of ERS and talked about the challenges of using these models to control for ERS in scales which are plagued by DIF beyond the impact of response styles.

In study 2, we provided the first empirical evidence that items with dichotomous response format are unaffected by ERS. We used three different ERS measures as covariates in Rasch trees (Strobl et al., 2013) and DIF Lasso models (Tutz & Schauberger, 2015) to analyze scales from the MMPI-2 (Engel & Hathaway, 2003) and the FPI (Fahrenberg et al., 2001) in a sample of psychology students: an ERS index from heterogeneous items, a self-report measure of ERS, and an ERS measure based on the classification of an ordinal mixed-Rasch model. We did not find any signs of ERS in the analyzed dichotomous item response data. Yet, control analyses with a NEO-PI-R facet showed that ERS was present in items with ordinal responses format. This emphasized that not finding an effect of ERS

in dichotomous items cannot be attributed to deficiencies of our ERS measures. Furthermore, our analyses using the self-report measure of ERS are the first to reveal that subjects have some conscious knowledge about their own response tendencies. Based on our positive results, we discussed possible advantages and disadvantages of using dichotomous item response formats to avoid the impact of ERS in practice.

In study 3, we tested whether individually targeted questionnaire instructions can be used to compensate for the impact of ERS in items with ordinal response format. We designed specific ERS instructions which urged subjects to give either more or less extreme responses, when they were uncertain about which category to choose. Respondents of our online questionnaire completed two NEO-PI-R scales under standard, extreme, and mid-reponding instructions in a within-subjects design. We created artificial datasets in which ERS instructions were matched to individual response tendencies, based on either a median split of an ERS index from heterogeneous items, or a self-report measure of ERS. The effectiveness of the ERS instructions were investigated in a prediction context, using the RF algorithm (Breiman, 2001) and general principles from the field of predictive modeling. If the instructions reduce the impact of ERS, we expect an increase in the criterion validity of the questionnaire items. A series of behavioral target variables, which are theoretically related to the analyzed NEO-PI-R scales Impulsivity and Order, were included in the questionnaire. The performance of predicting these criteria under the artificial settings in which ERS instructions were matched to compensate individual response tendencies was compared to an aggravated setting and to a control group with neutral instructions. Matching ERS instructions did not compensate for ERS: predictive performance did not differ between the compensation, aggravation, and control settings. Preliminary analyses with PC trees confirmed that ERS was still present in the compensation setting. Another analysis investigated the performance of predictive models with item responses under standard instructions and demographic variables as predictors. Adding the ERS index from heterogeneous items and self-reported ERS to the set of predictors did not increase predictive performance. This might indicate that ERS does not always have a high influence on criterion validity which would be in line with recent simulation studies (Plieninger, 2016; Wetzel, Böhnke, & Rose, 2016). We discussed alternative explanations for our results and suggested possible improvements for ERS instructions to reduce the impact of the response style.

5.2 Future Directions for Research on Response Styles and Test Practice

Response styles have been a flourishing topic in the psychometric literature since the middle of the 20th century. The three studies we reported in this thesis touched many questions which were already discussed by Cronbach (1950). Applying recent methods from IRT, we presented a more rigorous analysis of some topics (study 1), were the first to provide empirical evidence for some long held assumptions (study 2), and used methods from

predictive modeling to raise some new ideas which might push the field into uncharted territory (study 3). In the remainder, we discuss general directions for future research on ERS. We emphasize how theoretical research on ERS might be useful to applied researchers for constructing new psychological tests and applying existing measures more effectively.

5.2.1 Improving the Measurement of Psychological Constructs

ERS seems to be omnipresent whenever items with ordinal Likert scale format are used in psychological research (Vaerenbergh & Thomas, 2013). Theoretically, ERS poses a severe threat to the measurement of latent variables. Trait estimates can only be considered a valid representation of the underlying concept, if a measurement model with appropriate fit can be found. In the presence of ERS, common measurement models from CTT and IRT do not hold.

In study 1, we introduced a new method to detect ERS, which naturally illustrates this issue. By using an ERS index from heterogeneous items as a covariate in PC trees (Komboz et al., 2016), we detected ERS in any scale with ordinal response format that we analyzed. We observed multiple splits based on the ERS index, showing that a single PCM is not an adequate description of the item responses. As PC trees are a global model test for the PCM (Komboz et al., 2016), our analyses also formally rejected the PCM as an appropriate measurement model.

This emphasizes the necessity for more complex measurement models which take ERS into account. In practice, multidimensional IRT models that estimate both the trait and the response style from the item responses of only one scale of interest might be preferred. As discussed in chapter 2.4.3, multidimensional IRT models are very flexible. However, trait estimates can still be confounded by a variety of other influences, especially if the model does not offer a good description of the response style. We argue that exploratory methods which rely on an objective measure of ERS, like an index from heterogeneous items, can be used to investigate the structure of the response style. The results can then inform researchers about the structure of ERS in multidimensional IRT models.

Indeed, PC tree analyses in study 1 gave important insights into what an appropriate measurement model could look like. Our results suggest that ERS can be understood as a continuous trait, which influences the item thresholds in the PCM. This was illustrated by the large number of leafs as well as the smooth relationship between the ERS index and the estimated threshold parameters. In chapter 2.4.4, we concluded that the continuous IRT models by Jin and Wang (2014) and Tutz et al. (2016) might be promising candidates. Future empirical studies on a variety of scales should investigate which model reflects a more appropriate representation of ERS.

It has been noted that in order to find IRT models which reflect item response data in their "natural" form, "deeper insights into response processes and the precise way item properties and human abilities interact are needed" (Van Der Linden & Hambleton, 1997). In fact, most modeling strategies for ERS do not incorporate conceptual ideas about how participants choose between response categories (Zettler et al., 2015). This critique also applies to the continuous ERS models by Jin and Wang (2014) and Tutz et al. (2016).

Both models offer descriptions of item response data infested by ERS, but do not explain the cognitive aspects of the response process which lead to the observed item response pattern. In contrast, multiprocess models (Böckenholt, 2012; Böckenholt & Meiser, 2017; De Boeck & Partchev, 2012) overcome this caveat by modeling the response process more directly. These models put less focus on measurement and could be considered part of mathematical psychology, as they provide a cognitive model for the item response process (Van Der Linden & Hambleton, 1997). It would be interesting to reanalyze the data from study 1 with multiprocess models. NEO-PI-R items are measured on a five-point Likert scale. Thus, the corresponding multiprocess model to study ERS would contain three response processes (Zettler et al., 2015): The indifference process controls the selection of the midpoint category. If the mid category is not chosen, the direction process controls whether agreement or disagreement is expressed and the intensity process reflects whether the more or less extreme category is chosen. To validate the multiprocess model, the ERS index from heterogeneous items could be used as a covariate for each of the three response processes. The ERS index should be negatively related to the indifference process and positively related to the intensity process.

Multiprocess models can also be used to investigate the dimensionality of ERS (see chapter 1.1). In our studies, ERS was operationalized as a unidimensional construct, continuously ranging from a high preference for midpoint responses to a high preference for the most extreme response categories. The IRT models by Jin and Wang (2014) and Tutz et al. (2016) share this conceptualization of ERS. In contrast, other researchers postulate a bidimensional concept and differentiate between ERS and MRS. In study 3, we discussed that the inverse U-shape we observed in the individual partial prediction plots might indicate that item responses are affected by different ERS and MRS processes. The multiprocess model clearly differentiates between preferences for the midpoint category and preferences for the most extreme categories. If person characteristics that have a distinct effect on either the indifference of the intensity process could be found, this might indicate that MRS should be modeled separately from ERS. Moreover, the indifference and intensity processes in the multiprocess model could be restricted to represent the same latent variable. If the restricted model shows worse goodness of fit as the unrestricted model with separate MRS and ERS processes, this would be evidence for the bidimensionality of extreme responding.

Increasing knowledge about response styles should always be reflected in innovations in psychological test practice. Thus, one important goal of response style research is to guide applied researchers in developing new psychological measures, which are more robust against the impact of ERS. We studied multiple procedural remedies against ERS: As shown in study 2, designing questionnaires with dichotomous items might be a reasonable strategy to avoid the impact of ERS altogether. Unfortunately, dichotomous response formats are unpopular for several reasons, although some criticisms seem to be unfounded (see the discussion in chapter 3.4.4). In study 3, we tested a procedural approach to remove the influence of ERS using special instructions targeted at individual response tendencies in polytomous items, but were unable to compensate for ERS. Nevertheless, the ERS instructions indeed had an effect on item responses. It has been argued that the impact of response styles might be especially high when subjects are uncertain about their

item responses (Podsakoff et al., 2011; Cronbach, 1950). ERS instructions in study 3 were designed based on this assumption. Participants were told to adjust their answers only if they could not decide between two categories. The fact that subjects clearly altered their item responses under our instructions implies that they were frequently unsure about their item responses. This raises the important question whether decreasing the ambiguity of questionnaire items might be a useful strategy to reduce ERS in forthcoming psychological tests. Despite many attempts to reduce the impact of ERS, our results suggest that ERS is omnipresent in ordinal response formats. Hence, statistical control might be the only viable option to deal with ERS in items with ordinal response format at this point. We discussed two elegant IRT models which incorporate ERS as a latent person variable that has an influence on item thresholds (Jin & Wang, 2014; Tutz et al., 2016). In theory, such methods seem well suited to provide trait estimates for established psychological measures which are adjusted for ERS. However, using continuous models of response styles that do not depend on another item set to construct an independent ERS measure is problematic, when item responses are plagued by further unaccounted sources of systematic variance. In study 1 and 2, we also detected DIF induced by demographic variables. Removing these confounding effects might be a necessary first step in test construction, before appropriate statistical models can be used to deal with the more complex issue of ERS. In any case, larger samples are necessary to yield satisfying precision for ERS adjusted trait estimates in practical applications.

5.2.2 Increasing the Predictive Power of Psychological Theories

When the focus lies on the accurate measurement of some psychological trait, an appropriate measurement model is required. High criterion validity is still important in these settings. Otherwise, the psychological construct is almost useless in practice. Psychology is supposed to study human experiences and behavior but makes little effort to accurately predict these phenomena (Yarkoni & Westfall, 2017). Psychological science relies heavily on estimating large linear models. Predictive performance is evaluated based on the same data used in the estimation process, and the explained variance is often brought up to make impressive predictive claims (for example Schmidt & Hunter, 1998). However, it is known that this approach overestimates the predictive power of psychological research (Yin & Fan, 2001). If appropriate methods are used to assess predictive performance, psychological effects often decline (Yarkoni & Westfall, 2017). In light of the replication crisis in psychology (Open Science Collaboration, 2015), it has been argued that a heavier focus on predictive methodology might be useful to battle some of the field's problems (Yarkoni & Westfall, 2017; Chapman et al., 2016). Lately, predictive modeling has risen beyond its origin in computer science and statistics. This trend includes studies with high media coverage (Youyou et al., 2015) and special issues about the topic in psychological journals (Harlow & Oswald, 2016). We used predictive modeling in study 3 and gave a short introduction on some important principles (see chapter 4.1.2) in the hope that more psychological research can profit from these methods. In the following, we describe two settings in which predictive models might be useful for psychological research, and discuss

the possible role of response styles in these scenarios:

Psychological Constructs As Criteria Psychological concepts like personality traits are well established beyond the core field of psychology. Consequently, there are applications in which psychological constructs are the target of predictive modeling analyses. In a modern classic, Youyou et al. (2015) present simple linear predictive models for Big Five personality traits. Likes on the social network Facebook serve as predictors in their models. If based on enough likes, predictions for self-reported personality yield comparable accuracy to other-ratings of spouses. In this analysis setting, sum scores of personality questionnaires are used as the criterion in predictive models and treated as a manifest variable. This might be problematic, considering that sum scores are only an estimate of the latent personality trait, which is attenuated by measurement error and other confounding factors like response styles. When the model is trained to predict sum scores as accurately as possible, it also learns to predict those systematic response tendencies. Moreover, if questionnaire measures are additionally used as predictors, models might suggest overly optimistic performance in predicting psychological constructs. As response styles are included in both the predictors and the criterion, performance estimates reflect the ability to predict both response styles and the psychological trait of interest. This phenomenon has been frequently discussed in studies which examine the impact of ERS on the estimated relationship between latent constructs (Plieninger & Meiser, 2014; Böckenholt & Meiser, 2017). The situation should be more severe when using nonlinear predictive models which are well suited to capture the complex relations caused by response styles. In the worst case, predictive models might be really good in predicting the response biases contained in the criterion, but not in predicting the real psychological trait. Thus, when predicting psychological constructs, it seems especially important to adjust questionnaire measures for response styles before using them as criteria in predictive models. This closes the gap to the continuous models of ERS (Jin & Wang, 2014; Tutz et al., 2016), which were discussed in chapter 2.4.4. If DIF introduced by other covariates like demographic variables can be removed through more careful test construction, ERS adjusted trait estimates from these models could be used as criterion in predictive analyses.

In study 3, the inclusion of ERS measures did not increase the predictive performance of predicting self-reports of every day behaviors. These results are in line with recent simulation studies, which claim that the effect of ERS on criterion validity might be low in most practical applications (Plieninger, 2016; Wetzel, Böhnke, & Rose, 2016). Both findings might lead to the conclusion that ERS also plays a minor role when psychological constructs are used as criterion in predictive models. However, predictive performance was only moderate in our analyses and the sample was relatively small, considering the use of complex nonlinear models. It is possible that ERS can be savely ignored if the signal to noise ratio is low and only crude predictor-criterion relations can be implemented by the predictive model. However, Yarkoni and Westfall (2017) and Chapman et al. (2016) raise the possibility that applying predictive methods on large datasets could lead to an increase in the predictive power of psychological theories. At some point, response styles

might become important in improving predictive performance even further.

Psychological Constructs As Predictors In many areas of psychology, the study of latent constructs is driven by the ultimate goal to predict real life outcomes which are important for human society in general. In the first sentence of their famous meta analysis, Schmidt and Hunter (1998) write: "From the point of view of practical value, the most important property of a personnel assessment method is predictive validity: the ability to predict future job performance, job-related learning (such as amount of learning in training and development programs), and other criteria." In this context, psychological constructs are used as predictors in statistical models. Predictive modeling is the most natural methodological approach, as those tools are designed to maximize predictive performance (Yarkoni & Westfall, 2017). In a first step, combining classical linear models with the resampling strategies described in chapter 4.1.2 would give a more realistic estimate of the predictive performance of psychological constructs. As this will probably yield unsatisfying results, there are several directions to proceed further:

Most psychological concepts are assessed with questionnaires. Consequently, many derived measures are likely confounded by response styles. In study 1, we detected a strong impact of ERS on personality measures of the Big Five, which are frequently used in personnel assessment. Based on the theory of ERS, it is expected that including a measure of ERS as predictor in predictive models should increase the performance. On the one hand, item responses should be a better reflection of the measured construct, when controlling for different response tendencies. This should lead to more accurate predictions of related criteria. On the other hand, ERS might by itself be a useful predictor for some criteria. An increasing number of studies support the idea that response styles should be treated as personality traits (Aichholzer, 2013; Wetzel, Lüdtke, et al., 2016). This is in line with our analysis of the GESIS dataset in study 1, where ERS was found to be stable over time and across questionnaires which measure different constructs. Furthermore, several studies report that ERS is related to other personality characteristics (Naemi et al., 2009; Zettler et al., 2015). These findings suggest that measures of ERS could make a unique contribution to predictions in some cases.

Apart from maximizing the predictive performance of existing psychological measures, predictive modeling can also be used to construct new instruments which are optimized for criterion validity (Chapman et al., 2016). The MMPI-2 (Butcher et al., 1989) which was designed to predict psychiatric diagnosis, is a prominent example for criterion-oriented measures. Chapman et al. (2016) make a strong case that criterion oriented measures might be more powerful, if appropriate predictive methodology is used in the construction process. In some big data applications, a huge amount of data is easily available (Youyou et al., 2015). In contrast, questionnaires with a small number of items but high potential to predict relevant life outcomes might be appealing when actively collecting data is expensive. We think that response styles should also be considered in this process. On the one hand, it might be possible to increase predictive performance by avoiding ERS altogether. In study 2, we could not detect any effect of ERS in two scales with dichotomous

items, which included one scale from the MMPI-2. As discussed in chapter 3.4.4, adopting dichotomous response formats might be a useful strategy to avoid ERS in future test construction. On the other hand, items with a large number of response categories might yield higher predictive performance, if ERS is controlled for by including an appropriate measure in predictive models. Predictive modeling provides useful tools like cross-validation (see chapter 4.1.2) to reveal which strategies are more promising for researchers who want to construct effective criterion-oriented measures.

Yarkoni and Westfall (2017) explain how methods from predictive modeling can be useful even if predicting a practically relevant criterion is not the main focus of a research project. They point out that many topics in psychology could be investigated more clearly by formulating research questions which can be answered by predictive analyses. Study 3 is a good example for this process. To investigate if the impact of ERS can be reduced by special instructions, we compared predictive performance for external criteria which should be closely related to the analyzed scales. As our instructions were explicitly targeted to alter certain item responses, the traditional approach of using psychometric models to analyze item responses under different instructions would not have been meaningful on its own. Using item responses to predict related criteria was necessary to rule out that our instructions removed the pattern of ERS but simultaneously reduced the amount of valid information in the item responses.

Many predictive models are sometimes considered "black box" algorithms (Breiman et al., 2001) because predictions cannot be described by easily interpretable equations, as is the case in linear models. Nonetheless, shifting the focus on prediction might help to better understand psychological phenomena (Yarkoni & Westfall, 2017) and lead to the refinement of theories (Chapman et al., 2016). By estimating predictive models with and without measures of ERS in study 3, we achieved an interpretable statement about the impact of ERS on criterion validity in our dataset. Additionally, we used variable importance measures and individual partial dependence plots (Goldstein et al., 2015) to better understand the effect of ERS on item responses. We suggest that these methods could be applied in many psychological research projects. Variable importance measures and individual partial dependence plots might be especially useful in large scale test construction projects like the program for international student assessment (PISA), to better understand the effectiveness of questionnaire items. PISA datasets have already been used to study responses styles due to their large sample size (see for example Plieninger & Meiser, 2014). They would be good candidates to continue our work in study 3 by further investigating the impact of ERS on criterion validity.

5.3 Conclusion

Although an old topic (Cronbach, 1950), research on response styles has lost none of its importance. At first glance, responding to simple questionnaire items is one of the least complex human behavior imaginable. When studying ERS, however, it becomes clear that the whole item response process is still poorly understood. This is reason for concern,

given that the majority of psychological research relies on questionnaire measures with ordinal response formats. Applied psychological measurement is dominated by simple measurement models that completely ignore response styles. We illustrated that ERS can be detected in an objective way which leaves little doubt that it is a prevalent phenomenon, whenever using items with ordinal response format. By taking "Item Response Theory" literally, psychology should seek to find adequate models that draw a realistic picture of how psychological measurement with questionnaires can work. Until satisfying approaches are available, ERS can easily be avoided by using items with dichotomous response formats. Other procedural remedies against ERS, like targeted instructions to reduce the impact of the response style might be promising, but require further research. Last but not least, a shifted focus on predictive modeling in psychological science (Yarkoni & Westfall, 2017) might stimulate new research on response styles. In a predictive setting, a deeper understanding of the effect of response styles on the criterion validity of psychological measures could be extremely valuable. Hopefully, this will help to improve the usefulness of psychology as an applied science.

Appendices

Appendix A

Algorithm for the Automatic Selection of Uncorrelated Items

To find a set of N uncorrelated items in an item pool of M items for the computations of the ERS indices in study 1, we devised a handcrafted version of simulated annealing:

In the algorithm, the test statistic of the Bartlett test of sphericity (Bartlett, 1951) is used as the objective function to be minimized. In each of the B iterations, a new set of N items is proposed in which one of the previous items is replaced by a new one. The item to be replaced is found as follows:

1. For each item in the current set, the sum of the squared first order correlations with all other items in the current set is computed. Sums are then ordered decreasingly.
2. A random number x is drawn from a beta-binomial distribution with fixed parameters $n = N - 1$, $a = 1$ and parameter b which is changed dynamically during the optimization process.
3. The item corresponding to the position $x+1$ in the ordered list from Step 1 is replaced in the new candidate set.

The new candidate item is found as follows:

1. For each item not in the current set, the sum of the squared first order correlations with all items in the current set is computed. Sums are then ordered increasingly.
2. A random number y is drawn from a beta-binomial distribution with fixed parameters $n = N - 1$, $a = 1$ and dynamical parameter b which is changed dynamically during the optimization process.
3. The item corresponding to the position $y + 1$ in the ordered list from Step 1 is the new candidate item.

The parameter b in both beta-binomial distributions is changed dynamically during the optimization process. Starting with $b = 1$, it is increased by 1 after each $b.keep$ iterations.

Note that for $a = b = 1$ the beta-binomial distribution is the uniform distribution on the interval $[0, n]$. If b is increased, the distribution gets skewed and the probability for smaller values increases until the value 0 has almost probability 1. Thus, in the beginning of the optimization process, all items in the current set are equally likely to be replaced, whereas in later iterations, items with high correlations are much more likely to be replaced. Similarly, at first the N items showing the lowest correlations with the current item set are equally likely to be the new candidate item, whereas in later iterations items with lower correlations are much more likely to be selected.

A new candidate item set is always accepted if its Bartlett test statistic is lower than the test statistic of the current set ($Bart_{new} - Bart_{old} = \Delta Bart < 0$). If the test statistic of the candidate set is higher, the new set is probabilistically accepted according to the formula,

$$P_{Accept} = \exp\left(-\frac{\Delta Bart}{T}\right) \tag{A.1}$$

in which T is called the temperature. When the temperature is high, P_{Accept} is large, even for high values of $\Delta Bart$. For low temperatures, P_{Accept} is close to zero for ($\Delta Bart > 0$). The temperature is initiated with a high value $T.start$ and is reduced by a factor $T.fac$ after each $T.keep$ iterations.

The resulting algorithm has 5 configurable hyperparameters: B, $T.start$, $T.fac$, $T.keep$, and $b.keep$. To achieve good optimization performance, we set $B = 3000$, $T.start = 100$ and tuned the other parameters using iterated F-Racing (Birattari, Yuan, Balaprakash, & Stützle, 2010), a general procedure for configuring optimization algorithms, as implemented in the R package *irace* (López-Ibáñez et al., 2015).[1] We performed 1000 irace experiments. $T.keep$ and $b.keep$ were treated as integer parameters with feasible region between 1 and 500 each. $T.fac$ was treated as a real parameter between 0 and 1. An instance for the irace algorithm was generated by drawing 200 observations from a multivariate normal distribution with zero mean vector and covariance matrix equal to the sample correlation matrix of all items considered for the ERS index. The sample correlation matrix computed from these 200 observations was then used in the simulated annealing algorithm instead of the original correlation matrix of all considered items.

For the analysis of the NEO-PI-R dataset in chapter 2, irace tuning was only performed once, instead of tuning the algorithm for the computation of each of the five ERS indices separately. Instances were created by drawing from a multivariate normal distribution with a correlation matrix based on the sample correlation matrix of all NEO-PI-R items, instead of only four Big Five factors at a time. This resulted in one hyperparameter set which was then used in the optimization algorithm for finding the item sets of all five ERS indices. For the analysis of the GESIS dataset, a seperate tuning process was performed, as the GESIS dataset and the NEO-PI-R dataset differ with respect to several characteristics. Table A.1 shows the tuned parameter values used to find sets of uncorrelated items for the NEO-PI-R and the GESIS analyses.

[1] We refrain from describing iterated F-Racing here. Please refer to López-Ibáñez, Dubois-Lacoste, Cáceres, Birattari, and Stützle (2016) for further details.

Table A.1: Results of Tuning the Hyperparameters of the Item Selection Algorithm in Study 1 with Iterated F-Racing

	$T.keep$	$T.fac$	$b.keep$
NEO-PI-R	111.00	0.74	284.00
GESIS	355.00	0.52	104.00

Note. Remaining parameters were fixed to $B = 3000$ and $T.start = 100$. B = total number of iterations, $T.start$ = starting temperature, $T.keep$ = number of iterations until the temperature is changed, $T.fac$ = factor by which the temperature is reduced, $b.keep$ = number of iterations until the b parameter of the beta-binomial distribution is changed.

In the final item selection process, the item set with the smallest Bartlett test statistic was used for the computation of ERS indices. Note that this is not necessarily the item set from the last iteration of the optimization algorithm.[2] Also, simulated annealing is not guaranteed to find the global minimum. We refrained from running simulated annealing multiple times and choosing the best solution, as results only differed slightly and had no impact on the interpretation of any further analyses.

[2] For irace tuning, the Bartlett test statistic at the end of the optimization process was used to evaluate the performance of the run. This should give a slight advantage to parameter configurations that minimize the risk of the algorithm to get stuck in local minima.

Appendix B

Supplemental Material for Study 1

B.1 Description of ERS Items (GESIS)

Labels of the items which were used to compute the ERS index from heterogeneous items in the GESIS dataset.

Table B.1: Labels of Selected ERS Items in the GESIS Dataset

	ERS Items
1	EU-Integration-Einstufung: Die Linke
2	Civic duty: Keine Steuern hinterziehen
3	Big Five: Habe nur wenig künstlerisches Interesse
4	Wahrscheinlichkeit Wahl: SPD
5	Verbundenheit: Region
6	Verantwortung Wirtschaftliche Lage: Bundesregierung
7	Big Five: Bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen
8	Mit Schönheits-Ops auseinandergesetzt
9	Big Five: Erledige Aufgaben gründlich
10	NEP-Skala: Nähern uns Höchstzahl an Menschen
11	NEP-Skala: Menschen Naturgesetzen unterworfen
12	Idealer Beruf: Viel Freizeit
13	Finanzielle Situation Haushalt heute, Vergleich vor 12 Monaten
14	Beeinträchtigung Umwelteinflüsse:Lärmbelästigung
15	NEP-Skala: Genügend natürliche Rohstoffe
16	Parteiidentifikation Stärke
17	Urlaubsmotive: Zeit mit Anderen
18	Wochenmarkt: Lebensmittel hohe Qualität
19	Politische Entscheidungsfindung: Experten
20	Energiewende: Erneuerbare Energien machen Deutschland unabhängig von anderen Län
21	Werte: Anderen sagen was sie tun sollen
22	EU-Integration-Einstufung: CDU
23	Persönliche Priorität: Glückliche Ehe/Partnerschaft
24	Attraktivität: Attraktiven Menschen fliegt alles zu
25	Wochenmarkt: schlecht erreichbar
26	Urlaubsmotive: Unabhängig fühlen
27	Freizeitaktivität: entspannen und erholen
28	Wahrscheinlichkeit Wahl: FDP
29	Wichtigkeit: Arbeit
30	Big Five: Schenke anderen leicht Vertrauen, glaube an das Gute im Menschen

Note. Item labels are presented as in the official GESIS dataset.

B.2 Histograms of ERS Indices (NEO-PI-R)

Histograms of the extraERS, openERS, agreeERS, and conscERS indices in the NEO-PI-R dataset.

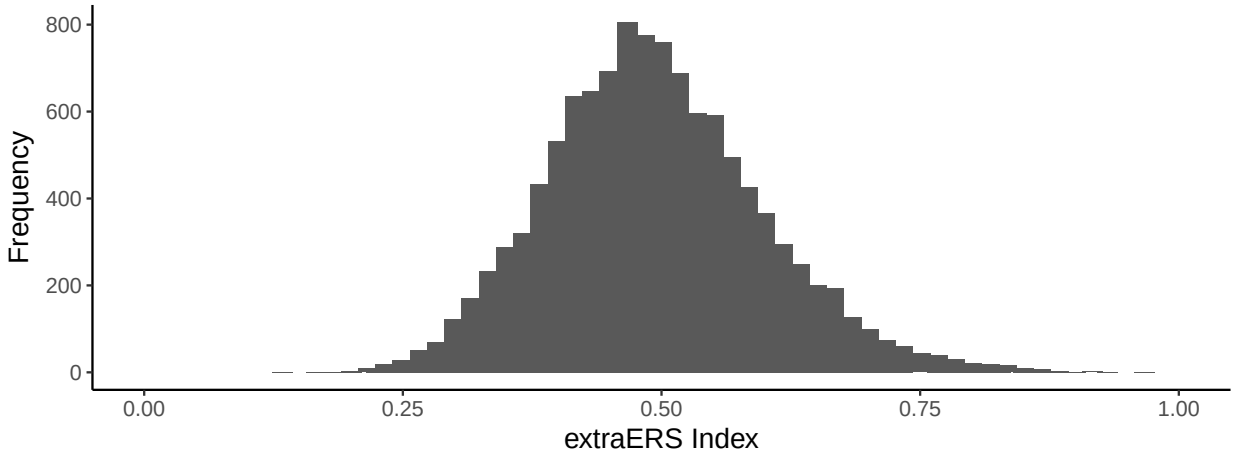

Figure B.1: Histogram of the extraERS index computed from heterogeneous items (except extraversion) in the NEO-PI-R dataset.

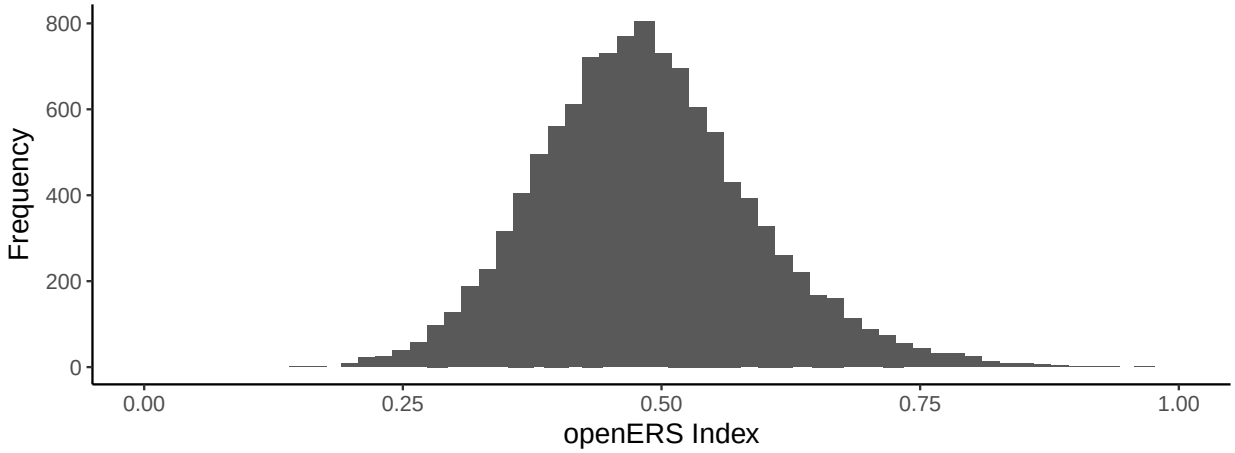

Figure B.2: Histogram of the openERS index computed from heterogeneous items (except openness) in the NEO-PI-R dataset.

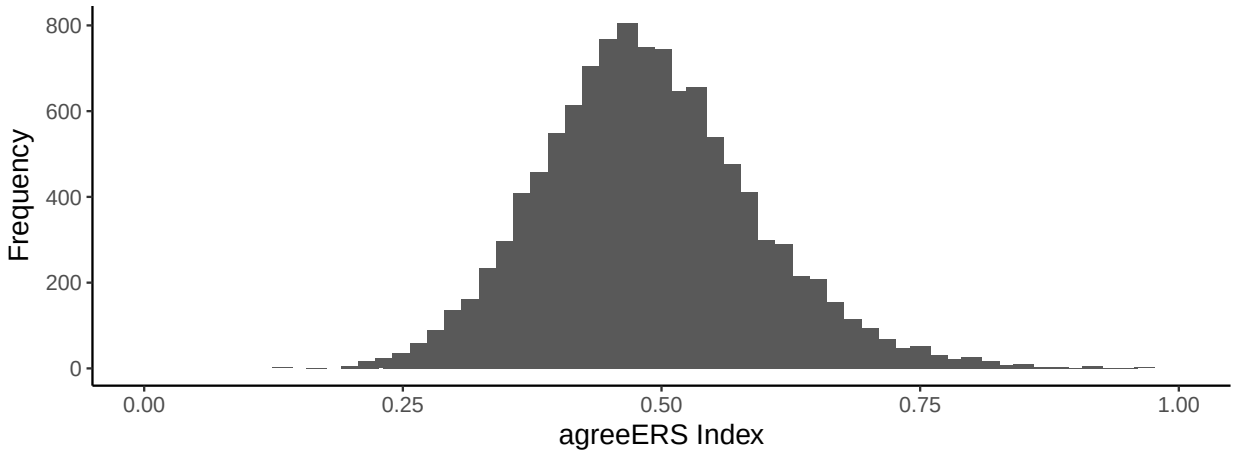

Figure B.3: Histogram of the agreeERS index computed from heterogeneous items (except agreeableness) in the NEO-PI-R dataset.

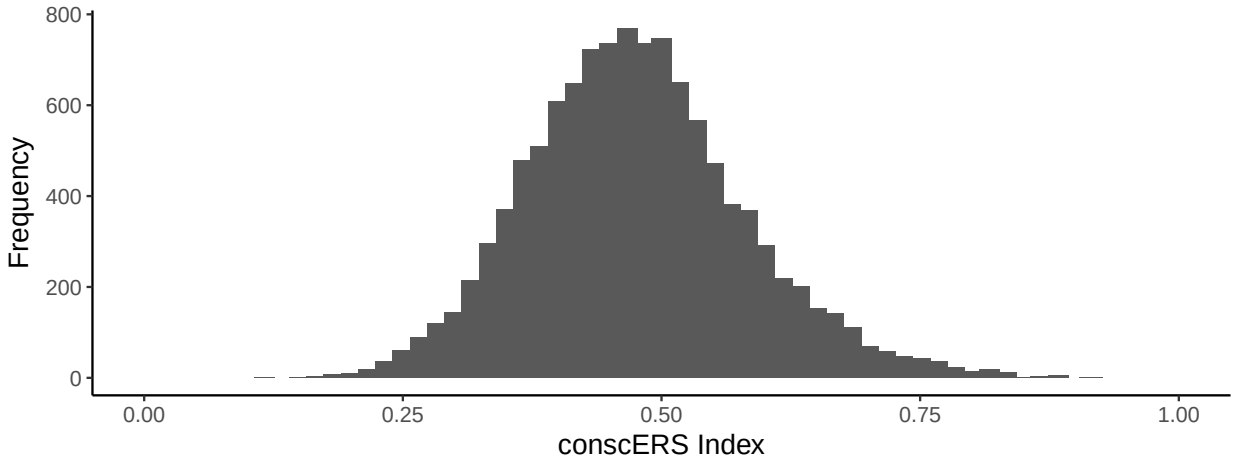

Figure B.4: Histogram of the conscERS index computed from heterogeneous items (except conscientiousness) in the NEO-PI-R dataset.

B.3 PC Trees Without Demographic Variables (NEO-PI-R)

PC trees for the analyzed personality facets in the NEO-PI-R dataset, with the respective ERS index from heterogeneous items as single covariate.

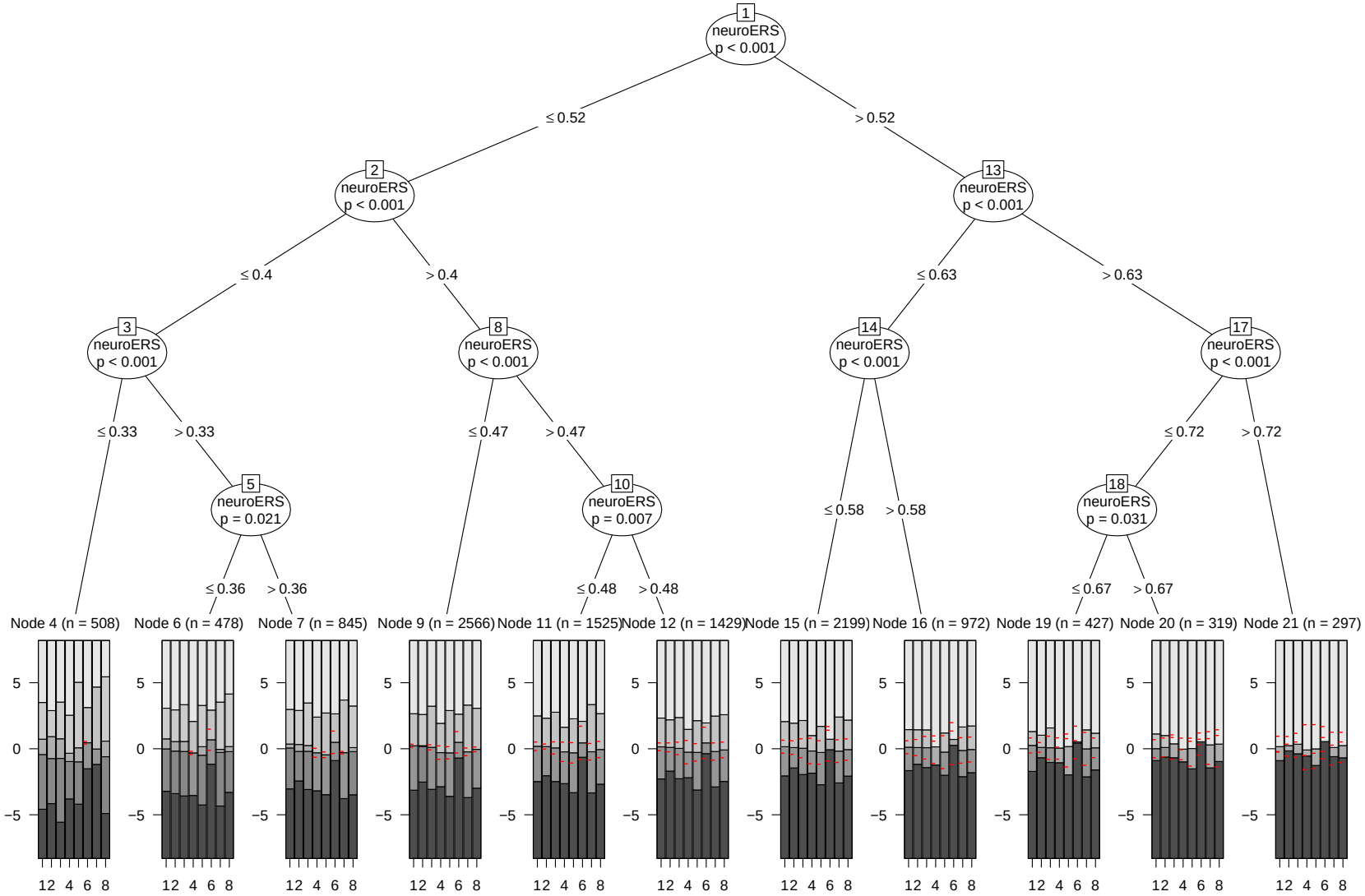

Figure B.5: PC tree of the NEO-PI-R facet Impulsivity (N5) with the neuroERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11565$.

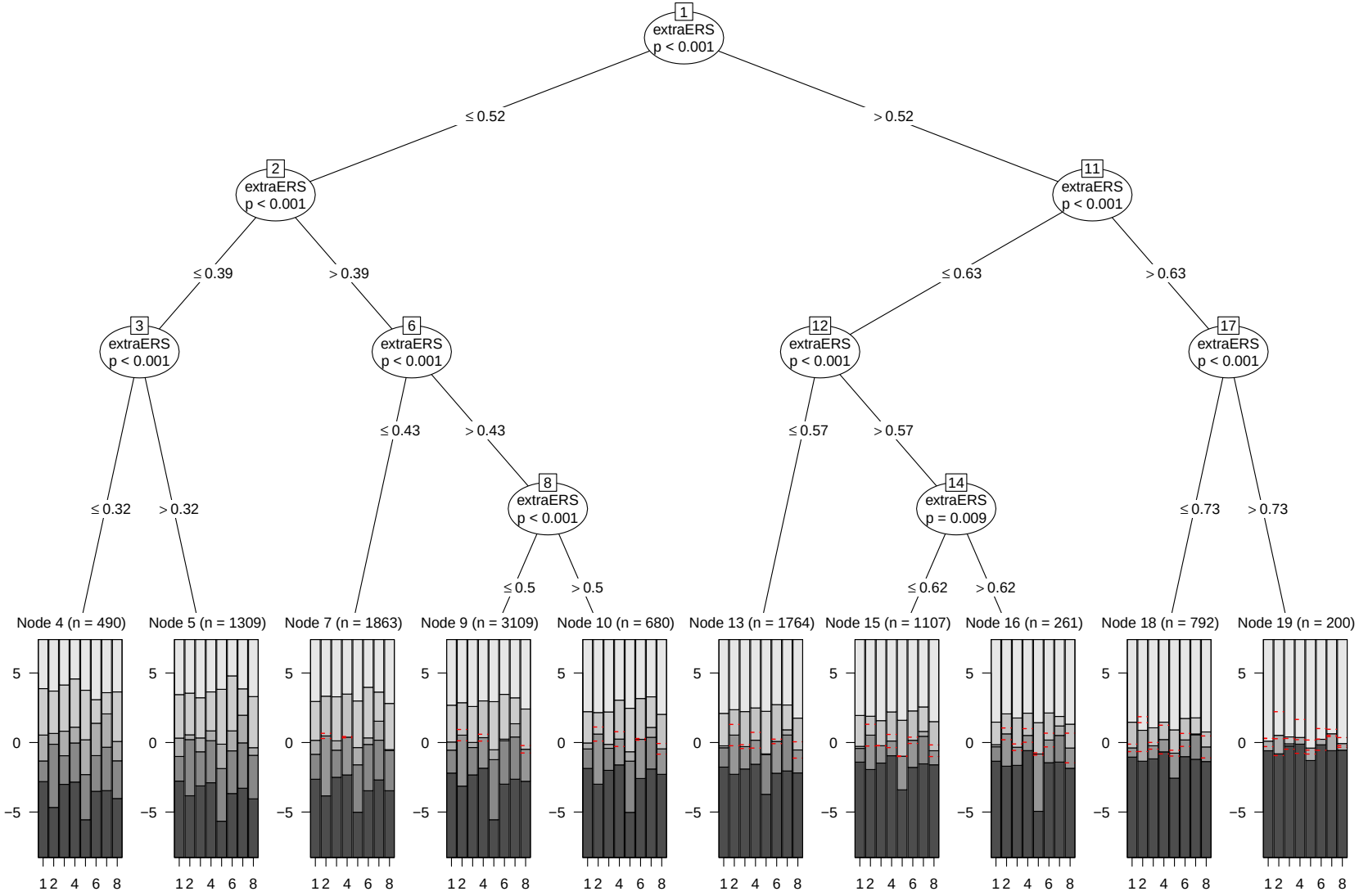

Figure B.6: PC tree of the NEO-PI-R facet Assertiveness (E3) with the extraERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11575$.

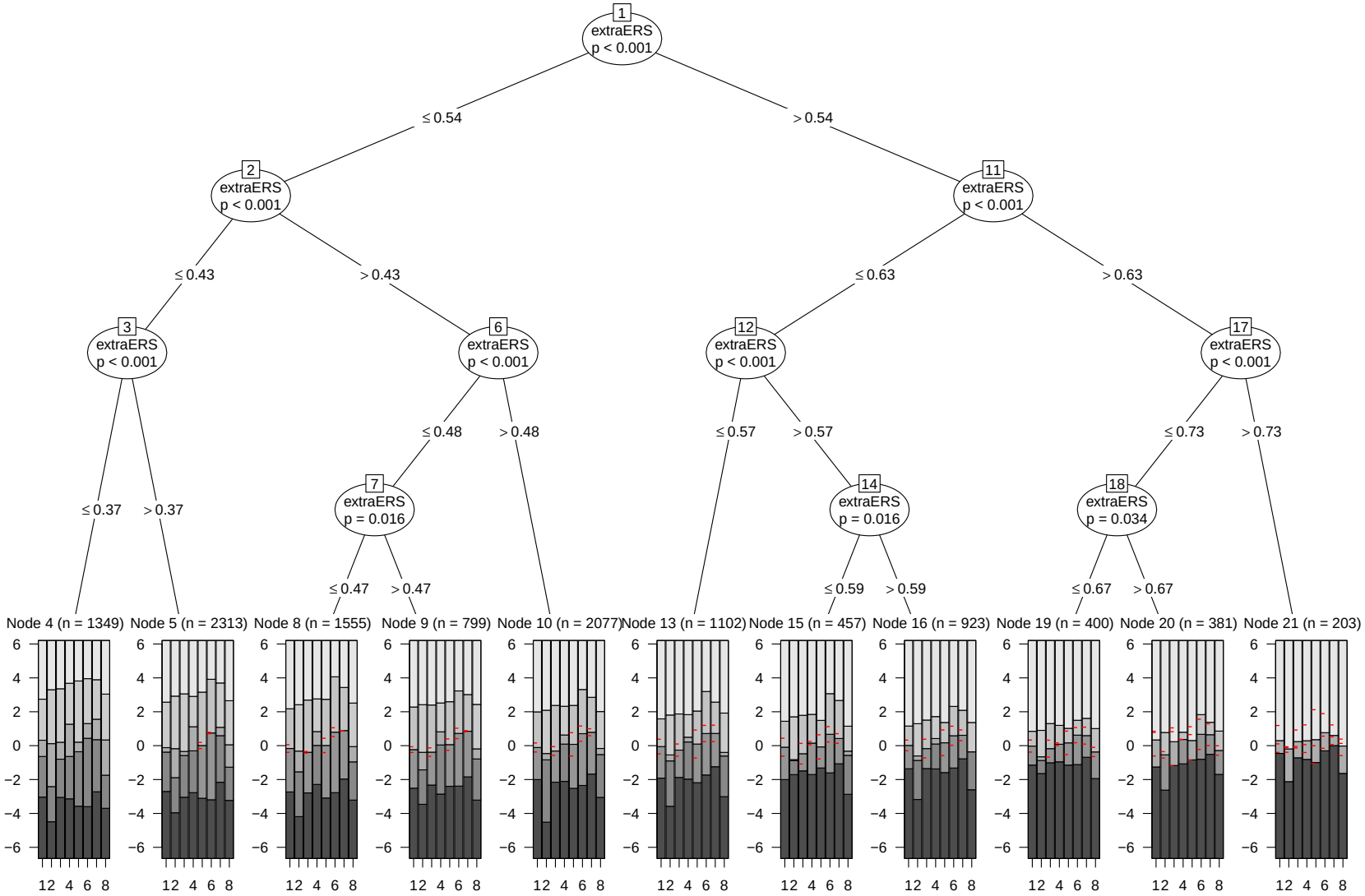

Figure B.7: PC tree of the NEO-PI-R facet Activity (E4) with the extraERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11559$.

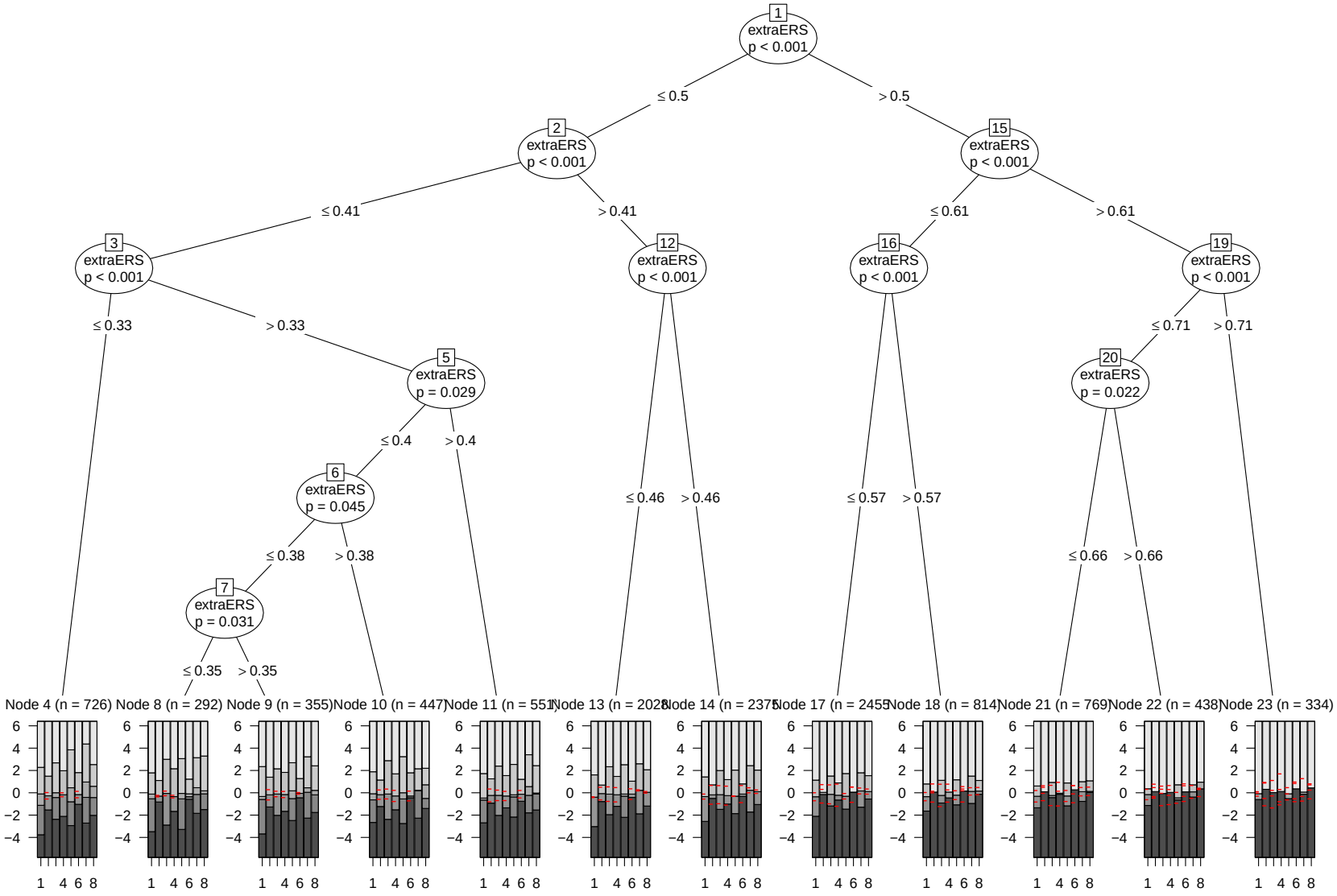

Figure B.8: PC tree of the NEO-PI-R facet Excitement-Seeking (E5) with the extraERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11584$.

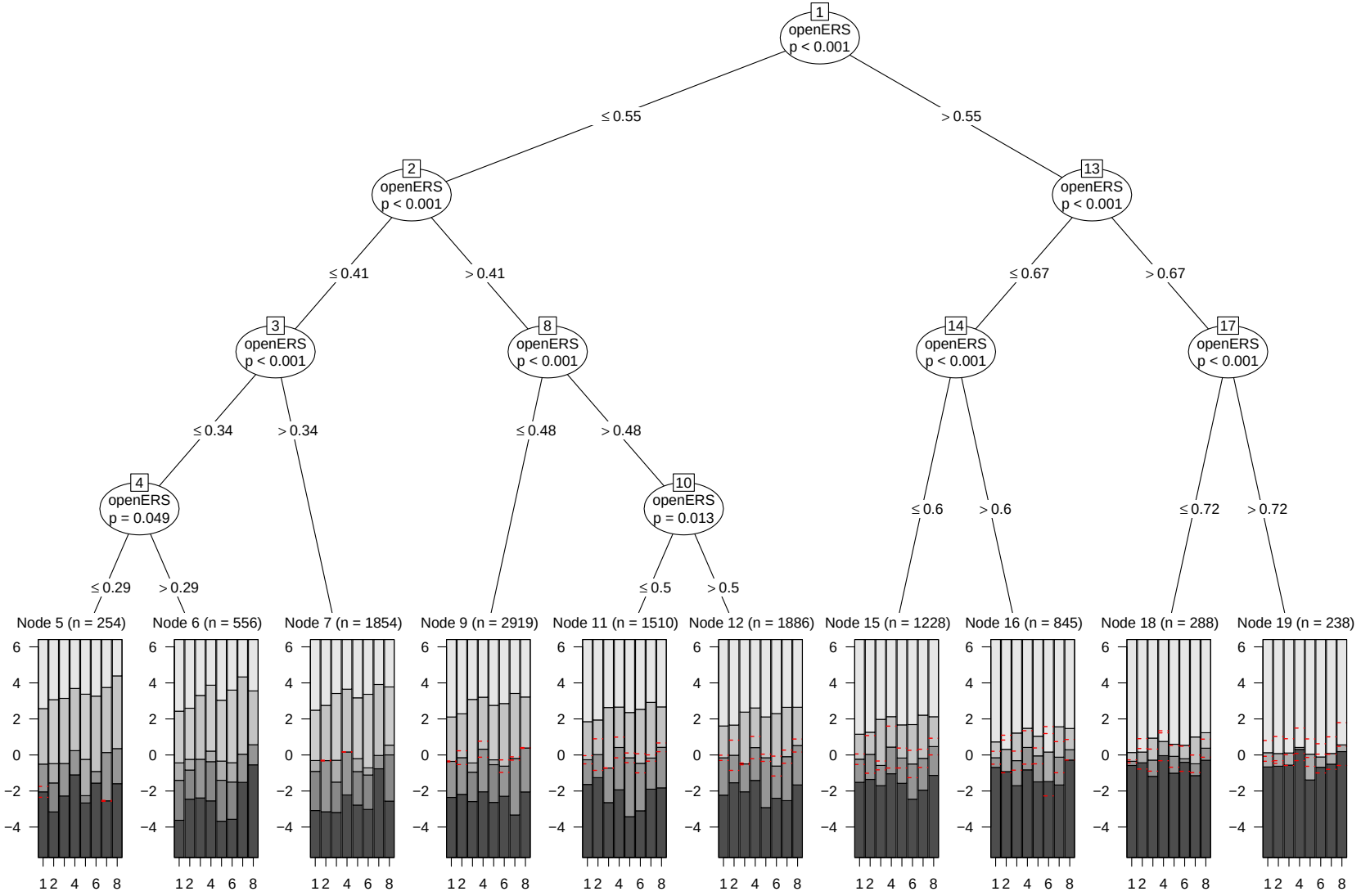

Figure B.9: PC tree of the NEO-PI-R facet Openness to Feelings (O3) with the openERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11578$.

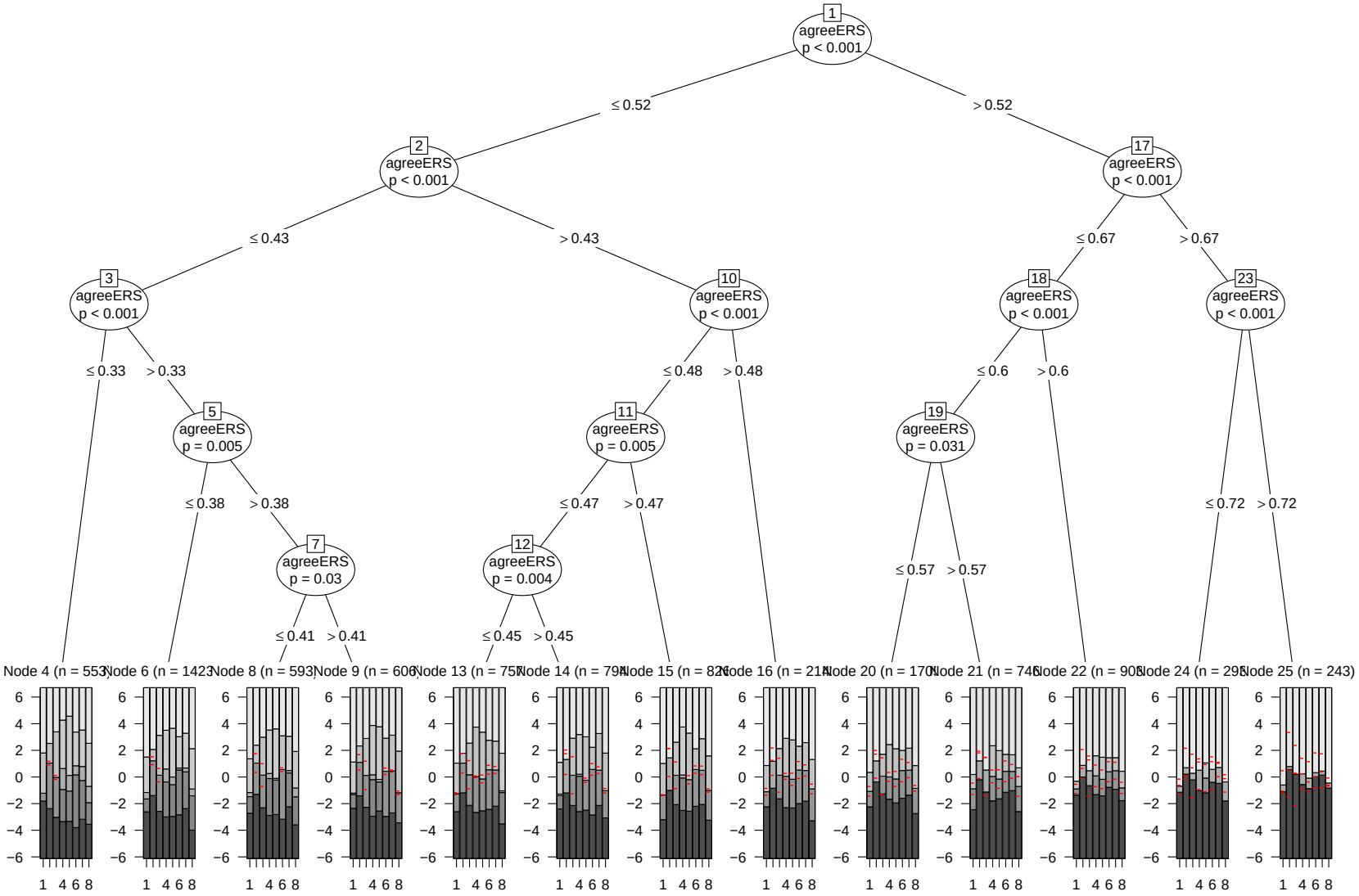

Figure B.10: PC tree of the NEO-PI-R facet Compliance (A4) with the agreeERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11583$.

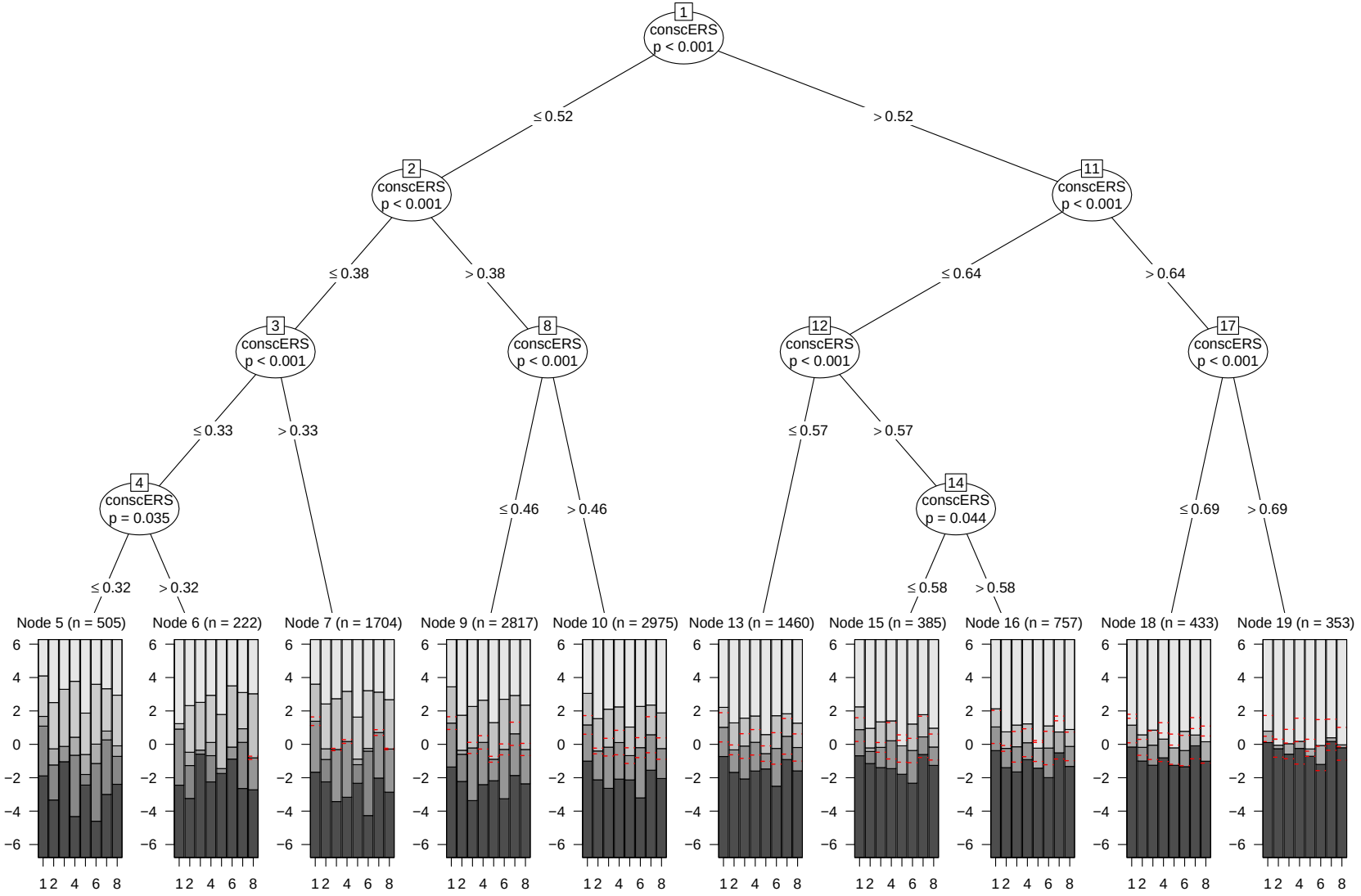

Figure B.11: PC tree of the NEO-PI-R Order (C2) with the conscERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11611$.

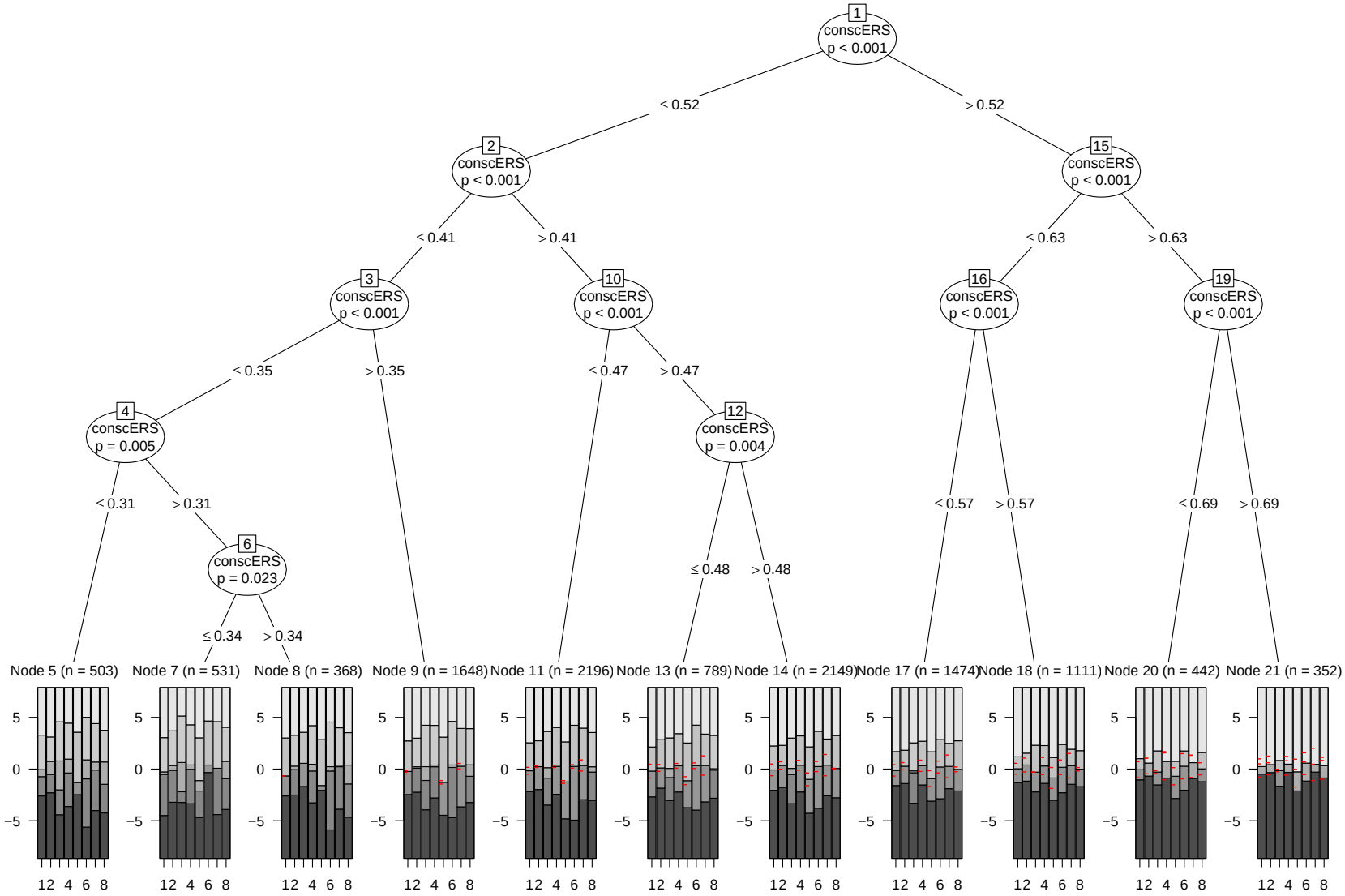

Figure B.12: PC tree of the NEO-PI-R Self-Discipline (C5) with the conscERS index from heterogeneous items as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11563$.

B.4 Relative Response Frequencies in PC Trees (NEO-PI-R and GESIS)

Relative response frequencies in the PC trees of analyzed scales in the NEO-PI-R and GESIS datasets.

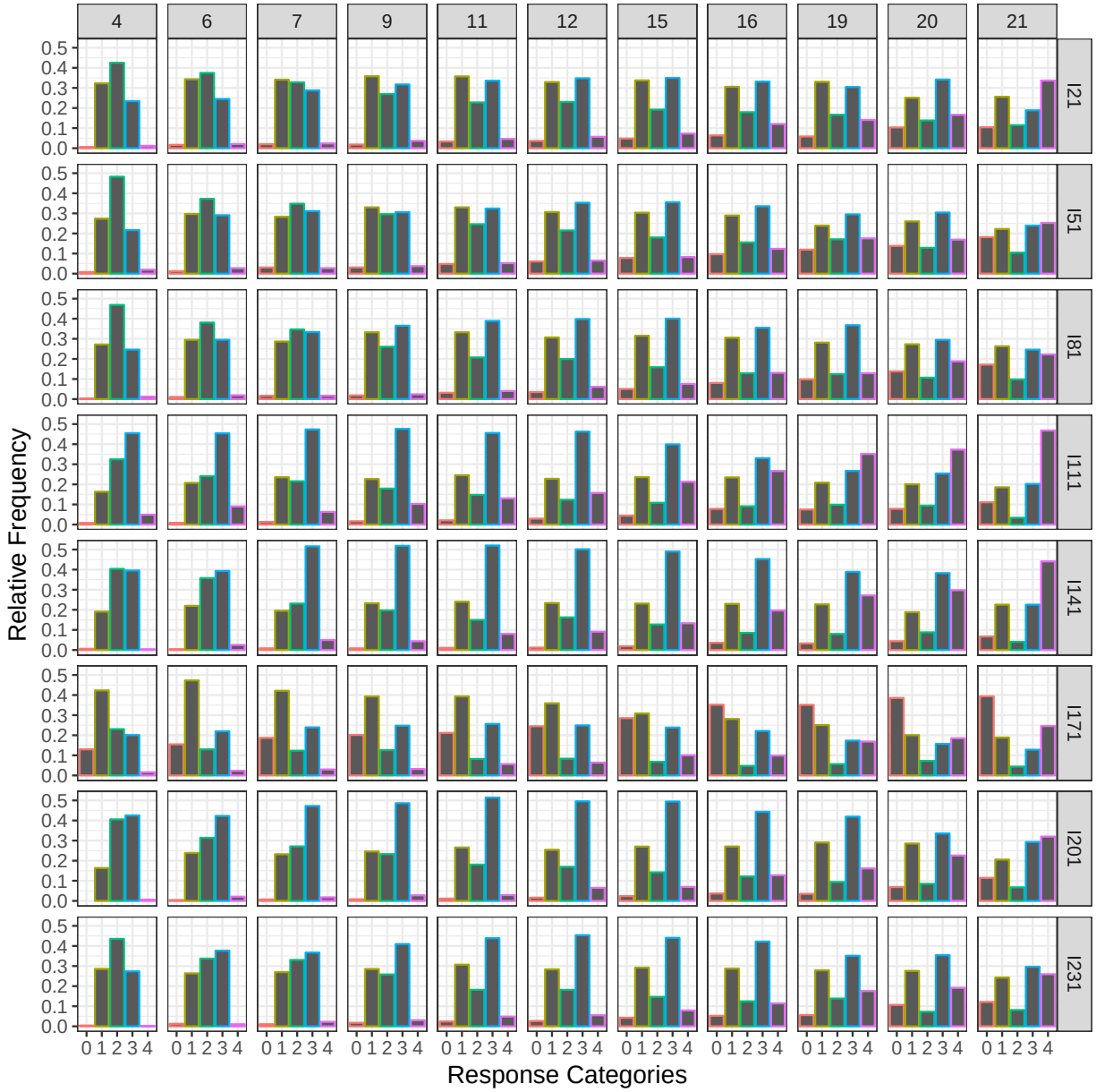

Figure B.13: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the NEO-PI-R facet Impulsivity (N5). Rows represent the items of the scale, numbered according to their position in the NEO-PI-R. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure B.5. Response frequencies sum up to one in each cell.

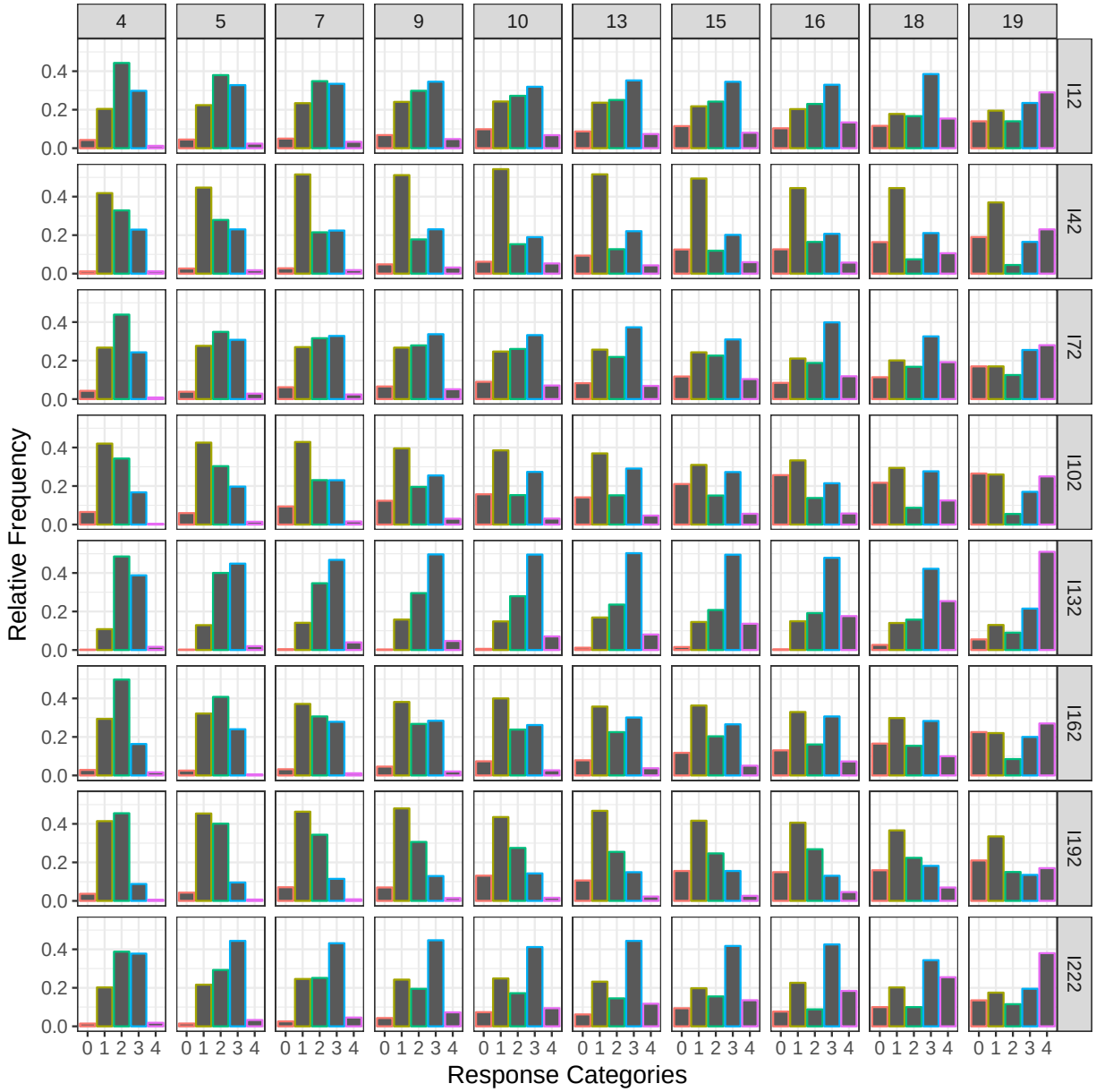

Figure B.14: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the NEO-PI-R facet Assertiveness (E3). Rows represent the items of the scale, numbered according to their position in the NEO-PI-R. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure B.6. Response frequencies sum up to one in each cell.

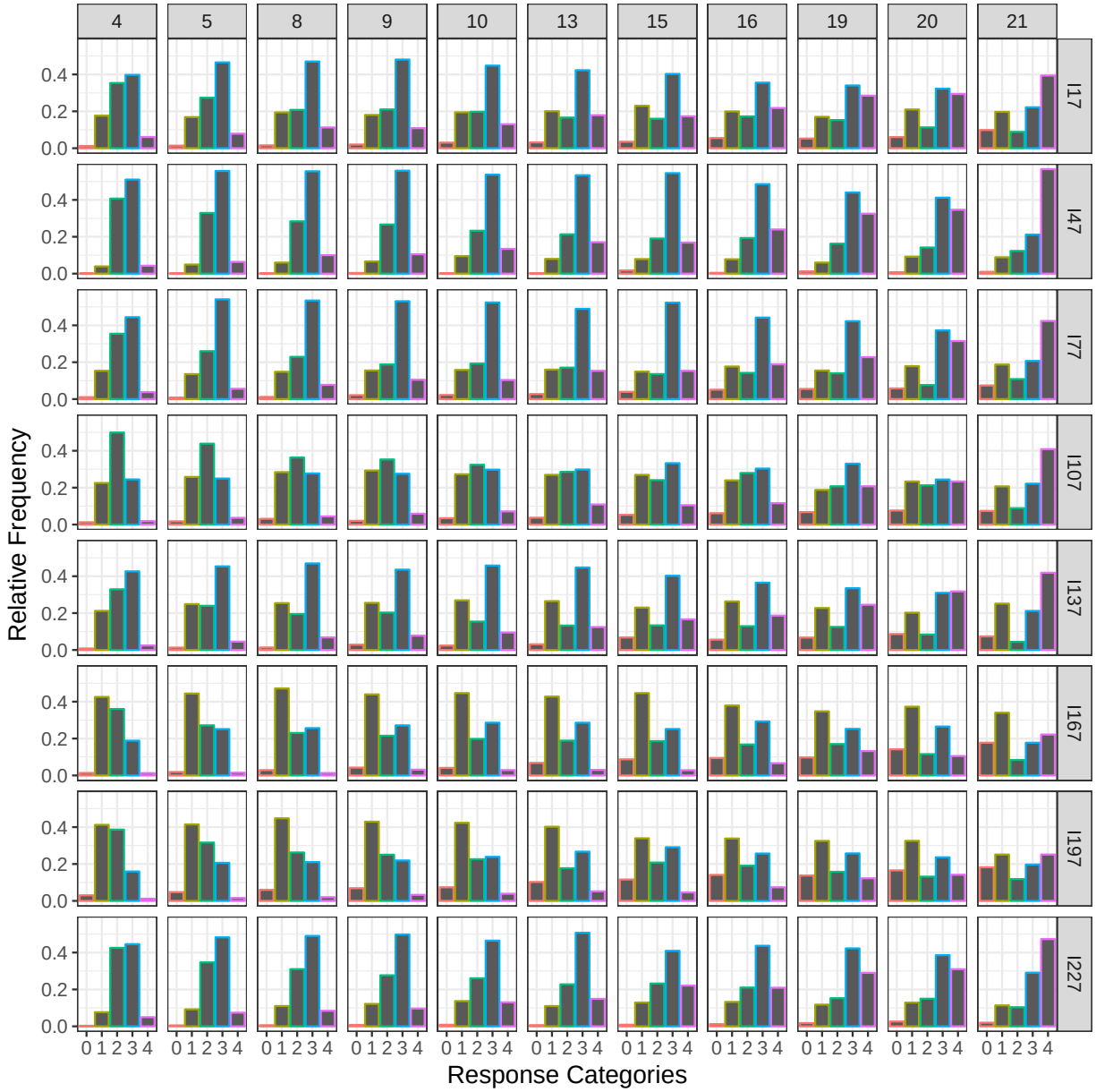

Figure B.15: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the NEO-PI-R facet Activity (E4). Rows represent the items of the scale, numbered according to their position in the NEO-PI-R. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure B.7. Response frequencies sum up to one in each cell.

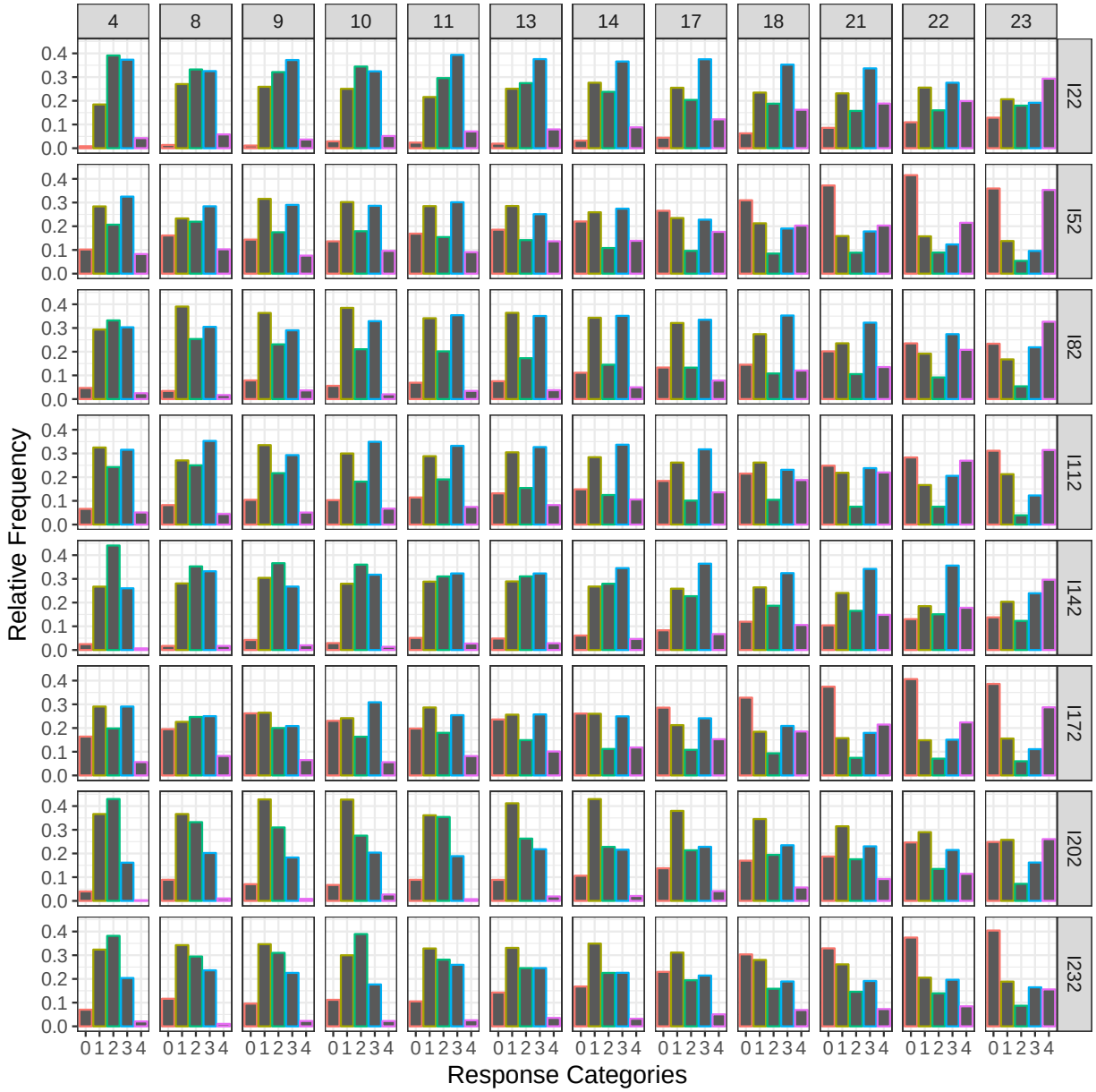

Figure B.16: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the NEO-PI-R facet Excitement-Seeking (E5). Rows represent the items of the scale, numbered according to their position in the NEO-PI-R. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure B.8. Response frequencies sum up to one in each cell.

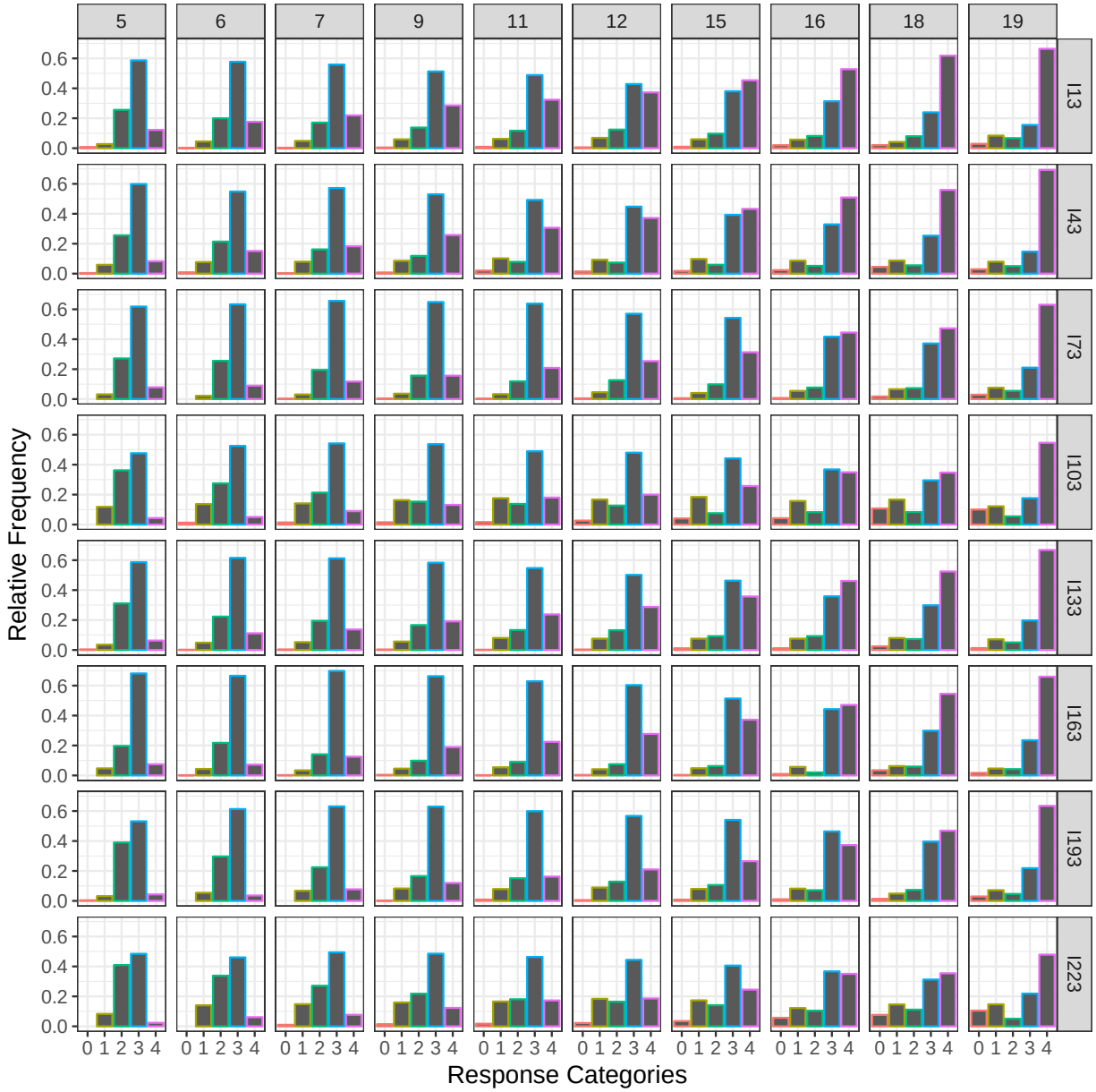

Figure B.17: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the NEO-PI-R facet Openness to Feelings (O3). Rows represent the items of the scale, numbered according to their position in the NEO-PI-R. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure B.9. Response frequencies sum up to one in each cell.

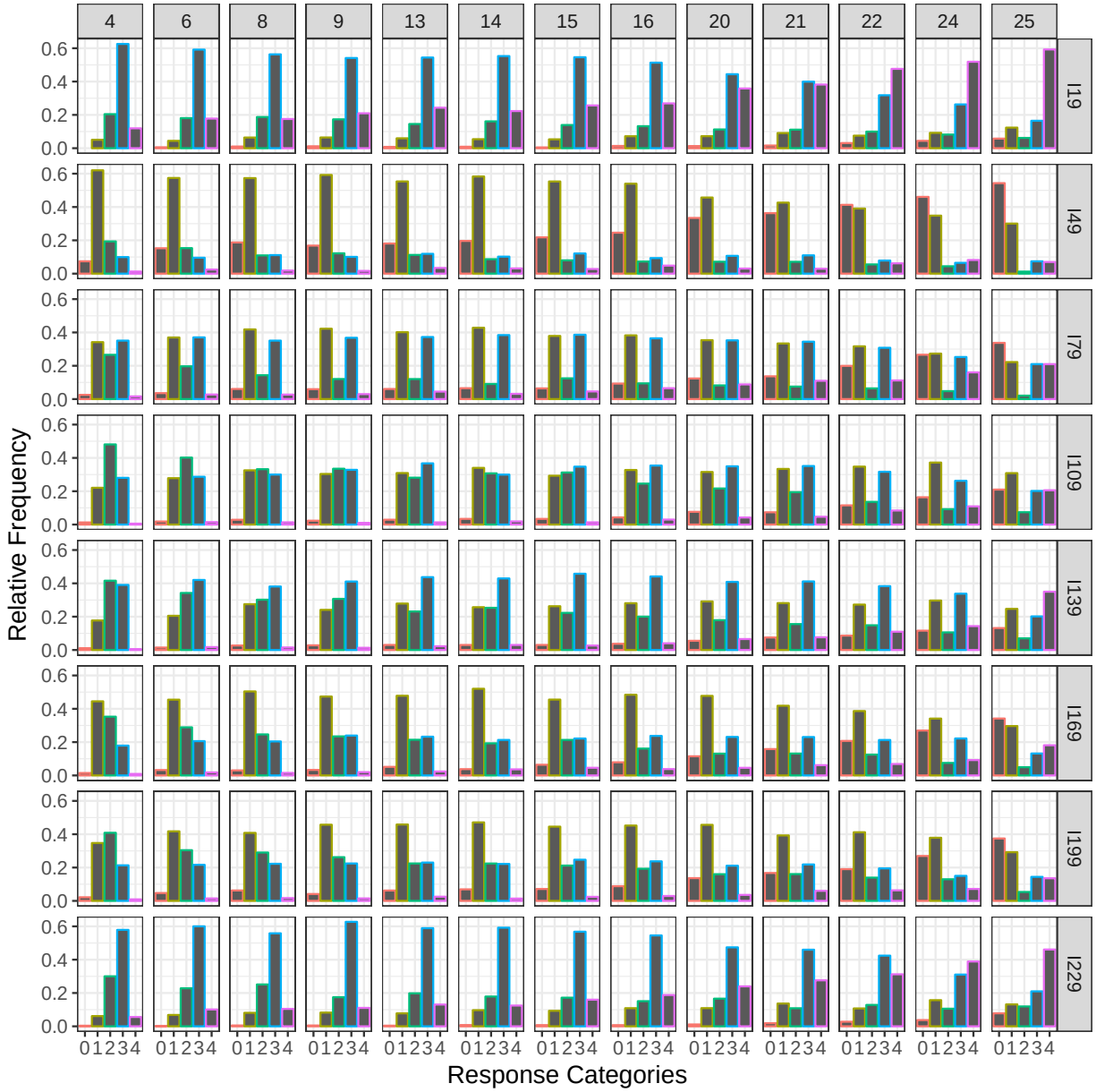

Figure B.18: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the NEO-PI-R facet Compliance (A4). Rows represent the items of the scale, numbered according to their position in the NEO-PI-R. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure B.10. Response frequencies sum up to one in each cell.

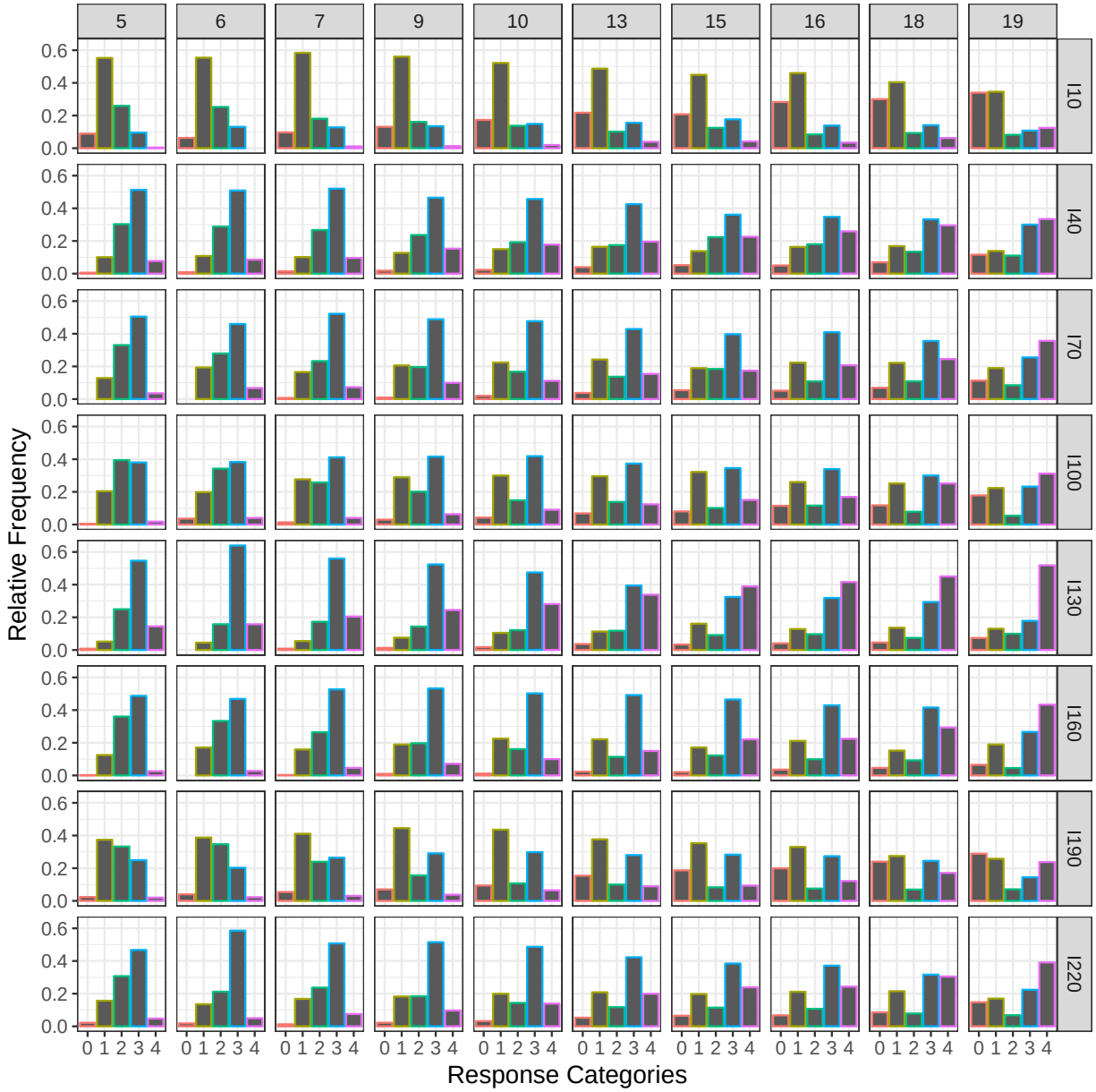

Figure B.19: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the NEO-PI-R facet Order (C2). Rows represent the items of the scale, numbered according to their position in the NEO-PI-R. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure B.11. Response frequencies sum up to one in each cell.

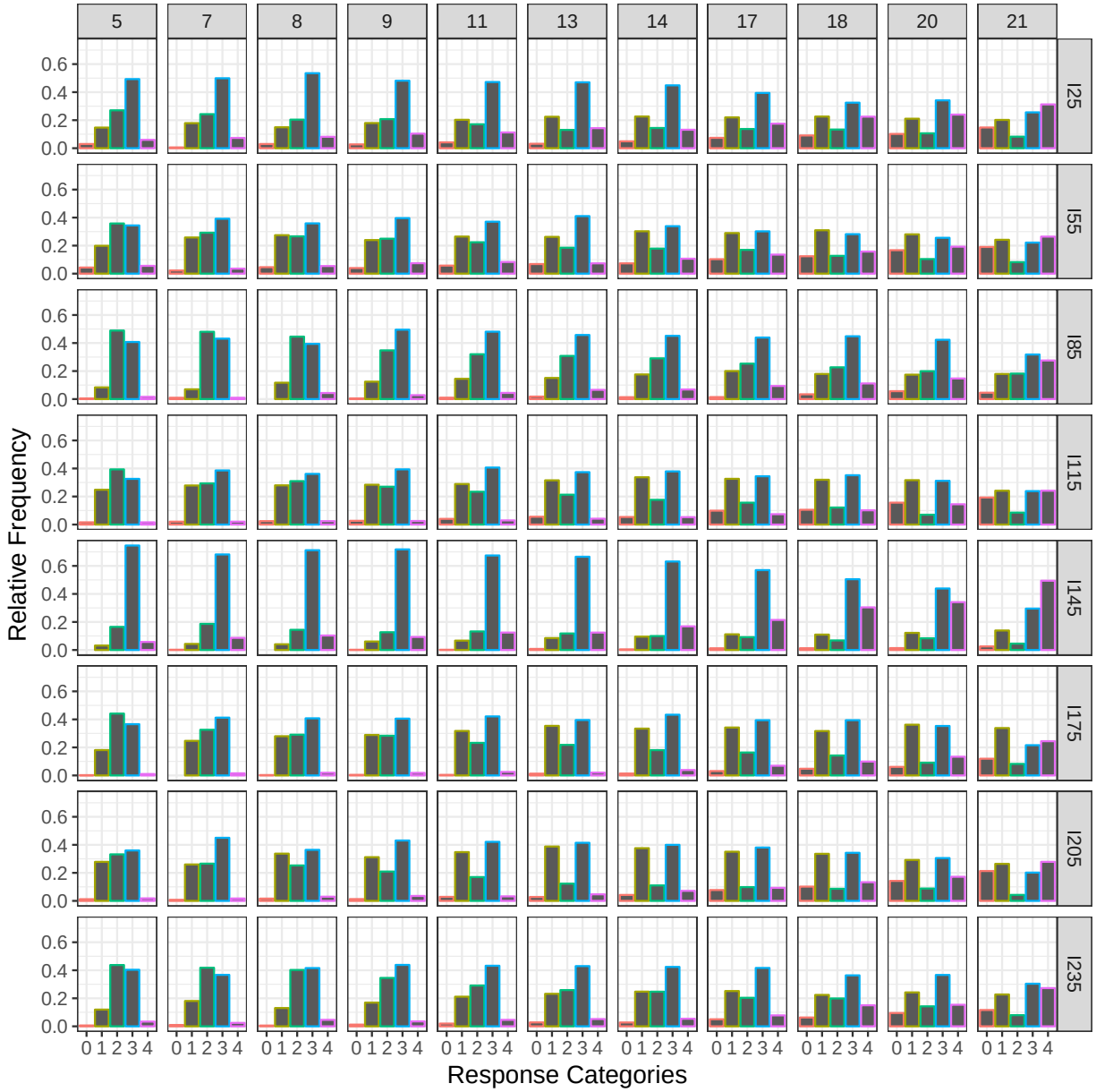

Figure B.20: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the NEO-PI-R facet Self-Discipline (C5). Rows represent the items of the scale, numbered according to their position in the NEO-PI-R. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure B.12. Response frequencies sum up to one in each cell.

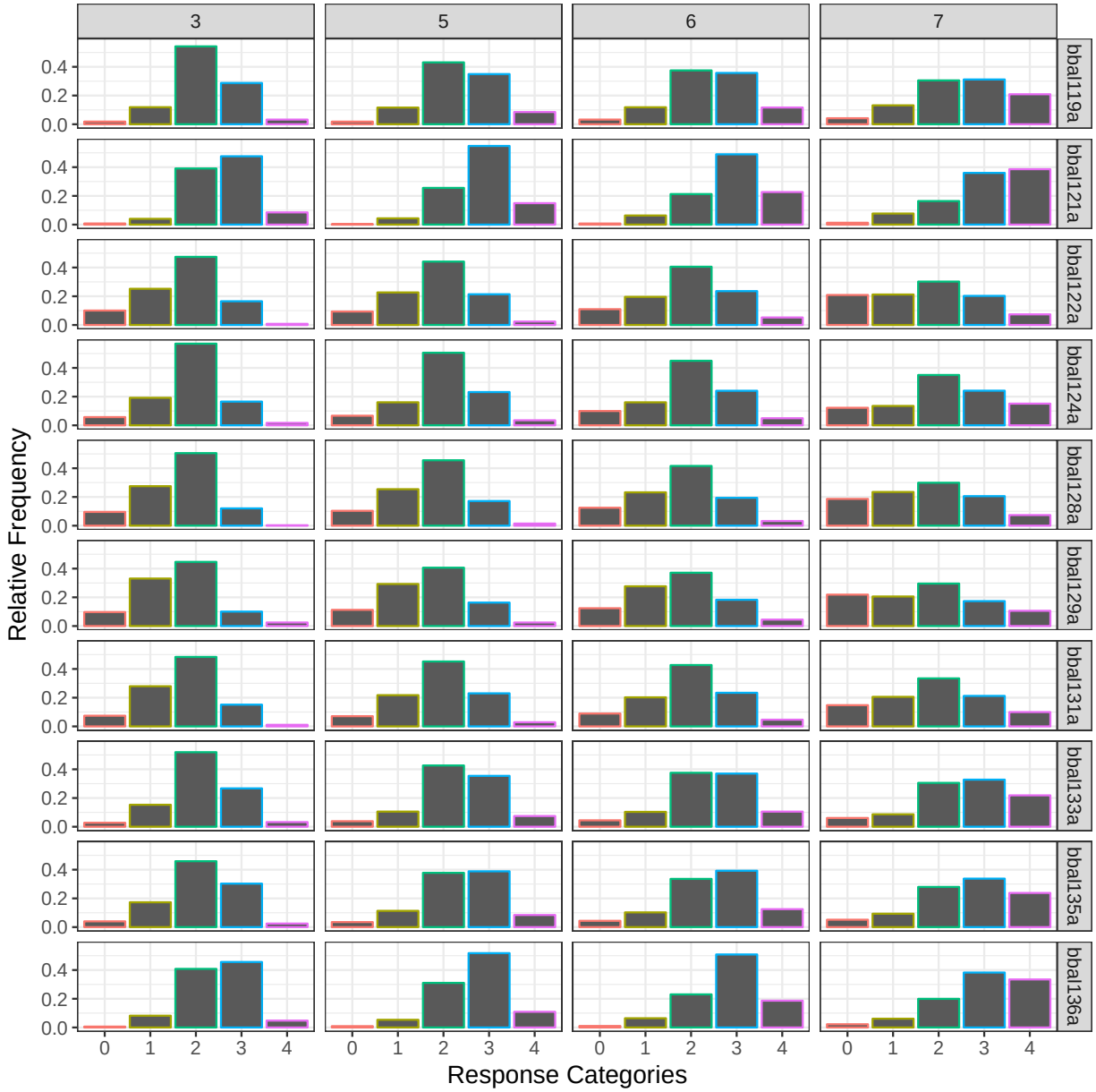

Figure B.21: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the Positive Affect (PA) scale in the GESIS dataset. Rows represent the items of the scale, labeled according to the official dataset. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure 2.6. Response frequencies sum up to one in each cell.

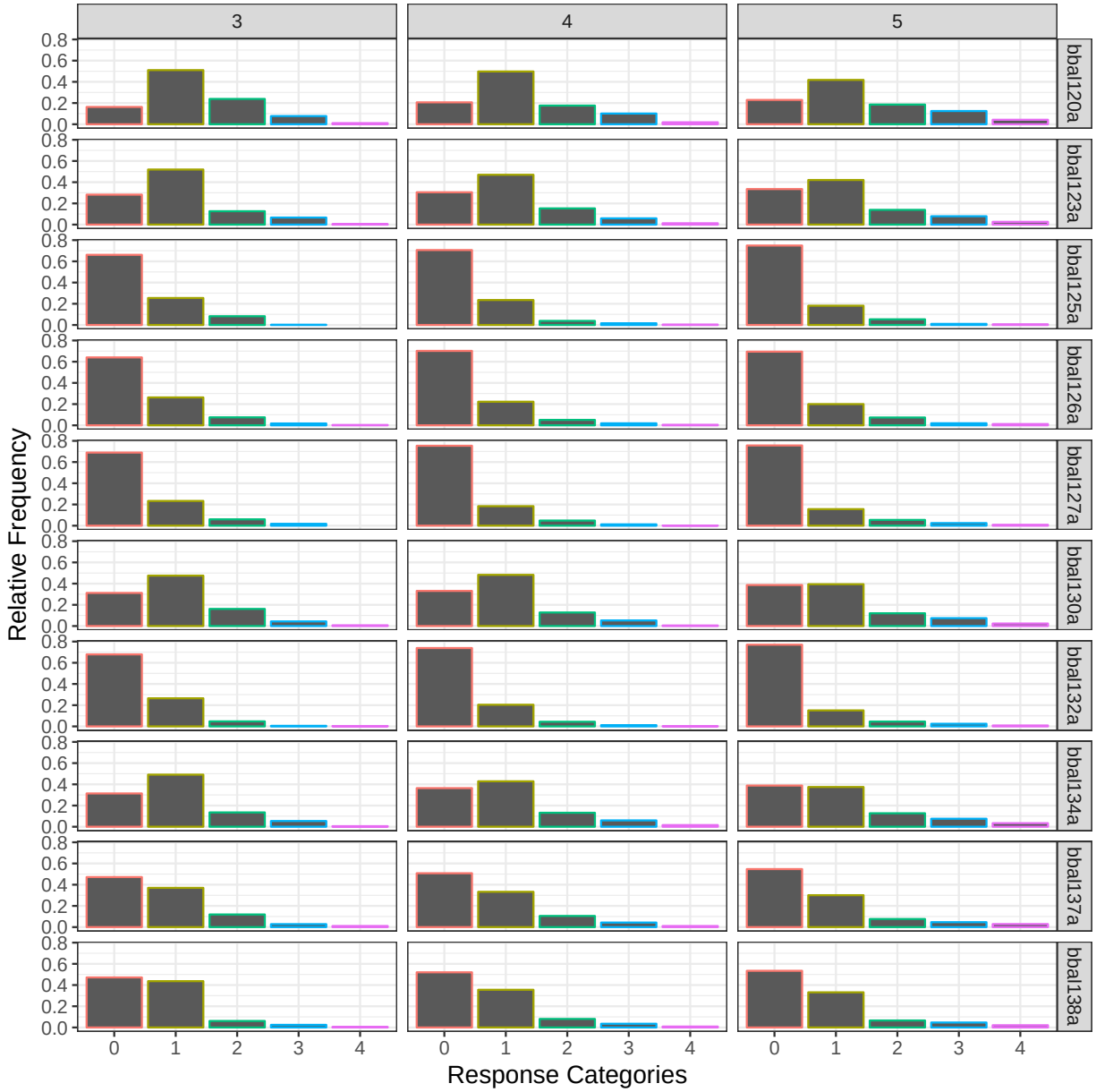

Figure B.22: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the Negative Affect (NA) scale in the GESIS dataset. Rows represent the items of the scale, labeled according to the official dataset. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure 2.7. Response frequencies sum up to one in each cell.

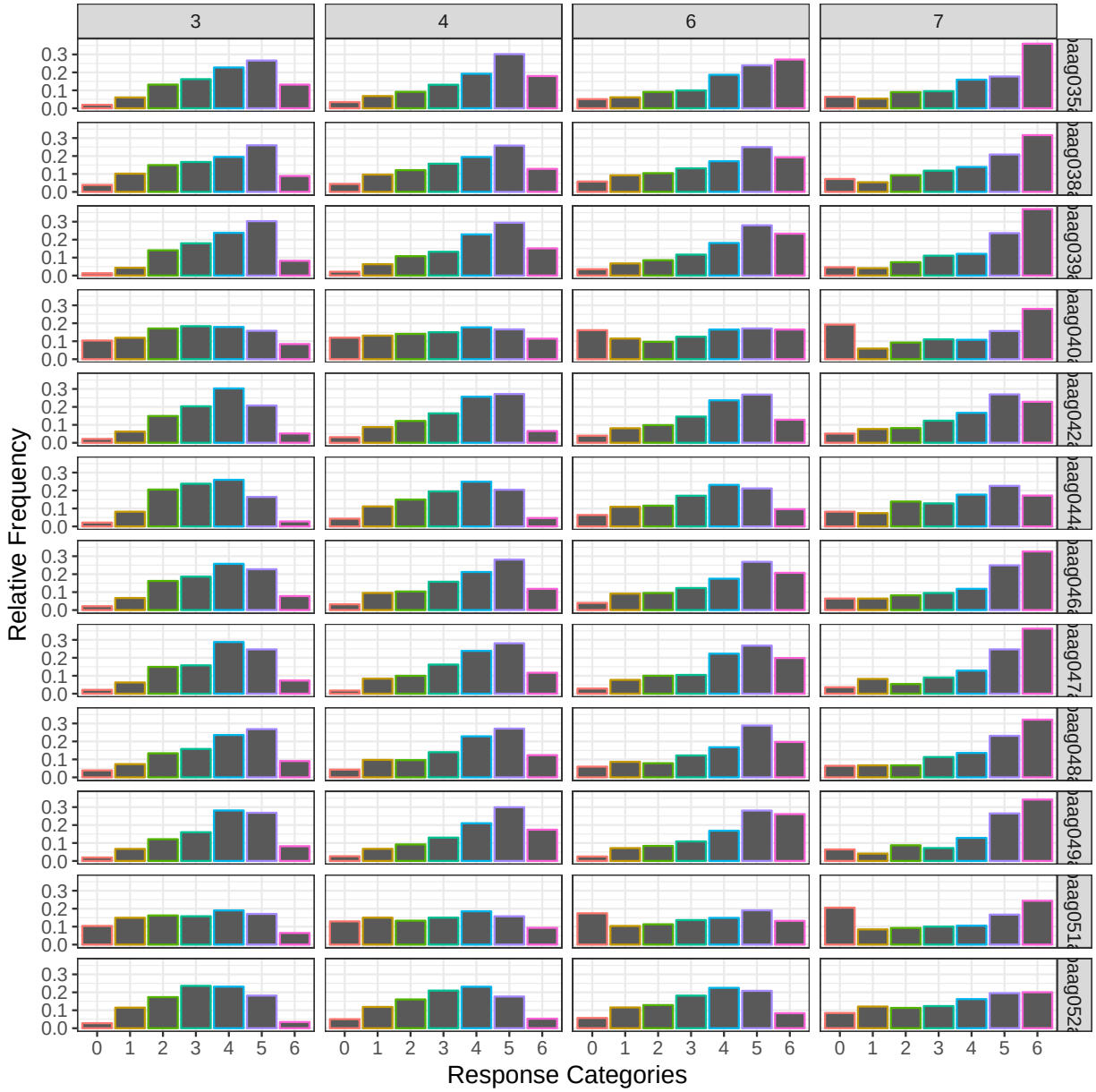

Figure B.23: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the Global/Egocentric Orientation (GEO) scale in the GESIS dataset. Rows represent the items of the scale, labeled according to the official dataset. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure 2.8. Response frequencies sum up to one in each cell.

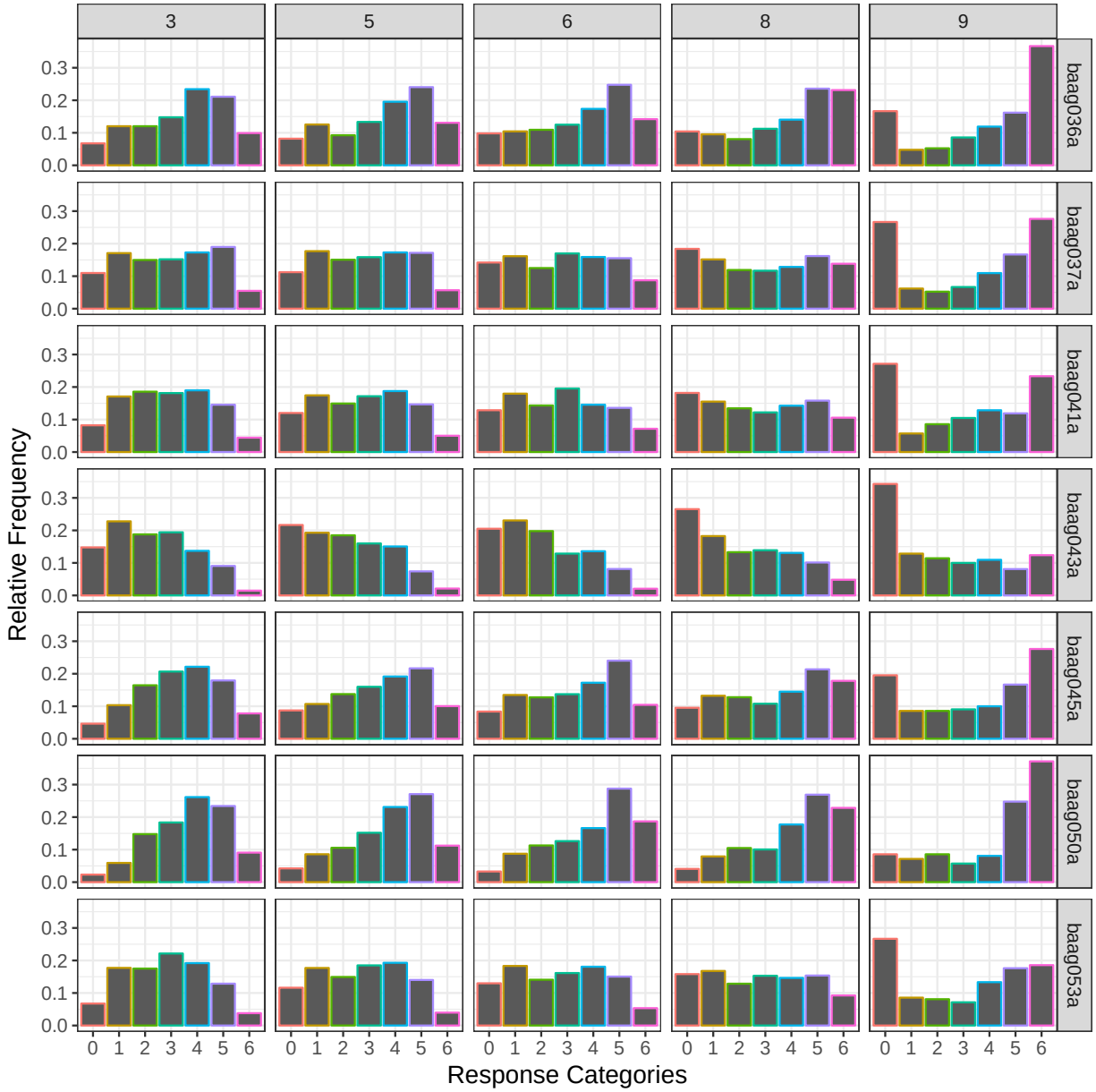

Figure B.24: Bar plots presenting the relative frequencies of response categories within the leaves of the PC tree for the Allocentric/Mental Map (AMM) scale in the GESIS dataset. Rows represent the items of the scale, labeled according to the official dataset. Columns represent the leaves of the respective PC tree. Numbers correspond to the node numbers in the leaves of Figure 2.9. Response frequencies sum up to one in each cell.

B.5 Threshold Plots of PC Trees (NEO-PI-R)

Threshold plots of PC trees in the NEO-PI-R dataset for which no thresholds were missing in any of the resulting leafs.

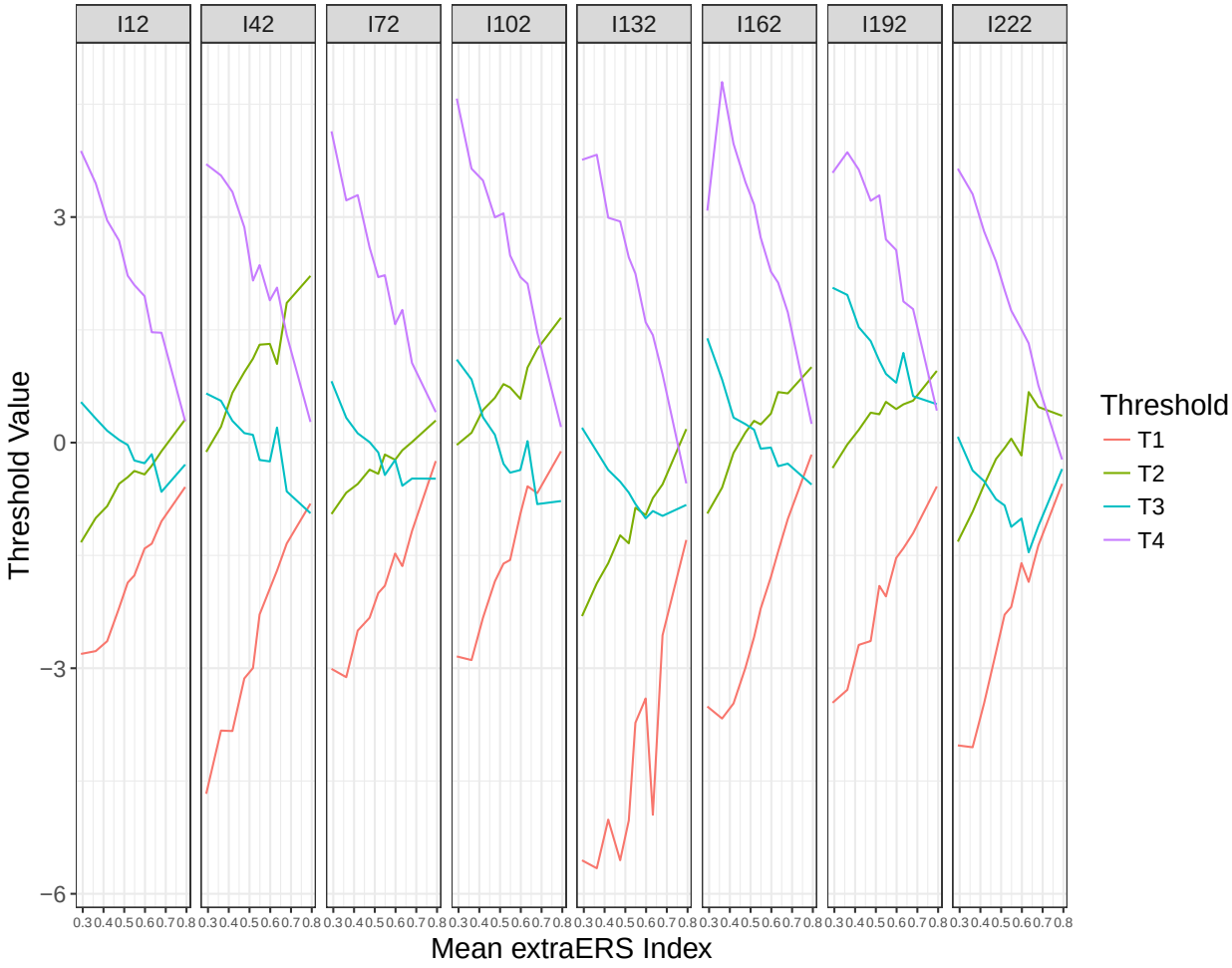

Figure B.25: Threshold plot for the PC tree of the NEO-PI-R facet Assertiveness (E3) with the extraERS index from heterogeneous items as single covariate. Mean extraERS values for each leaf of the PC tree in Figure B.6 are plotted against parameter estimates of the respective partial credit model. Items are shown in separate panels. Thresholds are depicted as lines. $N = 11575$.

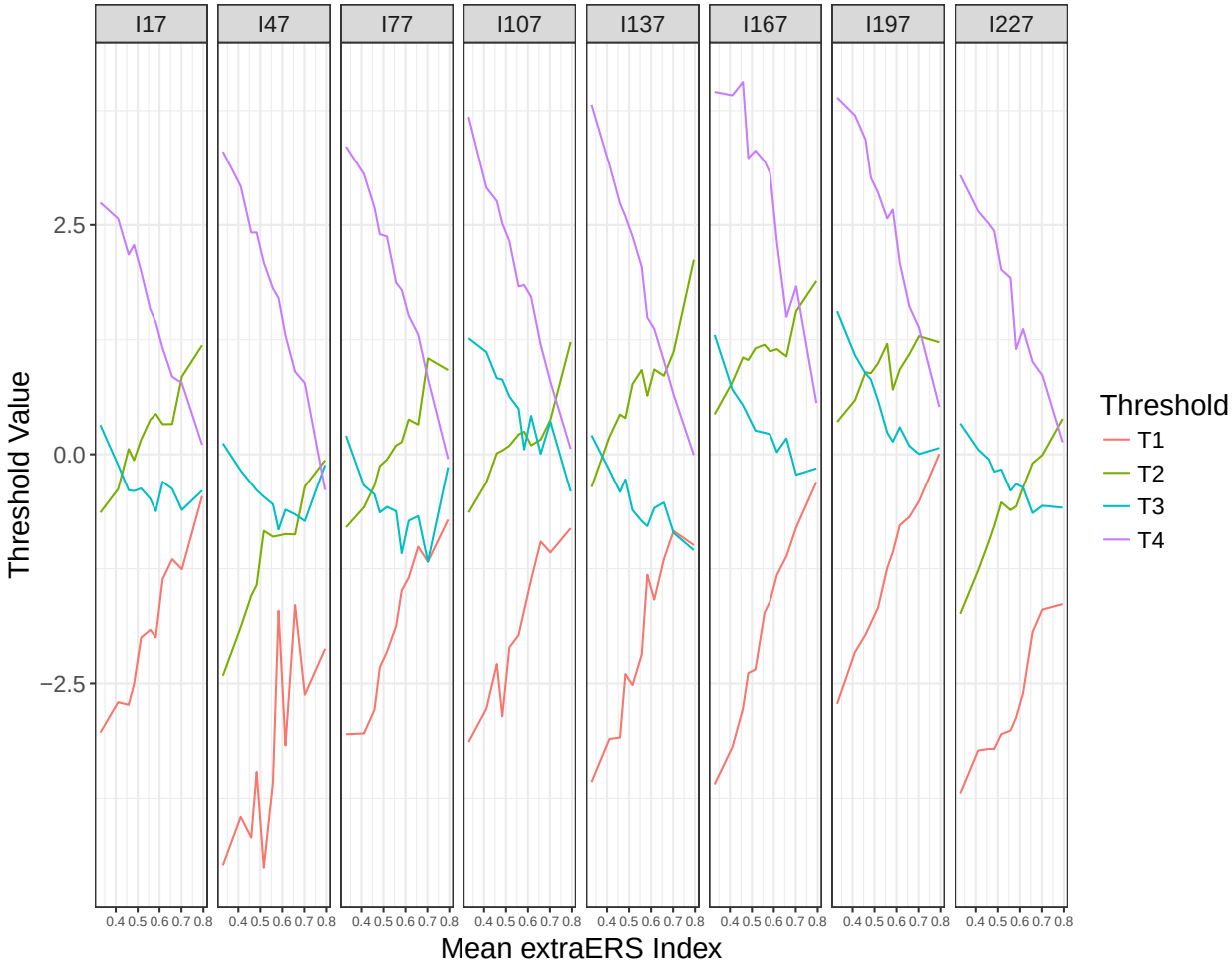

Figure B.26: Threshold plot for the PC tree of the NEO-PI-R facet Activity (E4) with the extraERS index from heterogeneous items as single covariate. Mean extraERS values for each leaf of the PC tree in Figure B.7 are plotted against parameter estimates of the respective partial credit model. Items are shown in separate panels. Thresholds are depicted as lines. $N = 11559$.

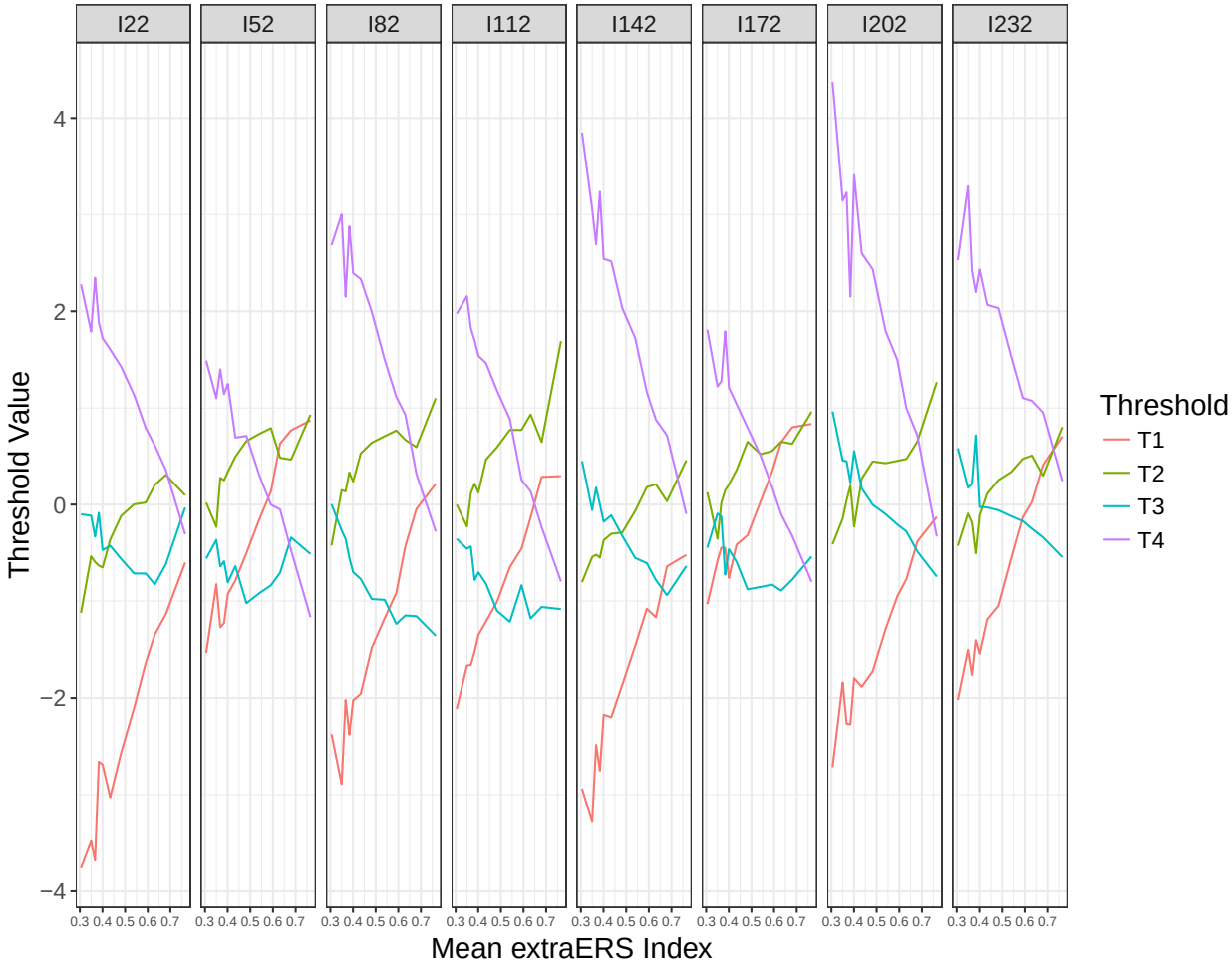

Figure B.27: Threshold plot for the PC tree of the NEO-PI-R facet Excitement-Seeking (E5) with the extraERS index from heterogeneous items as single covariate. Mean extraERS values for each leaf of the PC tree in Figure B.8 are plotted against parameter estimates of the respective partial credit model. Items are shown in separate panels. Thresholds are depicted as lines. $N = 11584$.

B.6 PC Trees Including Demographic Variables (NEO-PI-R and GESIS)

PC trees for the analyzed personality facets in the NEO-PI-R and GESIS datasets, with the respective ERS index from heterogeneous items, sex, and age as covariates.

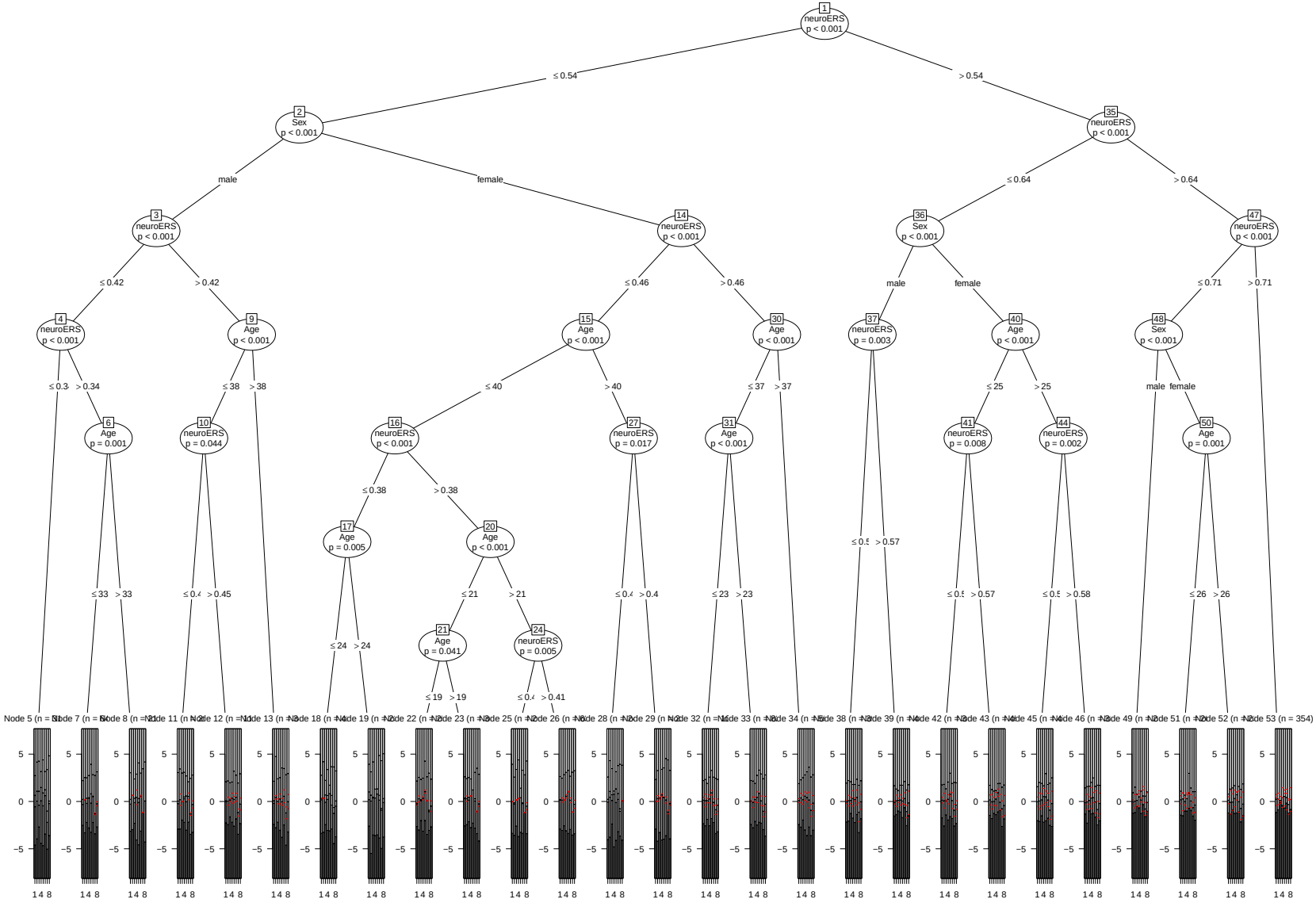

Figure B.28: PC tree of the NEO-PI-R facet Angry Hostility (N2) with the neuroERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11536$.

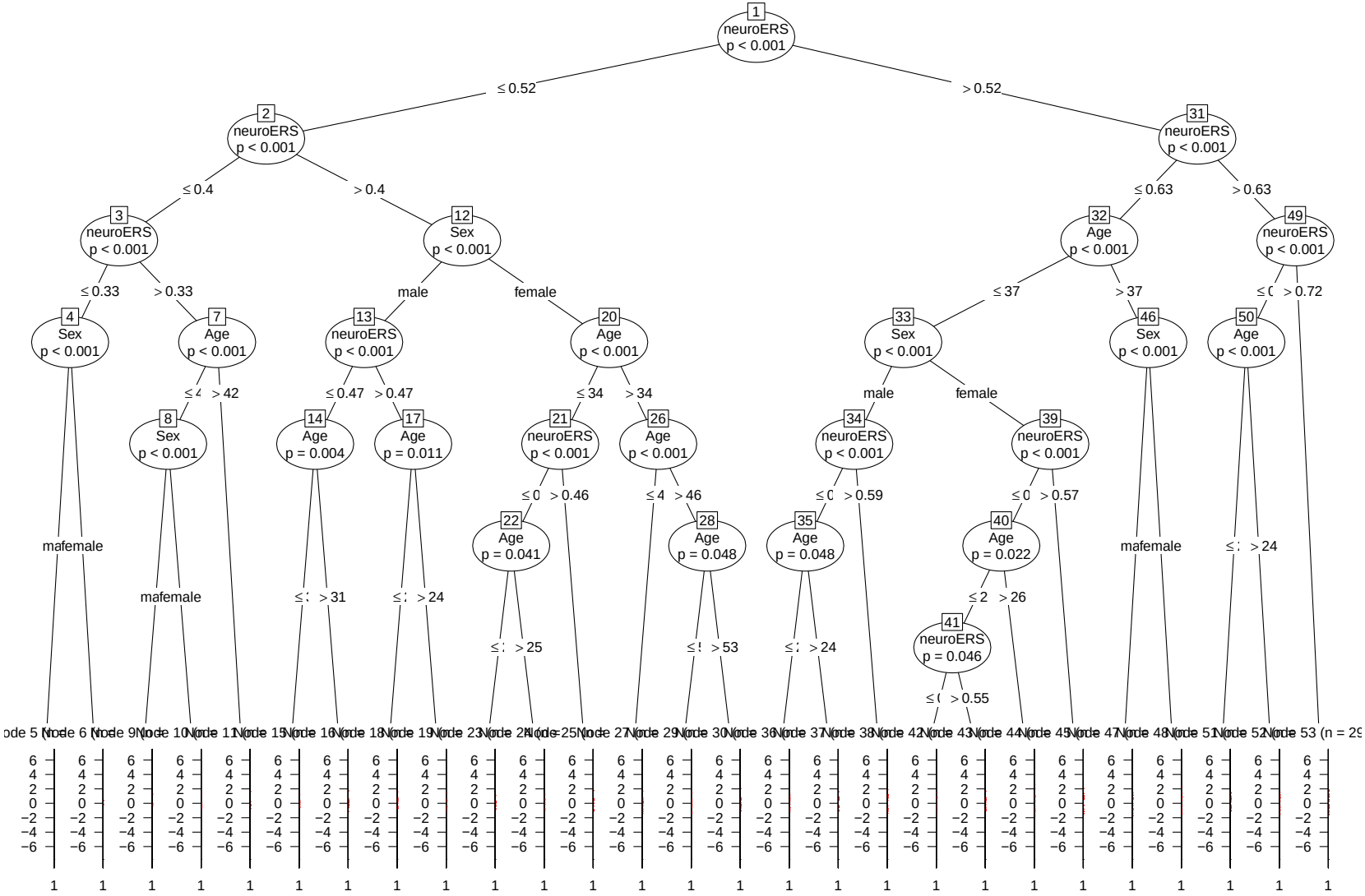

Figure B.29: PC tree of the NEO-PI-R facet Impulsivity (N5) with the neuroERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11565$.

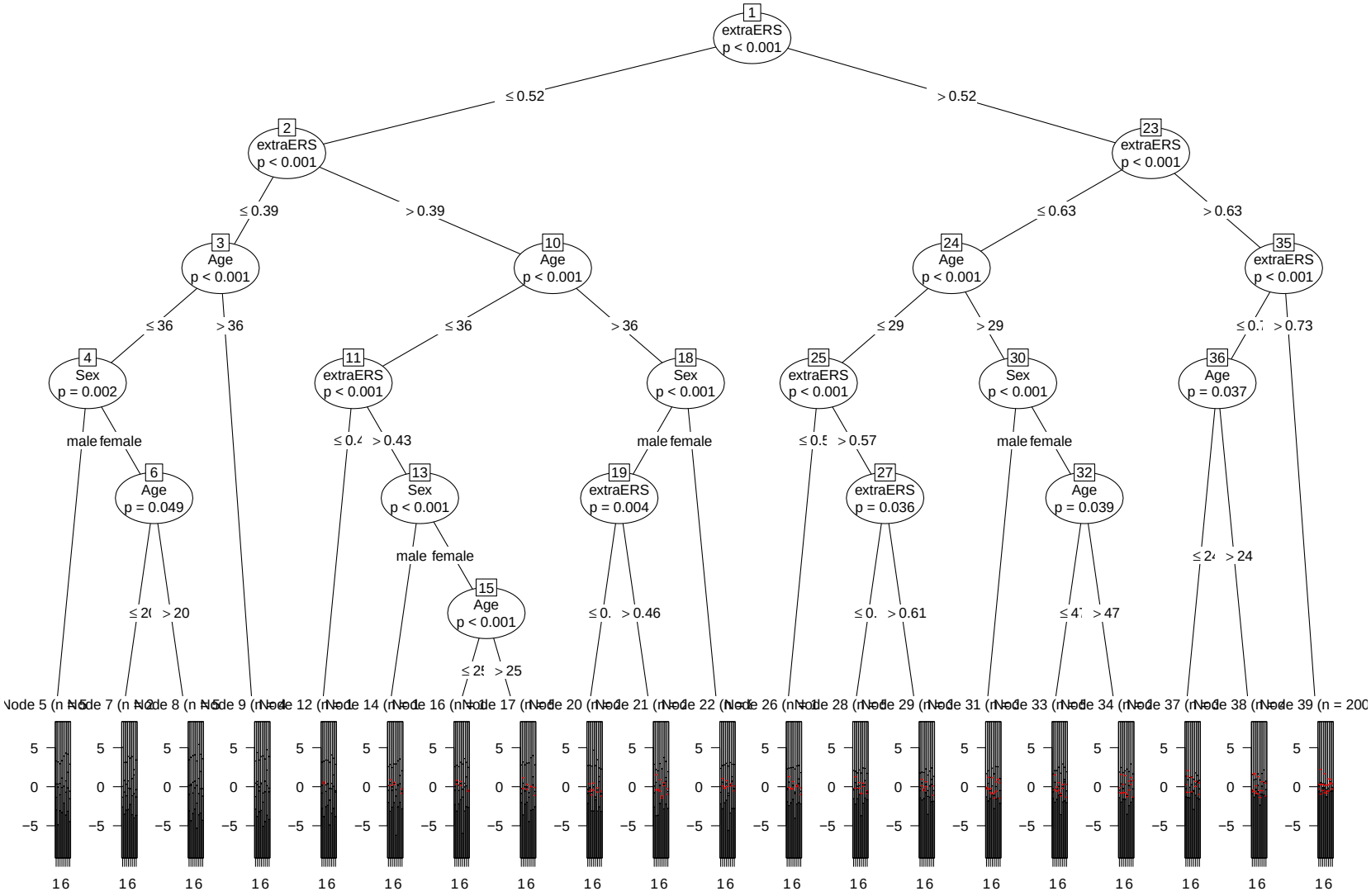

Figure B.30: PC tree of the NEO-PI-R facet Assertiveness (E3) with the extraERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11575$.

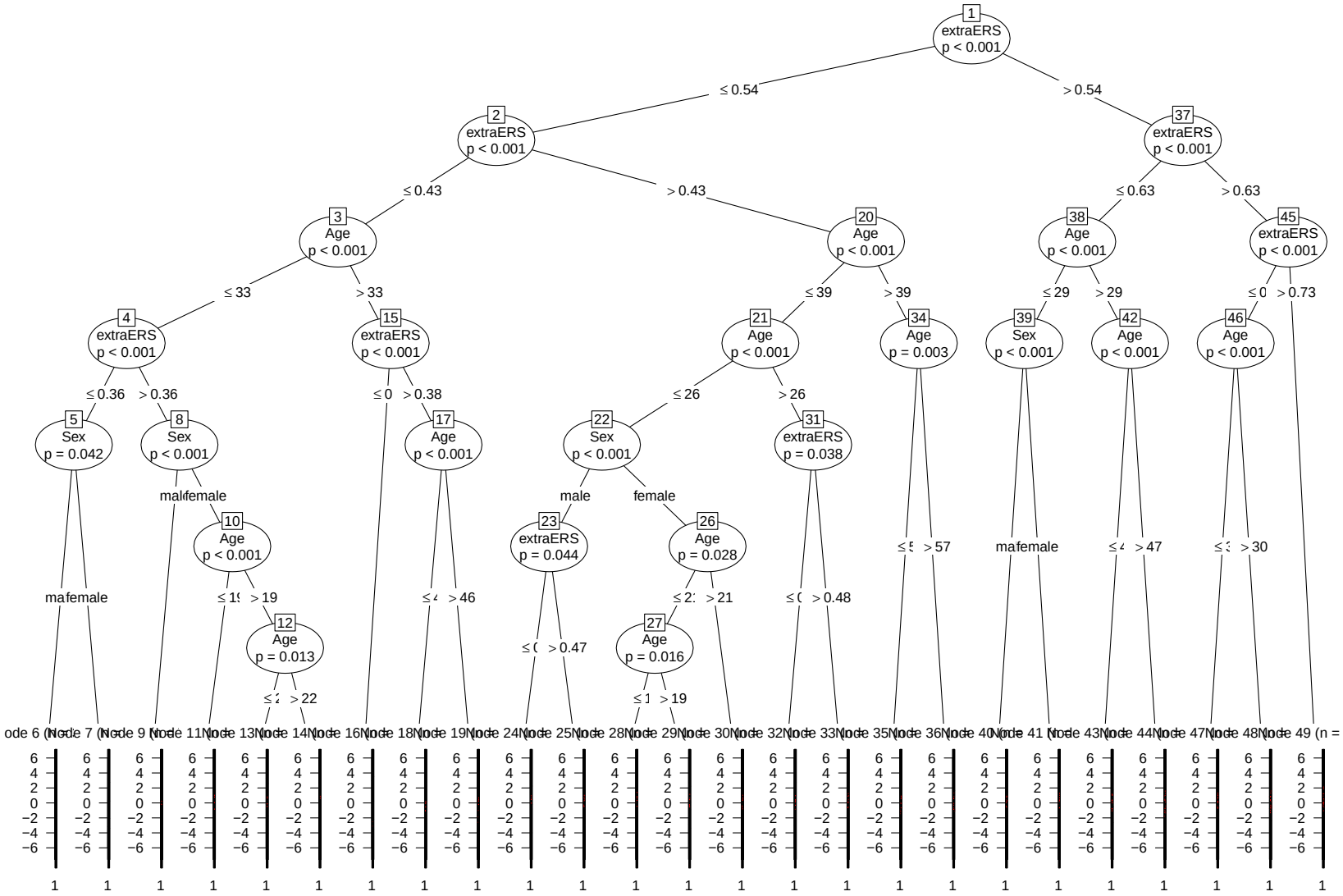

Figure B.31: PC tree of the NEO-PI-R facet Activity (E4) with the extraERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11559$.

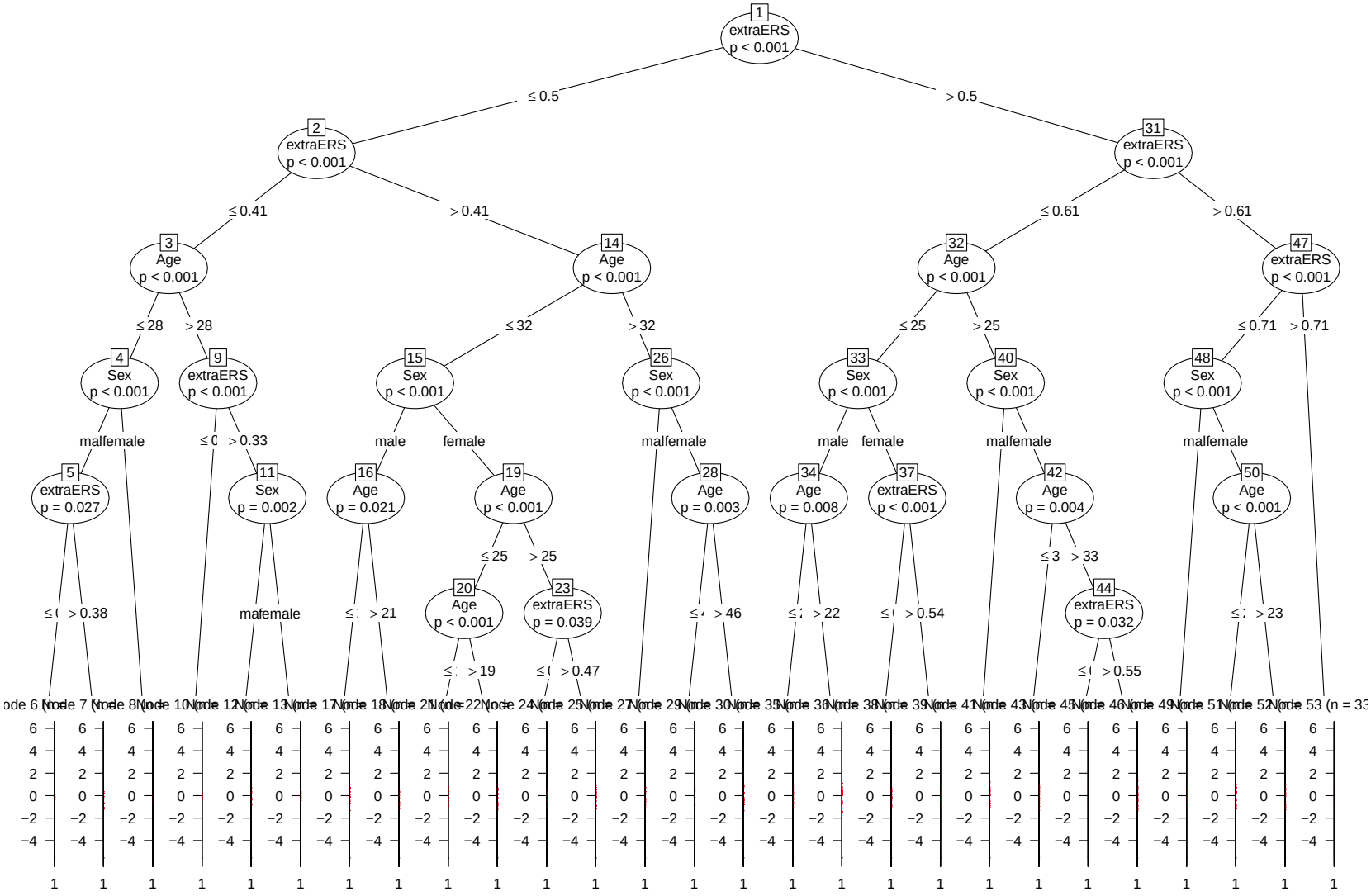

Figure B.32: PC tree of the NEO-PI-R facet Excitement-Seeking (E5) with the extraERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11584$.

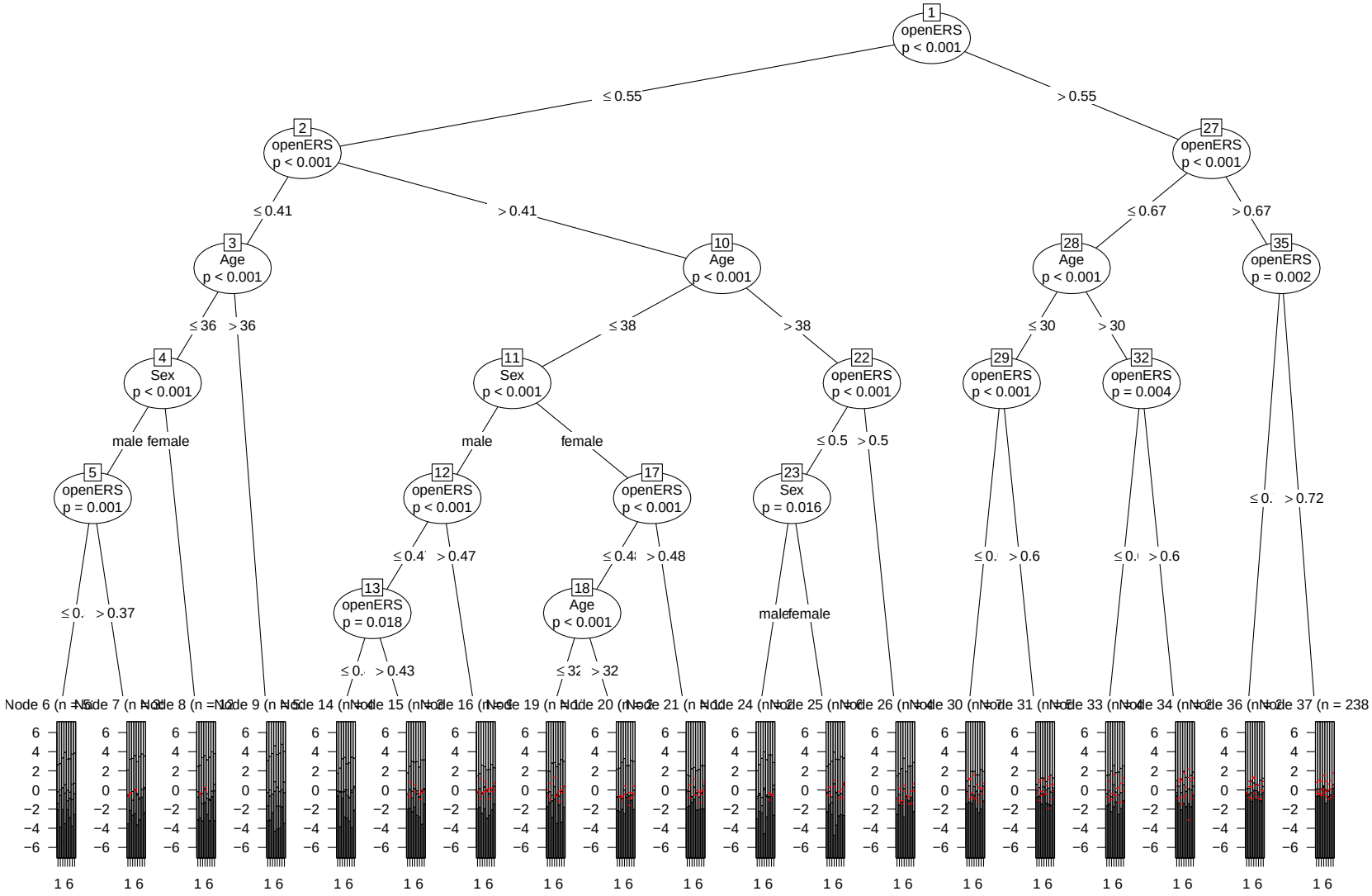

Figure B.33: PC tree of the NEO-PI-R facet Openness to Feelings (O3) with the openERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11578$.

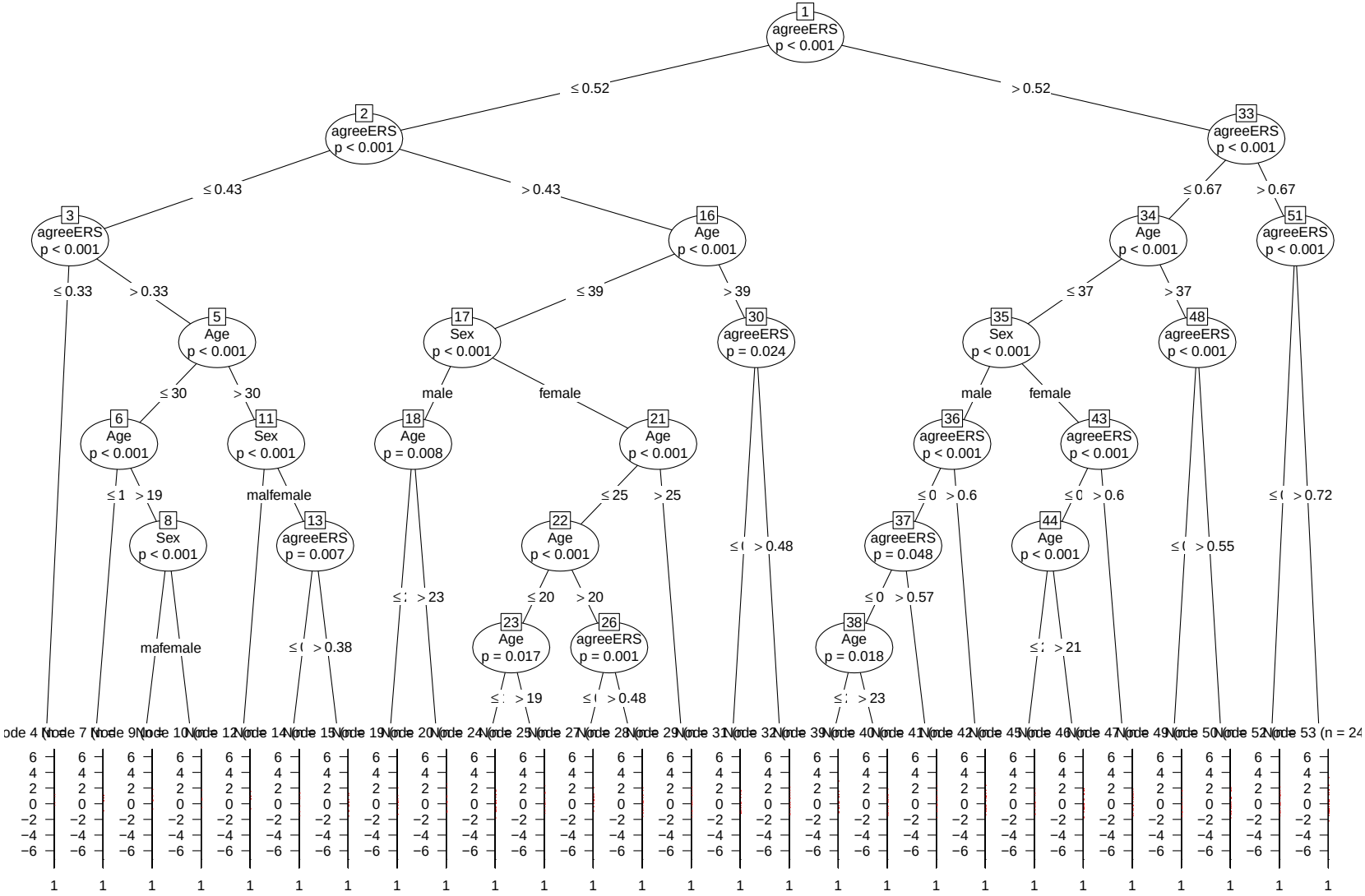

Figure B.34: PC tree of the NEO-PI-R facet Compliance (A4) with the agreeERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11583$.

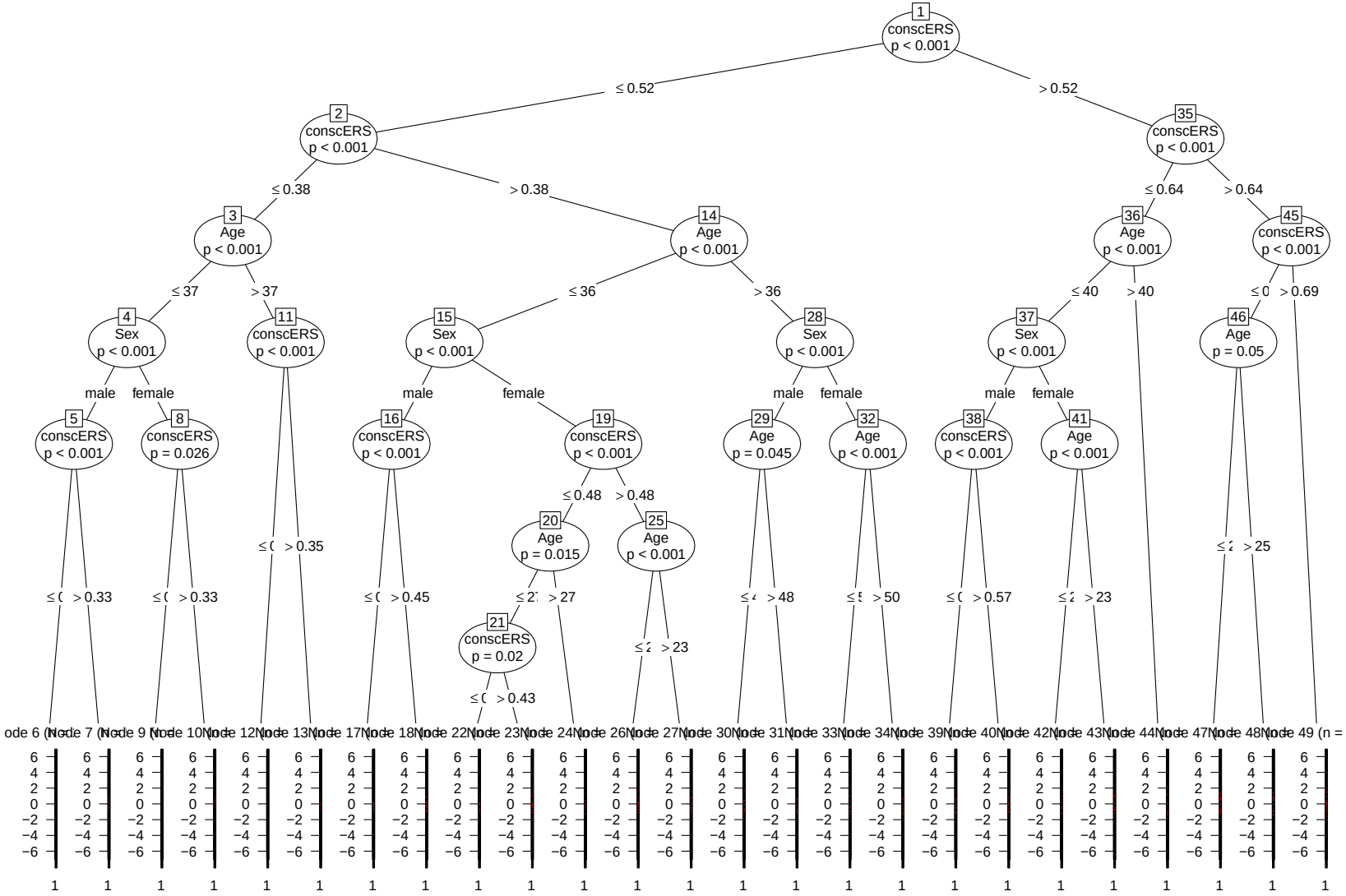

Figure B.35: PC tree of the NEO-PI-R Order (C2) with the conscERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11611$.

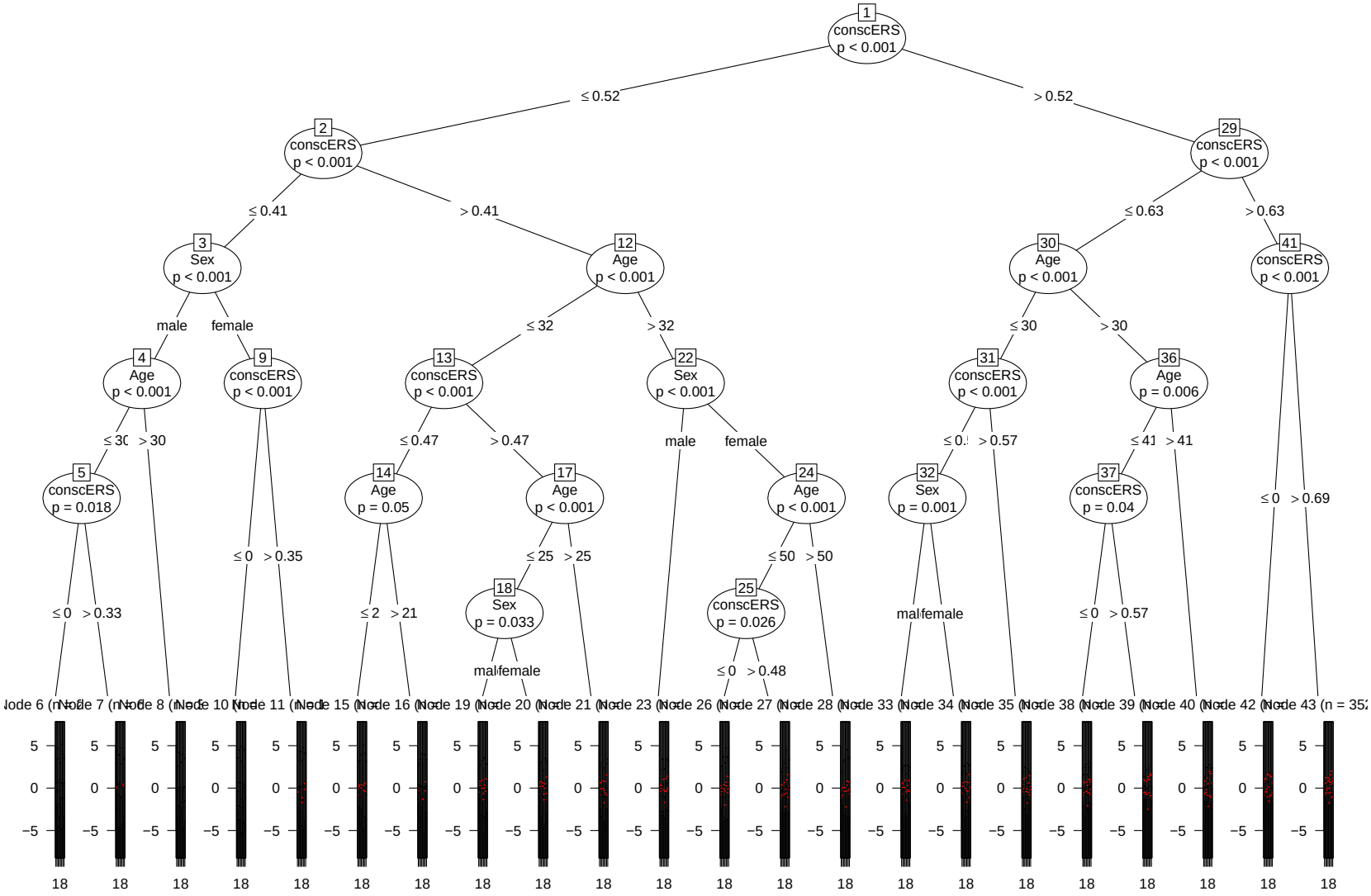

Figure B.36: PC tree of the NEO-PI-R Self-Discipline (C5) with the conscERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 11563$.

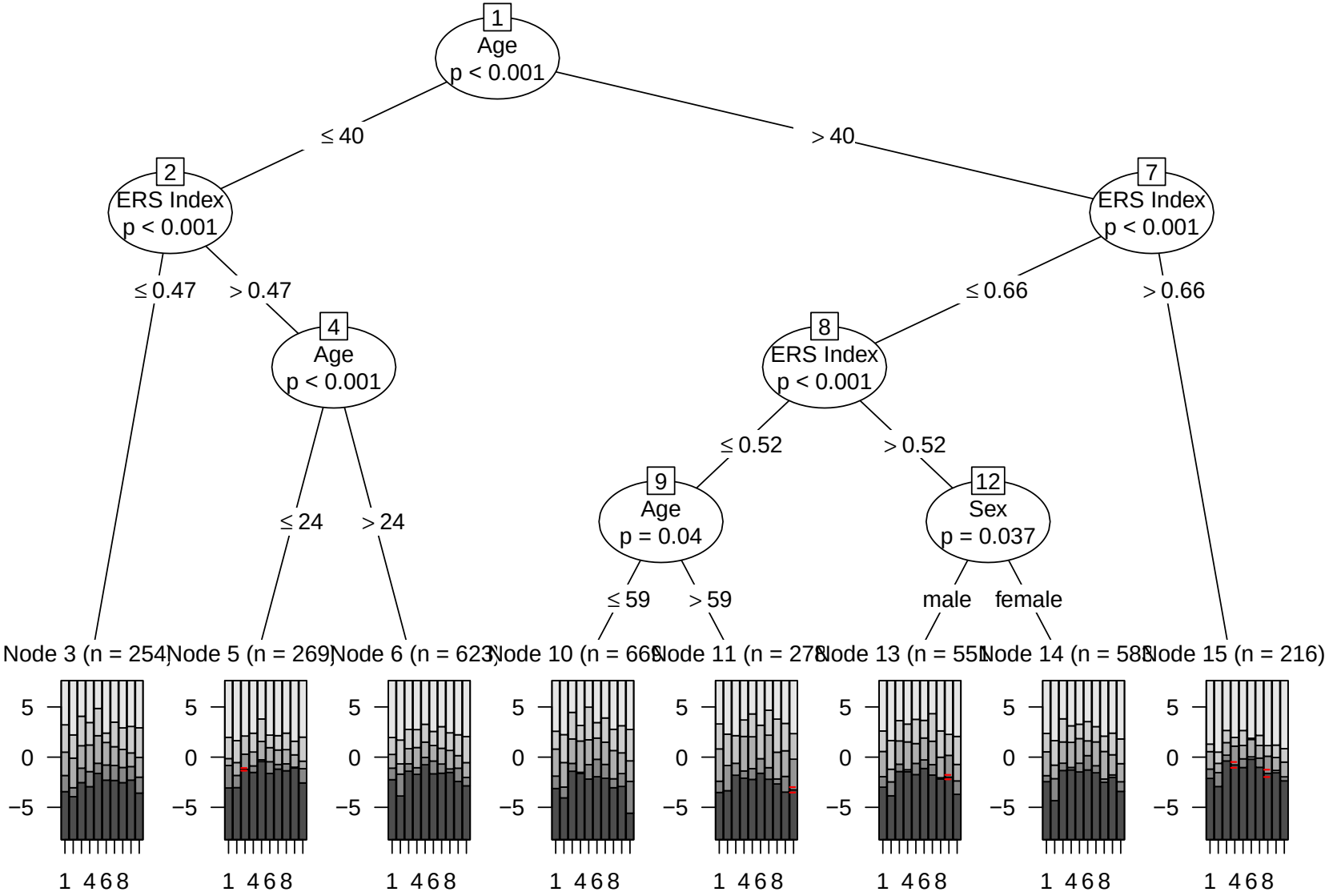

Figure B.37: PC tree of the Positive Affect scale (PA) with the ERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 3443$.

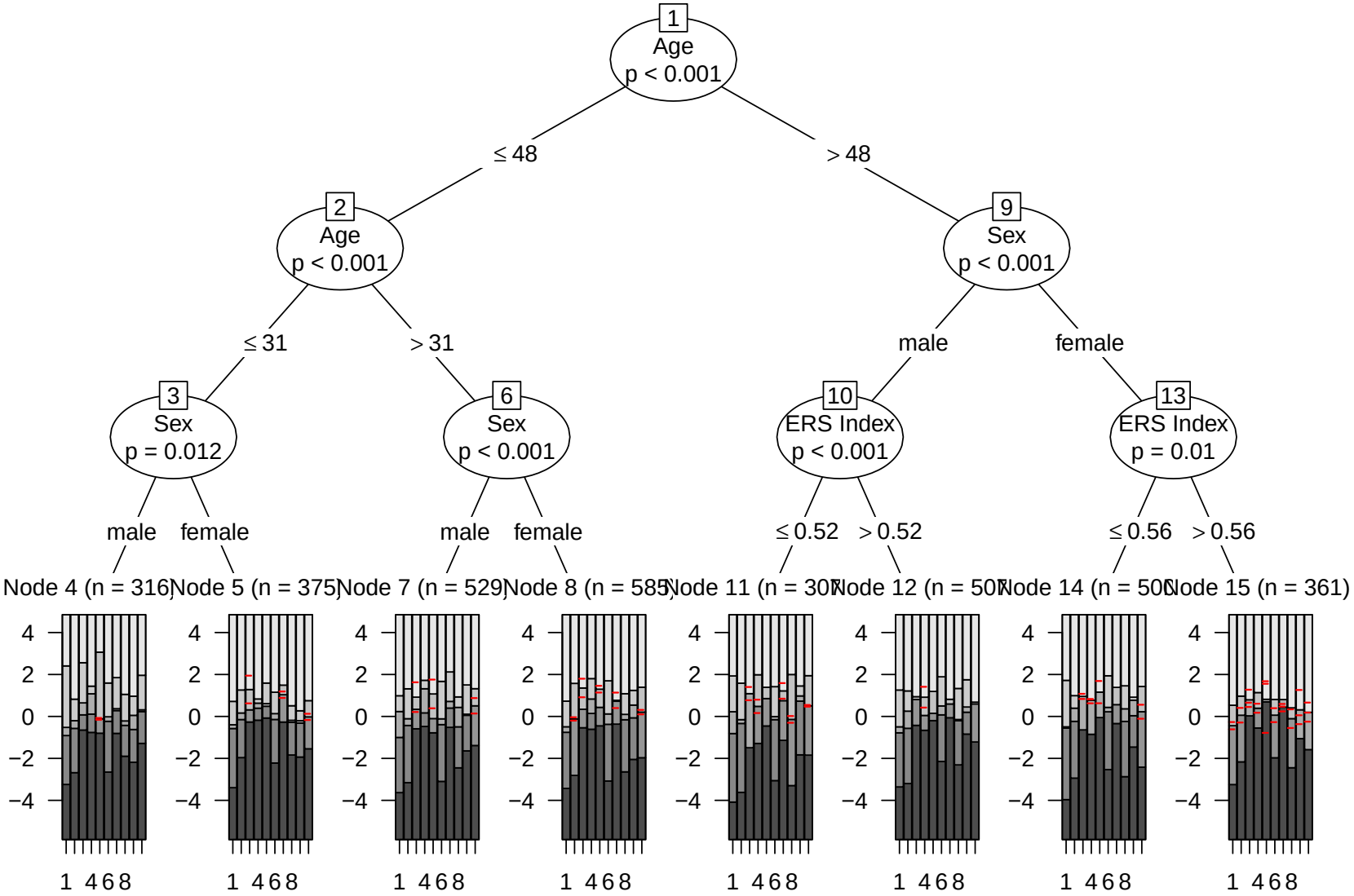

Figure B.38: PC tree of the Negative Affect scale (NA) with the ERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 3480$.

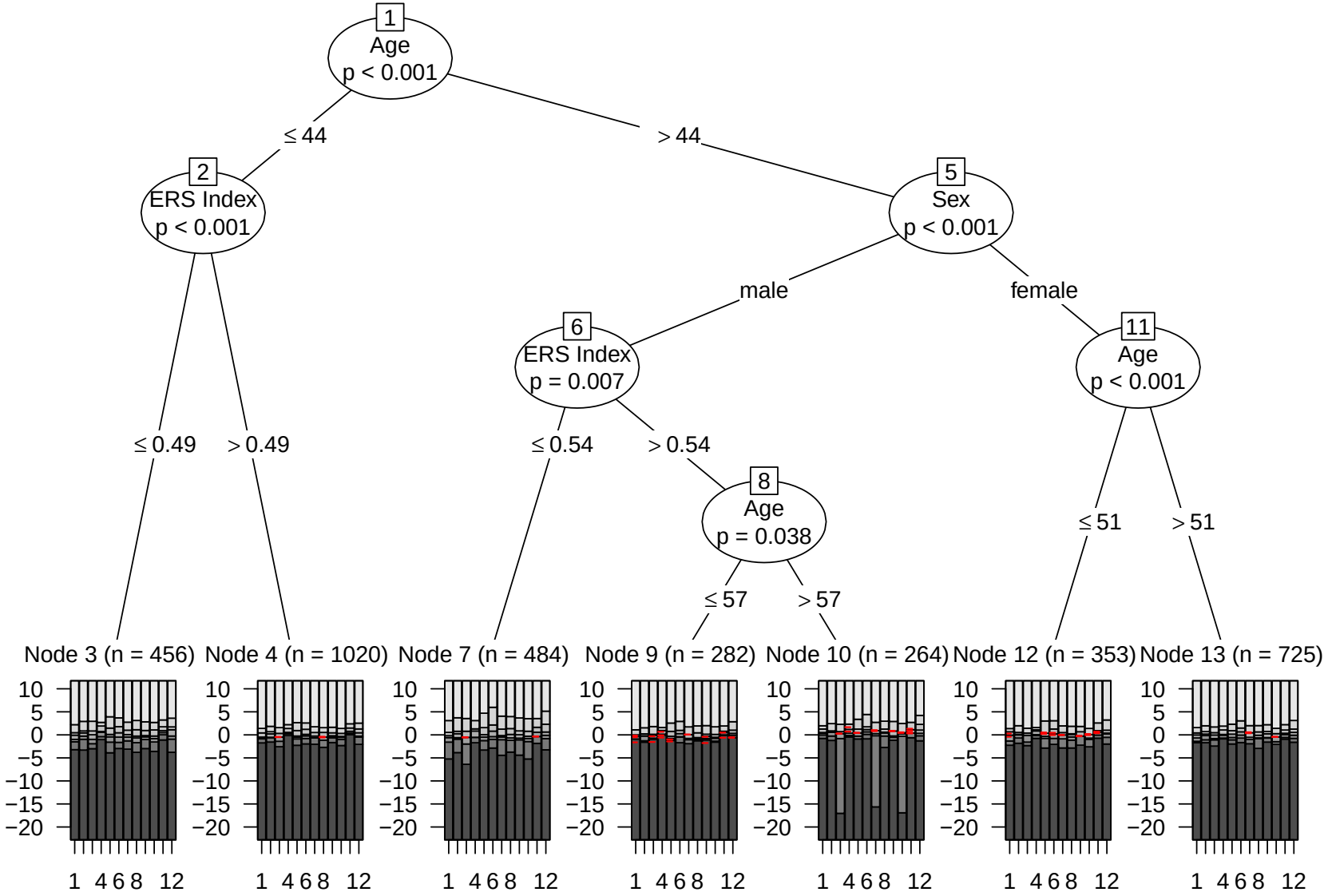

Figure B.39: PC tree of the Global/Egocentric Orientation scale (GEO) with the ERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 3584$.

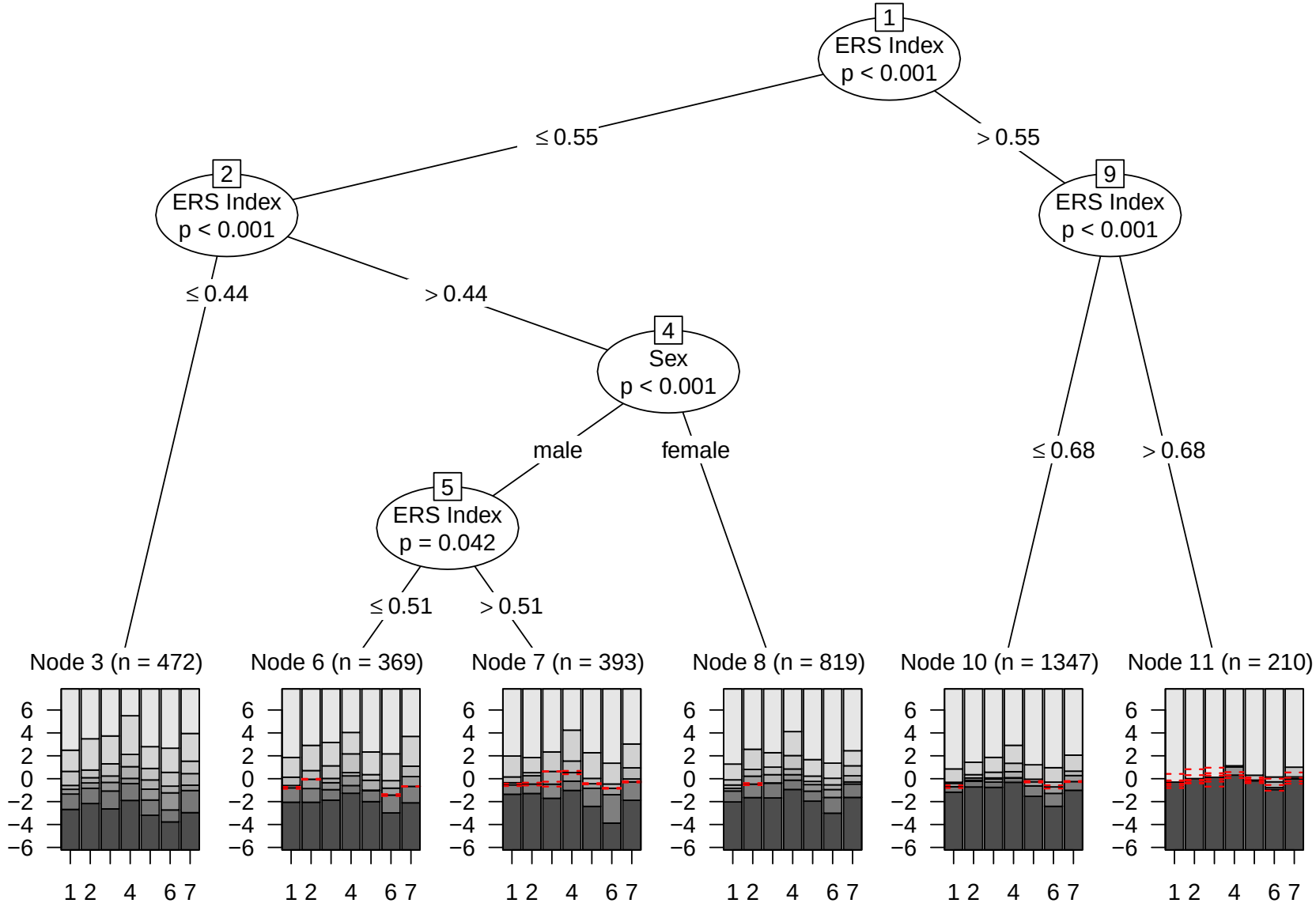

Figure B.40: PC tree of the Allocentric/Mental Map scale (AMM) with the ERS index from heterogeneous items, sex, and age as covariates. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. $N = 3610$.

Appendix C

Supplemental Material for Study 2

C.1 Description of ERS Items

Wording of the items which were used to compute the ERS index from heterogeneous items. Items are presented as they were used in study 2. Original wording differed in some cases. ERS items were sampled from the "Zusammenstellung sozialwissenschaftlicher Items und Skalen des GESIS - Leibniz-Institut für Sozialwissenschaften" (ZIS) and the "Elektronisches Testarchive des Leibniz-Zentrums für Psychologische Information und Dokumentation" (ZPID).

	Item Wording	Original Source
1	Ich habe Schwierigkeiten, mein Verhalten an verschiedene Menschen bzw. an verschiedene Situationen anzupassen.	Schyns, B. & Paul, T. (2014). Deutsche Self-Monitoring Skala. ZIS.
2	Ich bin jemand, der eine lebhafte Phantasie, Vorstellungen hat.	Schupp, J., & Gerlitz, J.-Y. (2014). Big Five Inventory-SOEP (BFI-S). ZIS.
3	Selbst wenn mein Partner ärgerlich auf mich ist, liebe ich ihn stark und bedingungslos.	Bierhoff, H.W. & Grau, I. (1997). Globalskalen zur Einschätzung von Beziehungseinstellungen (GSEB). ZPID.
4	Ich habe, innerhalb der letzten 4 Wochen, in Situationen anders gehandelt, als ich es ursprünglich geplant hatte.	Bankstahl, U.S. & Görtelmeyer, R. (2013). Attention and Performance Self-Assessment (APSA) - deutsche Fassung. ZPID.
5	Für wie wichtig halten Sie persönlich gute Aufstiegsmöglichkeiten für die berufliche Arbeit und den Beruf?	Zentralarchiv für empirische Sozialforschung (ZA) & Zentrum für Umfragen, Methoden und Analysen (ZUMA) e.V. (2014). Wichtigkeit verschiedener Berufsmerkmale. ZIS.

6	Wie wichtig ist Rücksichtslosigkeit, Härte damit jemand in unserer Gesellschaft Erfolg hat und sozial aufsteigt?	Sandberger, J. U. (2014). Aufstiegsmobilität. ZIS.
7	Ich schätze meine Fähigkeiten als hoch ein.	von Collani, G. & Schyns, B. (2014) Generalisierte Selbstwirksamkeitserwartung. ZIS.
8	Für wie wahrscheinlich schätzen Sie es ein, dass Ihnen während einer Flugreise der Magen knurrt?	Mühlberger, A., Herrmann, M.J. & Pauli, P. (1996). Gefahrenerwartungsfragebogen bei Flugreisen (GES). ZPID.
9	Es ist nur natürlich und richtig, wenn jeder seine Familie für besser hält als jede andere.	Lederer, G. (2014). Respekt für elterliche Autorität. ZIS.
10	Wie sehr beunruhigen, belasten oder stören Sie momentan Probleme mit Freunden und Nachbarn.	Jaekel, J. & Leyendecker, B. (2008). Everyday Stressors Index (ESI) - deutsche Fassung. ZPID.
11	Ich vergesse den Namen einer Person fast sofort nachdem er mir erstmals gesagt wurde.	Michalak, J., Heidenreich, T., Ströhle, G. & Nachtigall, C. (2008). Mindful Attention and Awareness Scale (MAAS) - deutsche Version. ZPID.
12	Ich weiß, was ich tun muss, um meine Karriereziele zu erreichen.	Rowold, J. (2004). Karriereplanung (KP). ZPID.
13	Wenn Sie an Ihre eigenen politischen Ansichten denken, wo würden Sie diese Ansichten auf einer Skala von links nach rechts einstufen?	Breyer, B. (2015). Left-Right Self-Placement (ALLBUS) - deutsche Fassung. ZIS.
14	Meine Eltern waren für mich da, wenn ich Probleme hatte.	Bühler, K.-E. (2014). Biographischer Fragebogen für Alkoholabhängige (BIFA-AL). ZIS.
15	Bei der Vorbereitung für eine wichtige Prüfung oder einen wichtigen Vortrag kann ich mich nicht lange auf den Lernstoff konzentrieren und schweife immer wieder ab.	Glöckner-Rist, A., Westermann, S., Engberding, M., Höcker, A., & Rist, F. (2014). Gründe für das Aufschieben von Prüfungslernen (GAP). ZIS.
16	Pessimisten sind Menschen, die voller Zweifel in die Zukunft blicken und meistens Schlechtes erwarten. Wie pessimistisch sind Sie im Allgemeinen?	Kemper, C.J., Beierlein, C., Kovaleva, A. & Rammstedt, B. (2012). Skala Optimismus-Pessimismus-2 (SOP2). ZPID.

17	Mit Grammatik in Englisch habe ich Schwierigkeiten.	Jensen, L. (2012). Skala zur Erfassung des Selbstkonzepts eigener Grammatikkompetenz in der ersten Fremdsprache Englisch (gramSK-L2E). ZPID.
18	In der letzten Woche setzte ich mich mit meinen Gefühlen auseinander.	Berking, M. & Znoj, H. (2008). Fragebogen zur standardisierten Selbsteinschätzung emotionaler Kompetenzen (SEK-27). ZPID.
19	Mir kommen regelmäßig Bilder in den Sinn, in denen ich mich als krank sehe.	Bailer, J. & Witthöft, M. (2014). Deutsches modifiziertes Health Anxiety Inventory (MK-HAI). ZIS.
20	Heutzutage kann man sich auf niemanden mehr verlassen.	Beierlein, C., Kemper, C., Kovaleva, A., J. & Rammstedt, B. (2014). Interpersonales Vertrauen (KUSIV3). ZIS.
21	Als Kind konnte ich kaum vorhersehen, ob meine Eltern sich freuen oder ärgern würden über etwas, was ich getan hatte.	Lederer, G. (2014). Autoritäre Familienstruktur. ZIS.
22	Wenn ich an Bankomaten denke, werde ich nervös.	Sinkovics, R. (2014). Technophobie-Skala. ZIS.
23	Mit dem Tod verliere ich mein Wesen.	Klug, A. (1997). Einstellungen zu Sterben, Tod und Danach (FESTD). ZPID.
24	So wie die Zukunft aussieht, kann man es kaum noch verantworten, Kinder auf die Welt zu bringen.	Zentralarchiv für empirische Sozialforschung (ZA) & Zentrum für Umfragen, Methoden und Analysen (ZUMA) e.V. (2014). Anomie (ALLBUS). ZIS.
25	Wenn ich so auf mein bisheriges Leben zurückblicke, bin ich zufrieden.	Dalbert, C. (1992). Habituelle subjektive Wohlbefindensskala (HSWBS). ZPID.
26	Wie gut fühlen Sie sich persönlich informiert über die Antibabypille.	Münch, K., Hübner, M., Reinecke, J., & Schmidt, P. (2014). Informationsgrad Sexualität und Verhütung. ZIS.
27	Die Sozialleistungen in Deutschland machen die Menschen faul.	Giza, A. & Scheuer, A. (2014). Bewertung der Sozialpolitik. ZIS.
28	Manchmal bin ich energielos.	Sellin, I., Schütz, A., Kruglanski, A.W. & Higgins, E.T. (2003). Locomotion-Assessment-Fragebogen (L-A-F). ZPID.
29	Jeder Mensch sollte etwas von seiner Zeit für das Wohl seiner Stadt oder Gemeinde aufbringen.	Bierhoff, H.W. (2000). Skala der sozialen Verantwortung (SV). ZPID.

30	Ich bin mir beim Kauf von Elektrogeräten darüber bewusst, wie umweltverträglich das jeweilige Gerät ist.	Hunecke, M., Blöbaum, A., Matthies, E., & Höger, R. (2014). Bewusstheit von Handlungskonsequenzen: global. ZIS.
31	Ich bin aggressiv.	Krahé, B., Berger, A. & Möller, I. (2007). Instrument zur Erfassung des Geschlechtsrollen-Selbstkonzepts im Jugendalter (GRI-JUG). ZPID.
32	Wie oft wurde bei Ihnen zu Hause gemeinsam gesungen bevor Sie 10 Jahre alt waren.	Lüdtke, H. & Neumann, P. (2014). Musikalische Frühsozialisation. ZIS.
33	In den vergangenen Jahren hatte ich ein einschneidendes Lebensereignis, an das ich häufig zurückdenke.	Linden, M., Baumann, K., Lieberei, B. & Rotter, M. (2009). Post-Traumatic Embitterment Disorder Selbstbeurteilungsfragebogen (PTED). ZPID.
34	Die PolitikerInnen haben die Pflicht, durch eine Förderung des Radverkehrs die Luftverschmutzung zu verringern.	Martens, T., Rost, J., & Gresele, C. (2014). Verantwortung für Umweltprobleme. ZIS.
35	Ich versuche, durch starke körperliche Betätigung mein Gewicht zu kontrollieren.	Böhm, B. (1993). Essverhalten-Test (EVT). ZPID.
36	Ich finde, dass mir bei wichtigen Entscheidungen im Großen und Ganzen Gerechtigkeit widerfährt.	Dalbert, C. (1999). Persönliche Gerechte-Welt-Skala (GWPER). ZPID.
37	Vorehelicher Geschlechtsverkehr ist etwas ganz Normales.	Hebler, M., Booh, A. T., Wieczorek, S., & Schneider, J. F. (2014). Right-Wing Autoritarismus. ZIS.
38	Ich zeige einem Partner nicht gern, wie es tief in mir aussieht.	Neumann, E., Rohmann, E. & Bierhoff, H.W. (2007). Bochumer Bindungsfragebogen (BoBi). ZPID.
39	Ich komme mir wie abgelöst von meinen Gedanken vor, so als ob diese unabhängig von mir existieren.	Michal, M., Sann, U., Niebecker, M., Lazanowsky, C., Kernhof, K., Aurich, S., Overbeck, G., Sierra, M. & Berrios, G.E. (2004). Cambridge Depersonalisation Scale (CDS) - deutsche Trait Fassung. ZPID.
40	Wenn ich von jemandem irgendwann einmal ungerecht behandelt oder verletzt wurde, werde ich es der Person heimzahlen.	Werner, R. & Appel, C. (2014). Deutscher Vergeltungsfragebogen. ZIS.

41	Ich fühle mich eng verbunden mit den Deutschen.	Leszczensky, L. & Gräbs Santiago, A. (2014). Ethnische und nationale Identität von Kindern und Jugendlichen. ZIS.
42	Bei einigen meiner Arbeitsaufgaben muss ich mich richtig darum bemühen, dass ich sie nicht zugunsten attraktiverer Aufgaben unerledigt lasse.	Neubach, B. & Schmidt, K.-H. (2007). Skalen Impulskontrolle, Überwinden von Widerständen, Ablenkungen Widerstehen (IK-ÜW-AW). ZPID.
43	Bewerten Sie die Qualität Ihrer Wohngegend bezüglich der Kinderfreundlichkeit insgesamt.	Schulz, W., Kaindl, M., Waska, B., Leskovich, I., & Rappold, E. (2014). ZIS.
44	An der Regierung soll jederzeit Kritik geübt werden können.	Weiss, H., Gravogl, B., & Oberforster, C. (2014). Civil-Rights (Wien). ZIS.
45	Rechtschreibregeln meiner Muttersprache verstehe ich nur schwer.	Faber, G. (2007). Fragebogen zur Erfassung des rechtschreibbezogenen Selbstkonzepts von Grundschulkindern (rsSK2, revidierte und erweiterte Form 07-3-24). ZPID.
46	Ich werde niemals zufrieden sein, bevor ich nicht alles das bekomme, was mir zusteht.	von Collani, G. (2014). Modifizierte deutsche Versionen des Narcissistic Personality Inventory (NIP-d). ZIS.
47	Abends nach der Arbeit bin ich erschöpft.	Weyer, G., Hodapp, V., & Neuhäuser, S. (2014). Subjektive Zufriedenheit und Belastung von Arbeit und Beruf. ZIS.
48	Wie sehr bemühen Sie sich der Einnahme von Suchtmitteln zu widerstehen?	Nakovics, H., Diehl, A., Geiselhart, H. & Mann, K. (2009). Mannheimer Craving Scale (MaCS). ZPID.
49	Wenn eine neue Mode herauskommt, dann gefallen mir meine älteren Sachen (nach einiger Zeit) nicht mehr so gut.	Scherhorn, G., Haas, H., Hellenthal, F., & Seibold, S. (2014). Gütergebundenheit. ZIS.
50	Ich bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen.	Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C. & Kovaleva, A., (2014). Big Five Inventory (BFI-10). ZIS.

C.2 Relative Response Frequencies for Dichotomous Scales

Observed relative response frequencies for the dichotomous Shyness and Achievement Orientation scales in study 2.

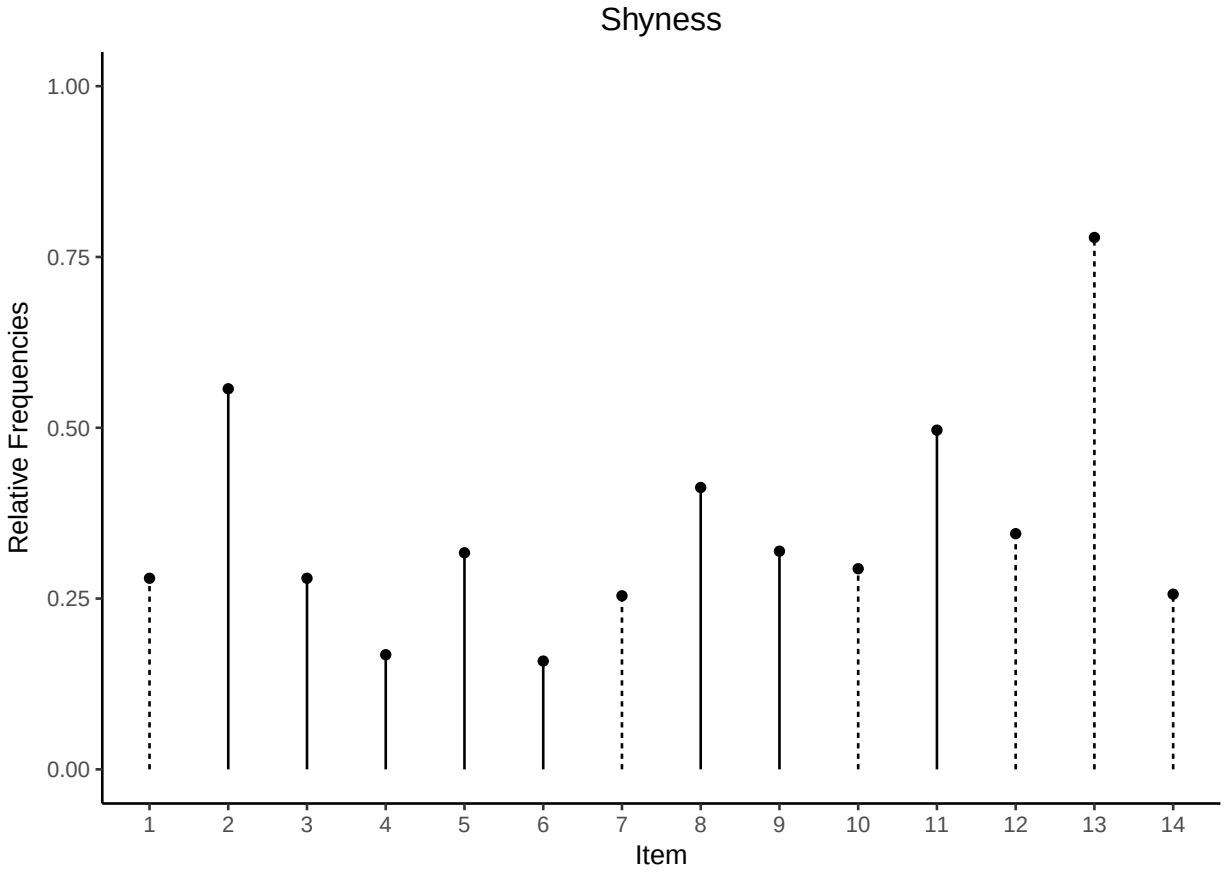

Figure C.1: Relative item response frequencies of Shyness (SHY) scale. An item response of one reflects higher values on the latent variable for all items. Originally negatively keyed items are indicated by dotted lines.

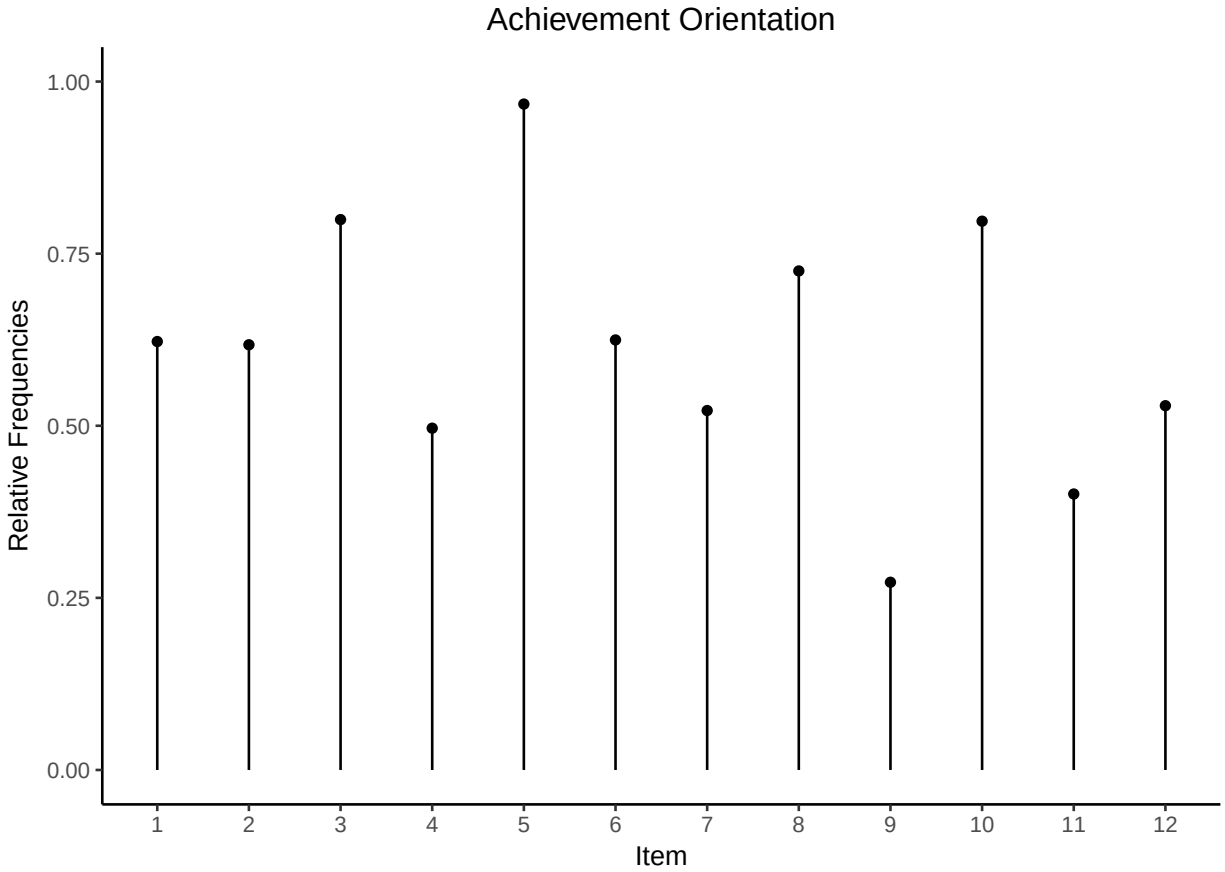

Figure C.2: Relative item response frequencies of the Achievement Orientation (ACHI) scale. An item response of one reflects higher values on the latent variable for all items. In the ACHI scale, all items are positively keyed.

C.3 PC Trees of the Order Scale with MRERS as Single Covariate

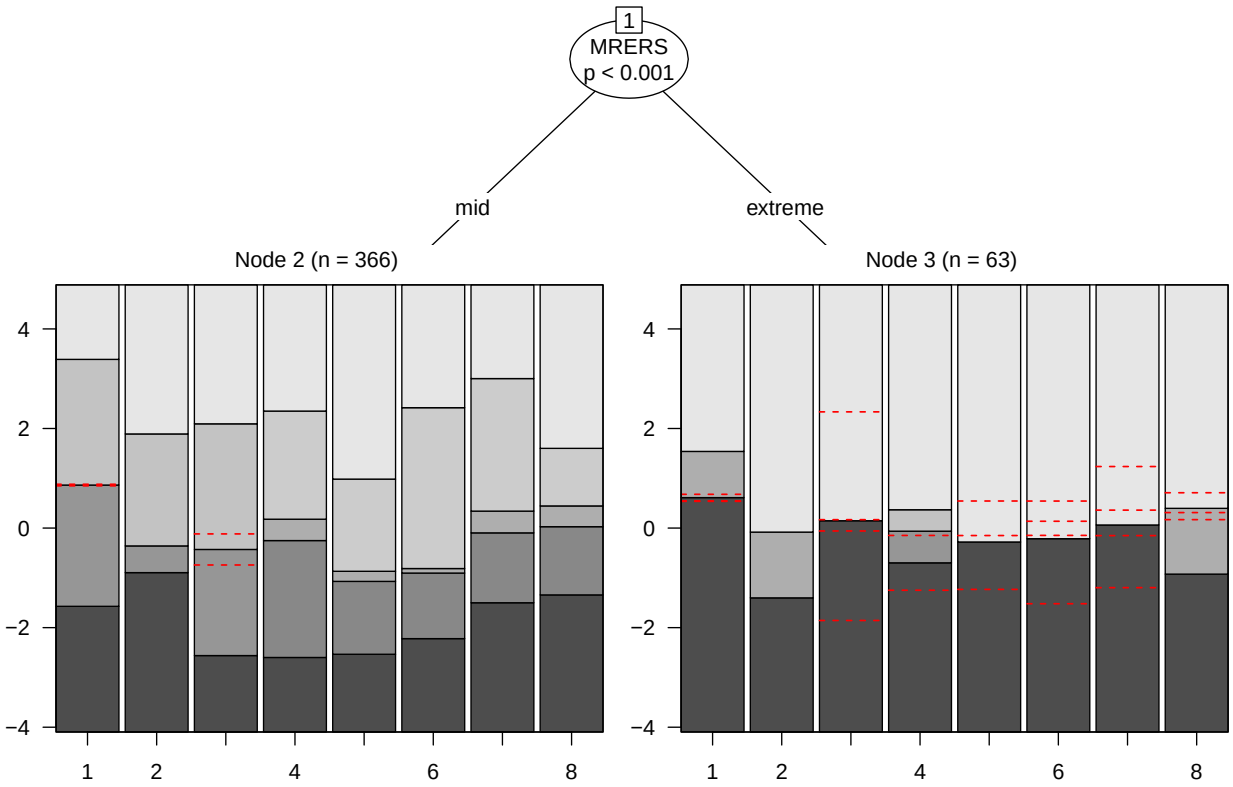

Figure C.3: PC tree of the Order (ORD) scale with MRERS as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. MRERS = extreme response style classified by a constraint mixed Rasch model of the Order scale.

C.4 Lasso Paths of ERS Index, SRERS, and MRERS

Lasso Paths, depicting the estimated DIF parameters γ for the ERS index from heterogeneous items, SRERS, and MRERS, as a function of the regularization parameter λ.

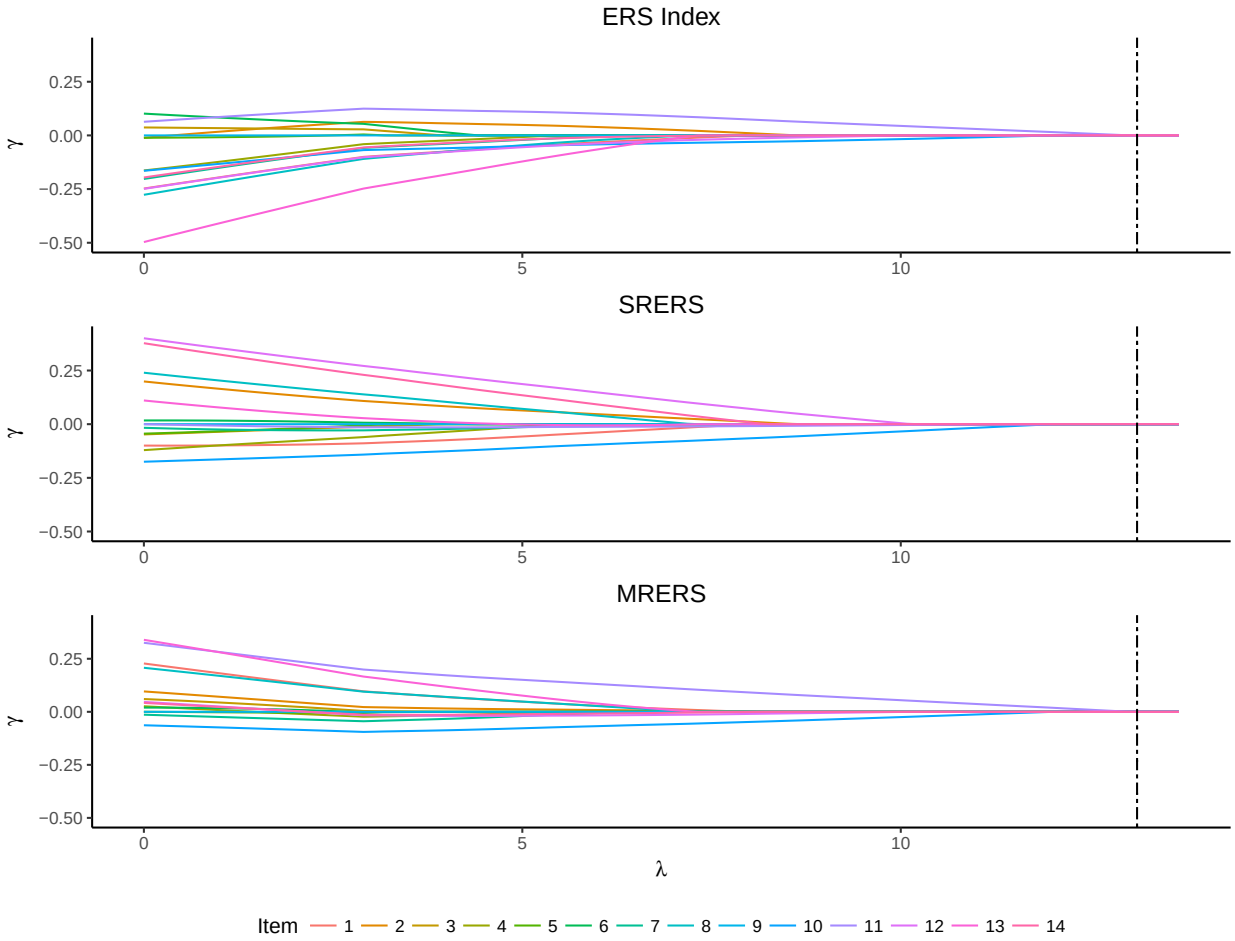

Figure C.4: DIF Lasso Paths of the three extreme response style indices for the Shyness (SHY) scale. The ERS index, SRERS, and MRERS were included as covariates in the model. Optimal values of the regularization parameter λ based on the BIC are indicated by vertical lines. For the dashed line, degrees of freedom were estimated by the L2-norm method whereas the method by Yuan and Lin (2006) was used for the dotted line. SRERS = self-reported extreme response style. MRERS = extreme response style classified by a constraint mixed Rasch model of the Order scale.

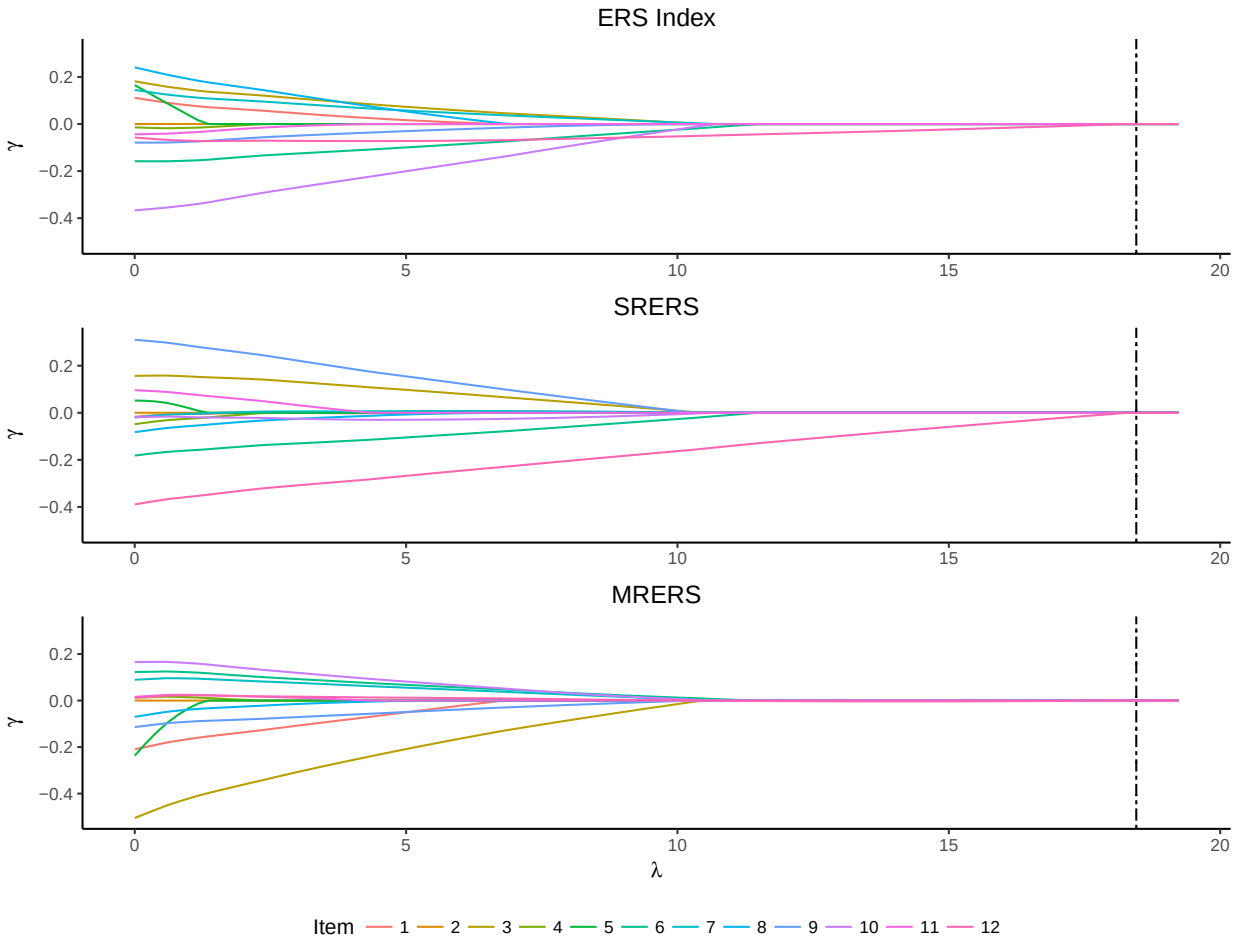

Figure C.5: DIF Lasso Paths of the three extreme response style indices for the Achievement Orientation (ACHI) scale. The ERS index, SRERS, and MRERS were included as covariates in the model. Optimal values of the regularization parameter λ based on the BIC are indicated by vertical lines. For the dashed line, degrees of freedom were estimated by the L2-norm method whereas the method by Yuan and Lin (2006) was used for the dotted line. SRERS = self-reported extreme response style. MRERS = extreme response style classified by a constraint mixed Rasch model of the Order scale.

Appendix D

Supplemental Material for Study 3

D.1 Description of ERS Items

Wording of the items which were used to compute the ERS index from heterogeneous items. Items are presented as they were used in study 3. Original wording differed in some cases. ERS items were sampled from the "Zusammenstellung sozialwissenschaftlicher Items und Skalen des GESIS - Leibniz-Institut für Sozialwissenschaften" (ZIS) and the "Elektronisches Testarchive des Leibniz-Zentrums für Psychologische Information und Dokumentation" (ZPID).

	Item Wording	Original Source
1	Ich halte die Anwendung der Antibabypille zur Schwangerschaftsverhütung für sinnvoll.	Münch, K., Hübner, M., Reinecke, J., & Schmidt, P. (2014). Determinanten der Pilleneinnahme. ZIS.
2	Zuhause ist es gemütlich.	Krämer, L. & Fuchs, R. (2010). Sportbezogenes Barrierenmanagement (SBBM). ZPID.
3	Es fällt mir schwer, meine Gedanken bei einer Aufgabe oder einer Arbeit zu behalten.	Lück, H. & Timaeus, E. (2014). Soziale Erwünschtheit (SDS-E). ZIS.
4	Es ist besser, sein Geld heute auszugeben, als es für das Alter aufzusparen.	Mühleck, K. & Scheller, P. (2014). Gerechtigkeitsideologien mit Bezug zur Altersvorsorge. ZIS.
5	Während einer Flugreise geht mir der Gedanke durch den Kopf, dass das Flugzeug vom Himmel fallen und abstürzen könnte.	Mühlberger, A., Herrmann, M.J. & Pauli, P. (1996). Gefahrenerwartungsfragebogen bei Flugreisen (GES). ZPID.
6	Die Ergebnisse von Umfragen sind in den meisten Fällen richtig.	Stocké, V. (2014). Einstellungen zu Umfragen. ZIS.

7	Wer sich für die Zwecke anderer ausnützen lässt, ohne es zu merken, verdient kein Mitleid.	Ulbrich-Herrmann, M. (2014). Machiavellistische Einstellungen. ZIS.
8	In einer Partnerschaft ist mir Freiraum für eigene Interessen sehr wichtig.	Münch, K., Hübner, M., Reinecke, J., & Schmidt, P. (2014). Bedeutung verschiedener Bereiche in einer Partnerschaft. ZIS.
9	Falls ein einziges Land die Welt regieren sollte, könnte die Bundesrepublik dies besser als alle anderen Nationen.	Lederer, G. (2014). Ausländerablehnung. ZIS.
10	Man muss Dinge entsprechend den Umständen handhaben.	Bierbrauer, G. & Klinger, E.W. (2000). Skala zur Messung Sozialer Axiome (SAS). ZPID.
11	Ich bin dankbar für Personen, die mir genau sagen können, was ich tun soll.	Lederer, G. (2014). Respekt für unspezifische Autorität. ZIS.
12	Ich bin ein Morgenmensch.	Randler, C. (2008). Composite Scale of Morningness (CSM) - deutsche Fassung. ZPID.
13	Ich mag das Gefühl, am Rande eines Abgrundes oder in großer Höhe zu stehen und herunterzuschauen.	Roth, M. & Mayerhofer, D. (2014). Deutsche Version des Arnett Inventory of Sensation Seeking (AISS-d). ZIS.
14	Manchmal verärgere ich mir nahe stehende Personen dadurch, dass ich wiederholt eine Versicherung von ihnen suche, dass ich ihnen wirklich etwas bedeute.	Schwennen, C. & Bierhoff, H. W. (2014). Skala zur exzessiven Bestätigungssuche (BSS). ZIS.
15	Ich habe den Eindruck, dass ich mir über meine inneren Probleme oft Gedanken mache.	Giegler, H. & Schürhoff, R. (2014). Gießen-Test (Kurzform). ZIS.
16	Die Welt ist heute so schwierig geworden, dass ich nicht mehr weiß was los ist.	Gümüs, A., Gömleksiz, M., Glöckner-Rist, A., & Balke, D. (2014). Anomie. ZIS.
17	Ich bin bereit auf Medikamente zu verzichten, deren Nebenwirkung das Krebsrisiko erhöht.	Montada, L., Kals, E., & Becker, R. (2014). Bereitschaften zur persönlichen Krebsvorsorge. ZIS.
18	Ich habe Angst davor zu sterben.	Klug, A. (1997). Einstellungen zu Sterben, Tod und Danach (FESTD). ZPID.
19	Momentan bin ich mit meinem Leben zufrieden.	Ciccarello, L. & Reinhard, M.-A. (2011). Lebensglückskala (LGS). ZPID.
20	Ich halte gerne an alten Traditionen fest.	Hermann, D. (2014). Individuelle reflexive Werte. ZIS.

21	Kinder sollten immer zu ihren Eltern stehen.	Gümüs, A., Gömleksiz, M., Glöckner-Rist, A., & Balke, D. (2014). Einstellungen zu Eigengruppenautoritäten. ZIS.
22	Beim Kauf eines Artikels fühle mich wohler, wenn ich diesen vorher durch Anfassen eingehend geprüft habe.	Nuszbaum, M., Voss, A., Klauer, K.C. & Betsch, T. (2010). Need for Touch Scale (NFT) - deutsche Fassung. ZPID.
23	Es stört mich momentan, dass ich nicht genug Zeit für die Dinge habe, die ich machen will.	Jaekel, J. & Leyendecker, B. (2008). Everyday Stressors Index (ESI) - deutsche Fassung. ZPID.
24	Es ist mir wichtig einen Beruf zu haben, der für die Gesellschaft nützlich ist.	Zentralarchiv für empirische Sozialforschung (ZA) & Zentrum für Umfragen, Methoden und Analysen (ZUMA) e.V. (2014). Wichtigkeit verschiedener Berufsmerkmale. ZIS.
25	Wenn ich Zeitungsberichte über Umweltprobleme lese oder entsprechende Fernsehsendungen sehe, bin ich oft empört und wütend.	Wingerter, C. (2014). Allgemeines Umweltbewusstsein. ZIS.
26	Eine verheiratete Frau, die lieber im Beruf weiterkommen möchte und keine Kinder haben will, sollte deswegen kein schlechtes Gewissen haben.	Krampen, G. (2014). Geschlechtsrollenorientierung. ZIS.
27	Ich mag Gedichte.	Lederer, G. (2014). Kernautoritarismus. ZIS.
28	Ich bin froh, dass ich für Reisen, Wanderungen oder fürs Picknick Erfrischungsgetränke in leichten Dosen oder Plastikflaschen kaufen kann.	Scherhorn, G., Haas, H., Hellenthal, F., & Seibold, S. (2014). Naturverträglichkeit. ZIS.
29	Meine Intuition ist ziemlich gut, wenn es um das Verständnis der Gefühle und Motive anderer geht.	Schyns, B. & Paul, T. (2014). Deutsche Self-Monitoring Skala. ZIS.
30	Ich interessiere mich für Politik.	Zentralarchiv für empirische Sozialforschung (ZA) & Zentrum für Umfragen, Methoden und Analysen (ZUMA) e.V. (2014). Politikinteresse. ZIS.

D.2 Original Wording of Target Variables

Original German wording of the behavioral self-report questions that were used as target variables in the predictive modeling analyses.

Table D.2: Original German Wording of Target Variables

	Impulsivity
Argument	Haben Sie in der letzten Woche einer Person im Streit etwas an den Kopf geworfen, das Sie am liebsten zurück genommen hätten?
Clothes	Haben Sie in der letzten Woche online oder im Geschäft spontan Kleidungsstücke gekauft, ohne dass Sie den Kauf geplant hatten?
Electronics	Haben Sie in den letzten zwei Wochen spontan elektronische Geräte oder Unterhaltungselektronik gekauft, ohne dass Sie den Kauf geplant hätten?
Lying	Haben Sie in den letzten beiden Wochen gelogen, um einen Termin oder eine Verabredung nicht wahrnehmen zu müssen?
Series	Erinnern Sie sich an das letzte mal zurück, als Sie ihre Lieblingsserie gesehen haben. Haben Sie mehr Folgen am Stück angeschaut, als Sie eigentlich vorhatten?
Snacking	Denken Sie an gestern zurück. Welche der folgenden Snacks/Süßigkeiten, die aufgelistet sind, haben Sie gestern konsumiert und wie viel davon?
	Einen Schokoriegel/Müsliriegel
	Eine Hand voll Chips/Nüsse/Salzstangen
	Eine Portion Schokolade (drei Stück)/eine Praline
	Einen großen Cookie bzw. zwei kleine Kekse
	Ein Stück Kuchen
	Eine Hand voll Gummibärchen/Fruchtgummi/Bonbons
	Eine Kugel Eis/ein Eis am Stiel
	Ein Becher Joghurt/Pudding
Height	Wie groß sind Sie?
Weight	Wie viel wiegen Sie momentan?
	Order
Bed	Haben Sie heute morgen ihr Bett gemacht?
Document	Erinnern Sie sich an das letzte mal, als Sie ein wichtiges Dokument erhalten haben (z.B. Versicherungsunterlagen, Zeugnis, Vertrag). Haben Sie dieses Dokument sofort in einem Ordner abgeheftet?
Dishes	Befindet sich in diesem Moment von Ihnen genutztes, ungespültes Geschirr außerhalb des Küchenbereichs?
Desk	Ist ihr Schreibtisch bzw. Arbeitsplatz in diesem Moment aufgeräumt?
Vacuum	Denken Sie an die letzten zwei Wochen. Wie häufig haben Sie in dieser Zeit in ihrem Schlafzimmer Staub gesaugt?

Note. Target variables are listed with respect to their primary scale (Impulsivity or Order). Height and Weight were combined into body mass index (BMI) for all analyses.

D.3 Screenshots of ERS Manipulations

Screenshots depicting the original layout and German wording of the extreme, mid, and neutral-responding instructions.

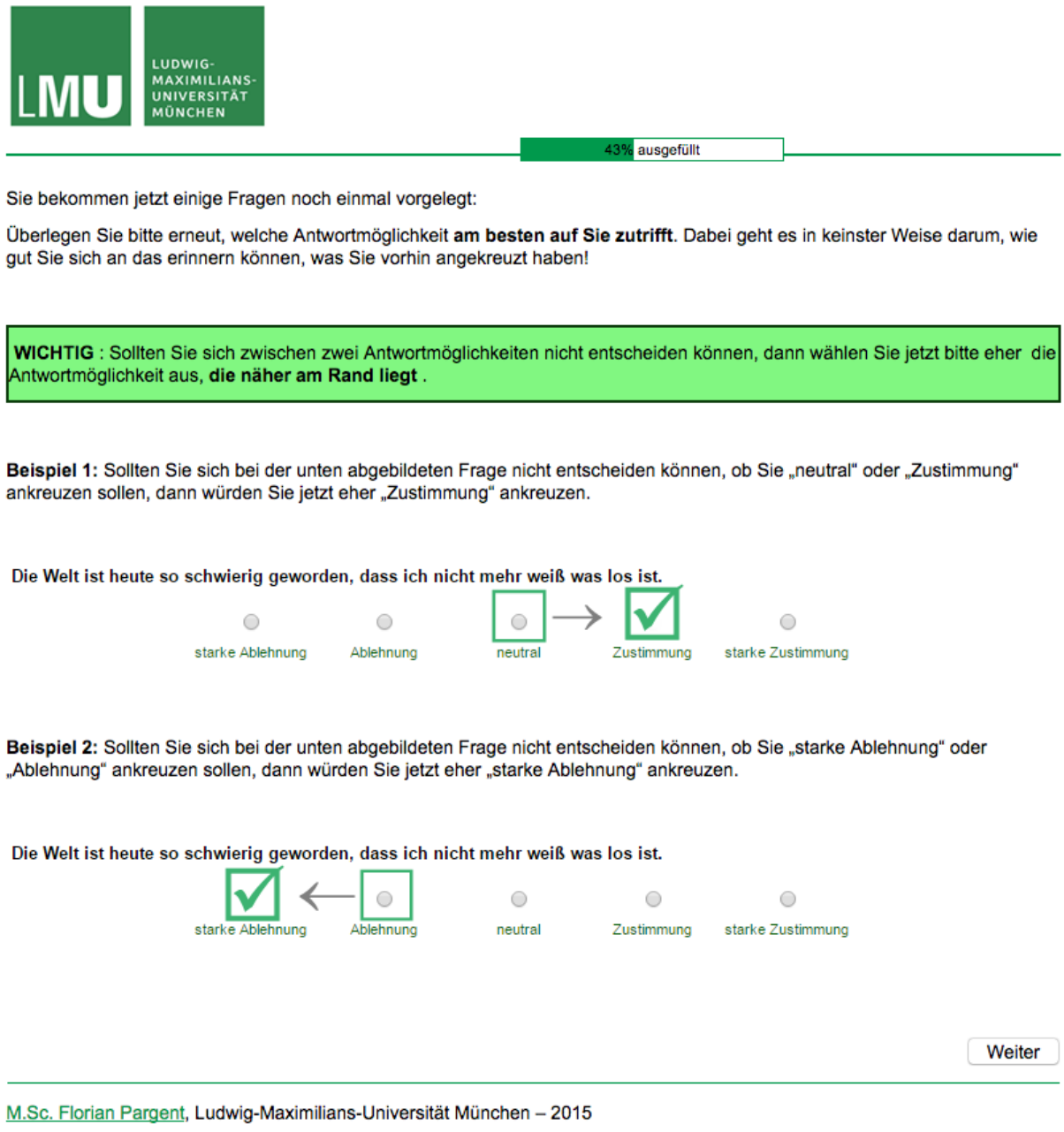

Figure D.1: Screenshot of the extreme-responding instruction. The green box was repeated on the following questionnaire pages.

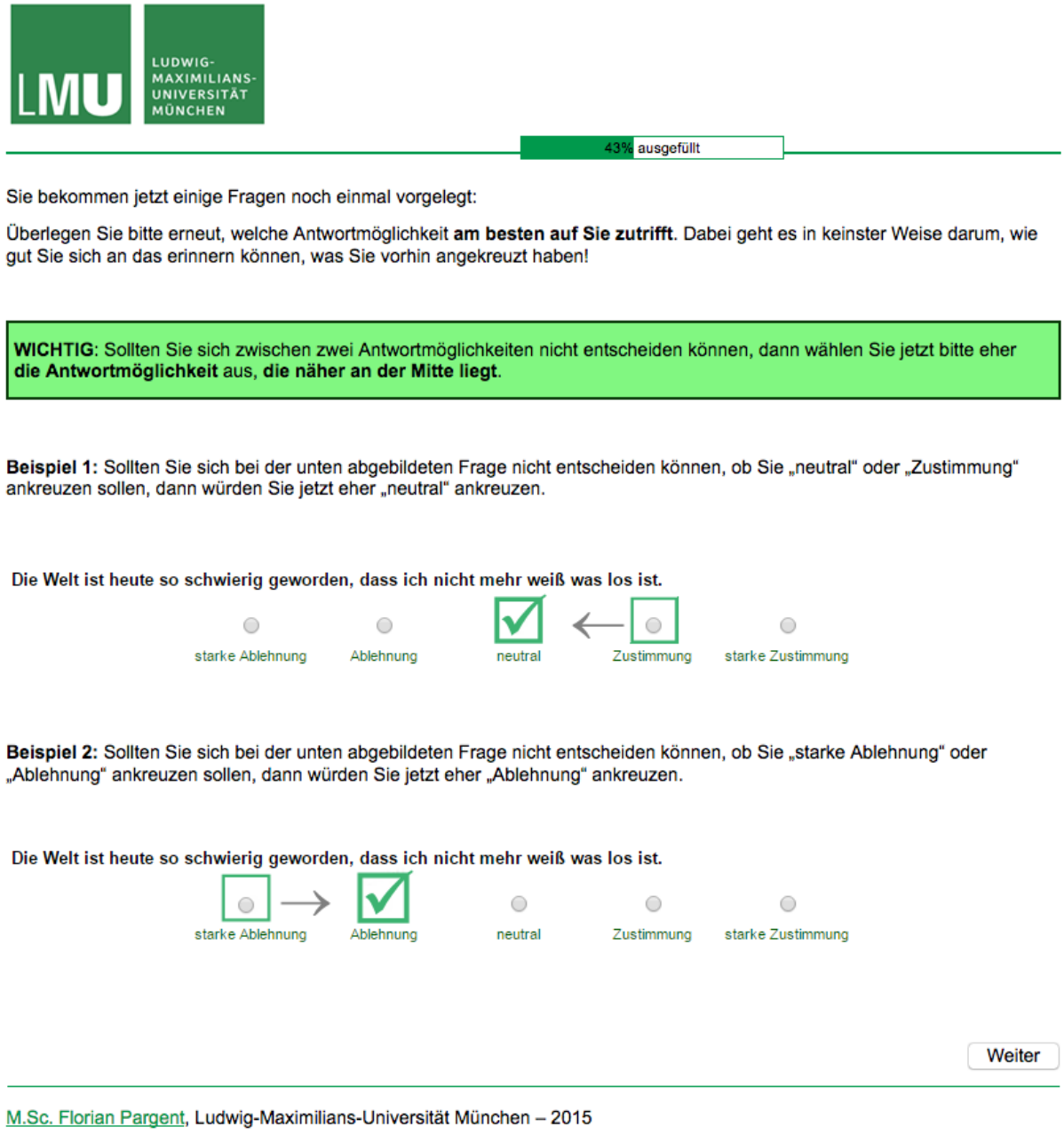

Figure D.2: Screenshot of the mid-responding instruction. The green box was repeated on the following questionnaire pages.

Figure D.3: Screenshot of the neutral instruction. The green box was repeated on the following questionnaire pages.

D.4 PC Trees of Impulsivity and Order from Part A, with the ERS Index as Single Covariate

PC trees of the Impulsivity and Order scales from part A of the questionnaire, with the ERS index from heterogeneous items as single covariate.

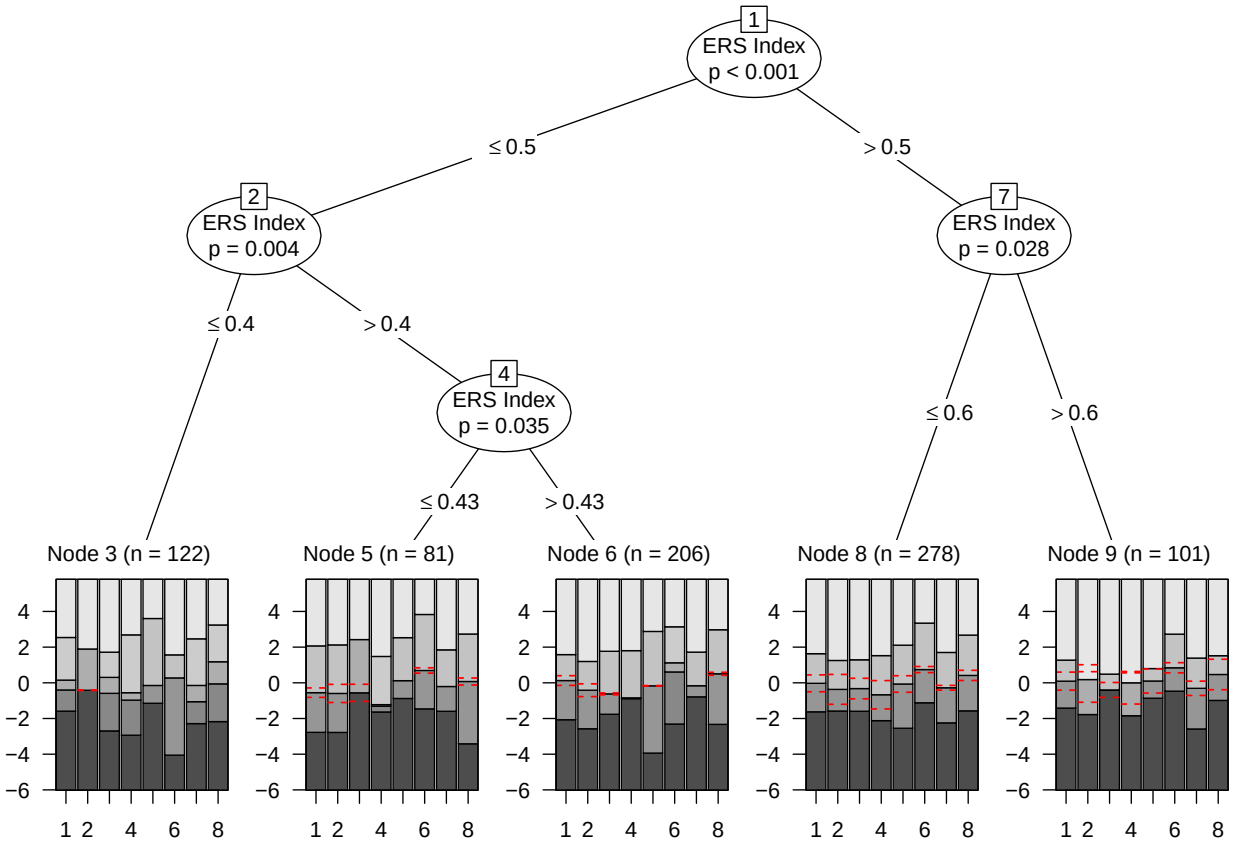

Figure D.4: PC tree of the Order (ORD) scale under standard instructions, with the ERS index from heterogeneous items as single covariate. The analyses is computed for the whole sample with item responses from part A of the questionnaire. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

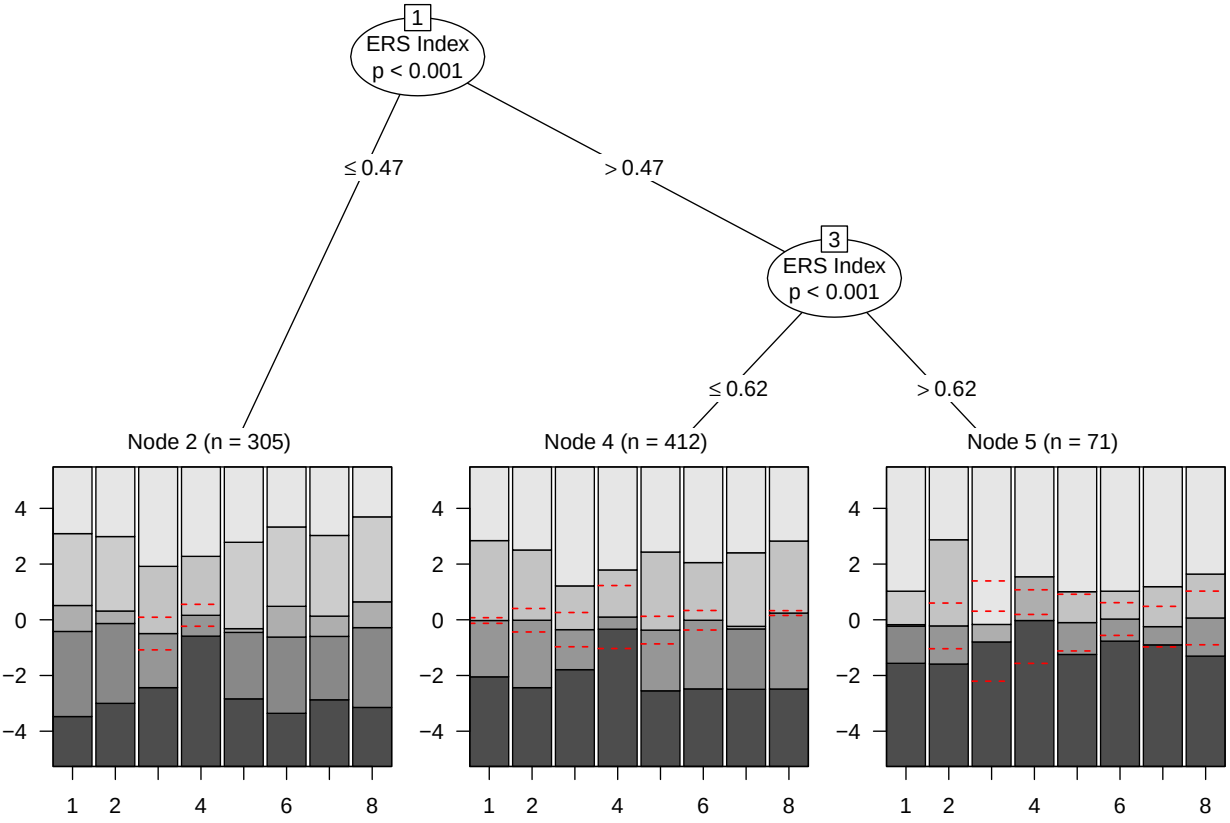

Figure D.5: PC tree of the Impulsivity (IMP) scale under standard instructions, with the ERS index from heterogeneous items as single covariate. The analyses is computed for the whole sample with item responses from part A of the questionnaire. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

D.5 PC Trees of Impulsivity and Order from Part A, with SRERS as Single Covariate

PC trees of the Impulsivity and Order scales from part A of the questionnaire, with SRERS as single covariate.

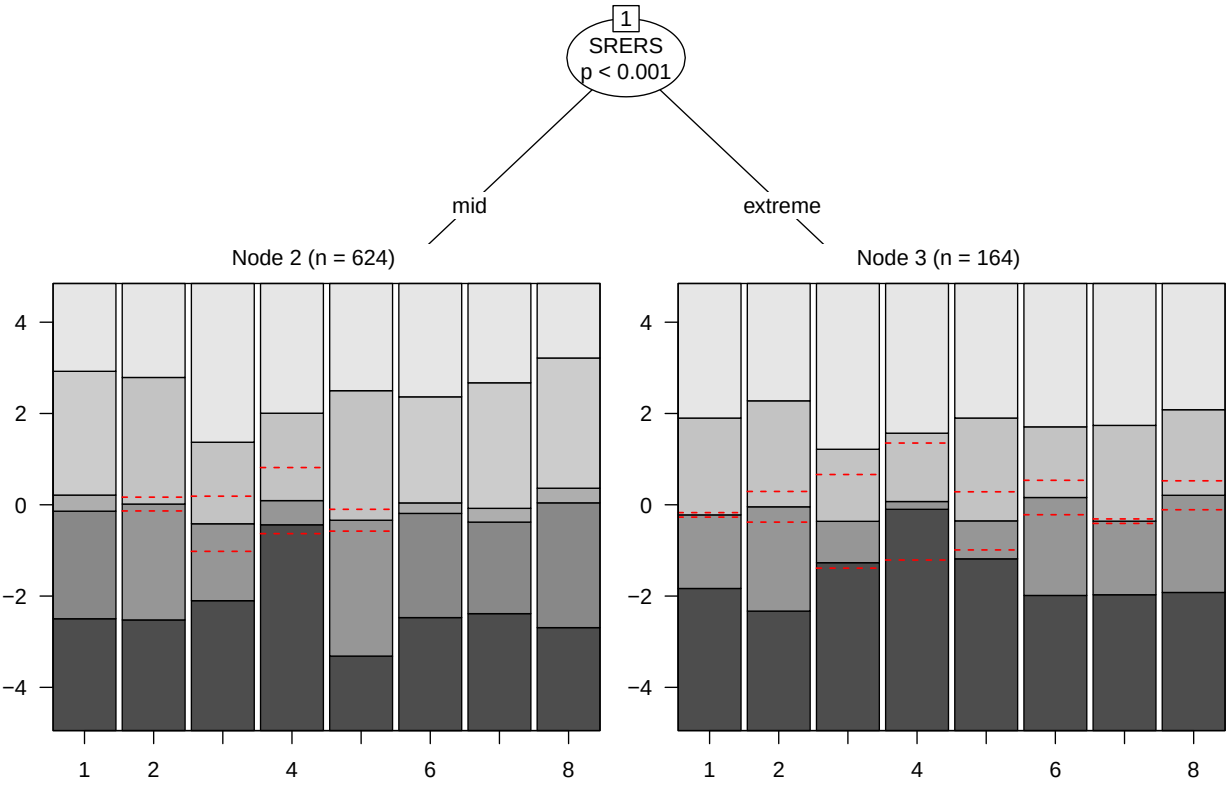

Figure D.6: PC tree of the Impulsivity (IMP) scale under standard instructions with SRERS as single covariate. The analyses is computed for the whole sample with item responses from part A of the questionnaire. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. SRERS = self-reported extreme response style.

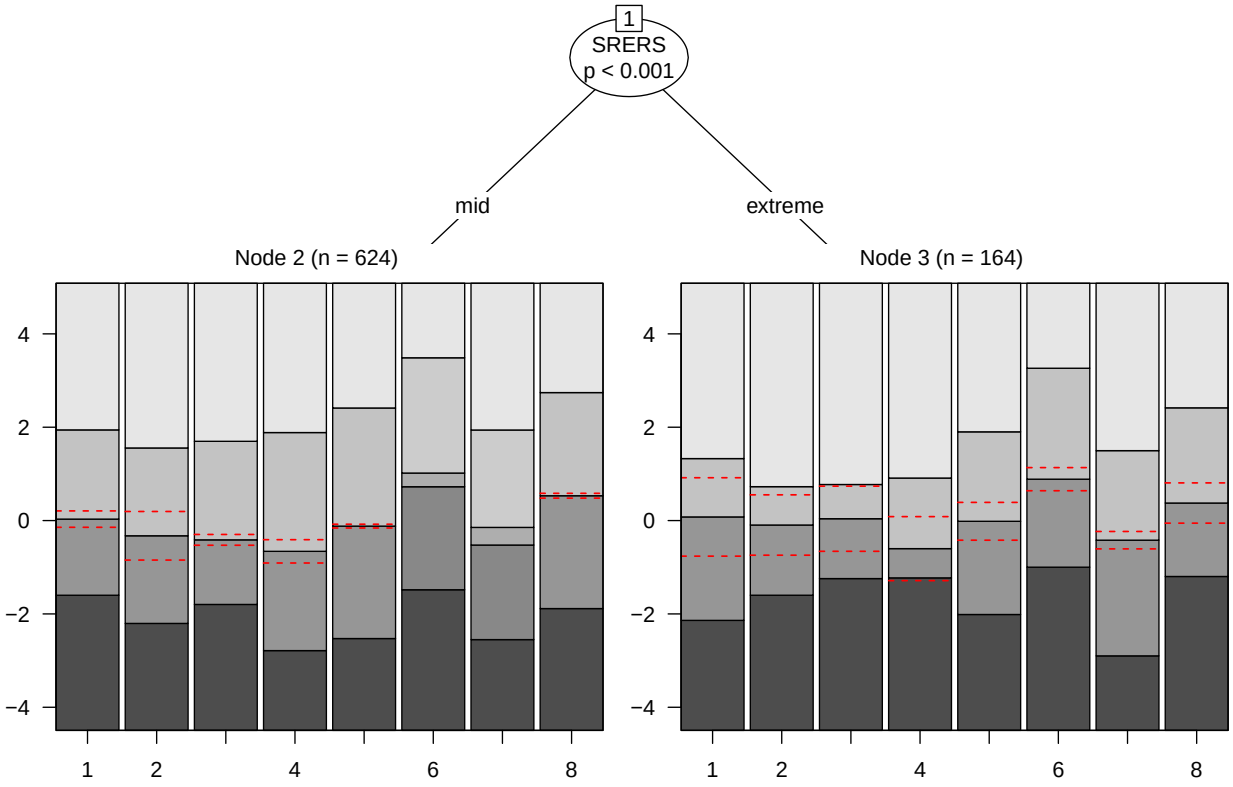

Figure D.7: PC tree of the Order (ORD) scale under standard instructions with SRERS as single covariate. The analyses is computed for the whole sample with item responses from part A of the questionnaire. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. SRERS = self-reported extreme response style.

D.6 PC Trees of Impulsivity and Order from Part C, with the ERS Instruction as Single Covariate

PC trees of the Impulsivity and Order scales from part C of the questionnaire, with the type of ERS instruction as single covariate.

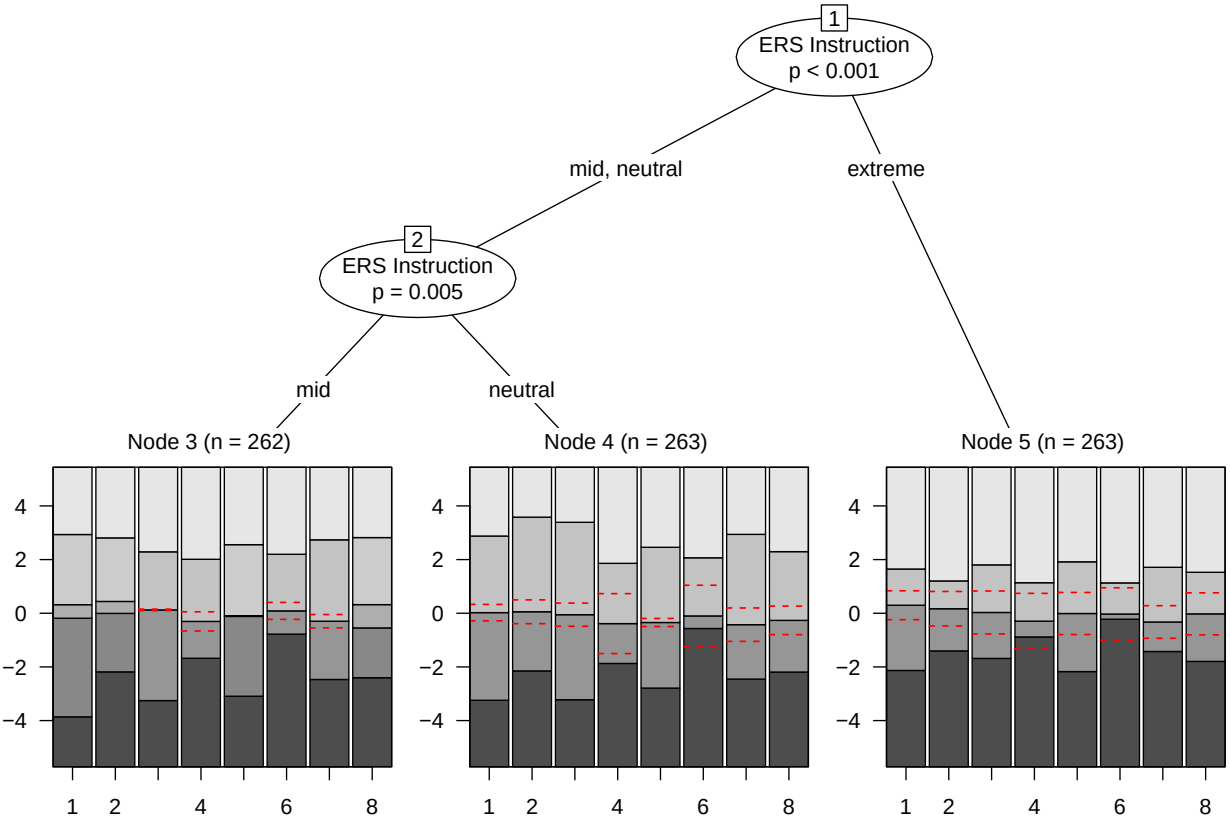

Figure D.8: PC tree of the Impulsivity (IMP) scale from part C of the questionnaire, with the type of ERS instruction as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

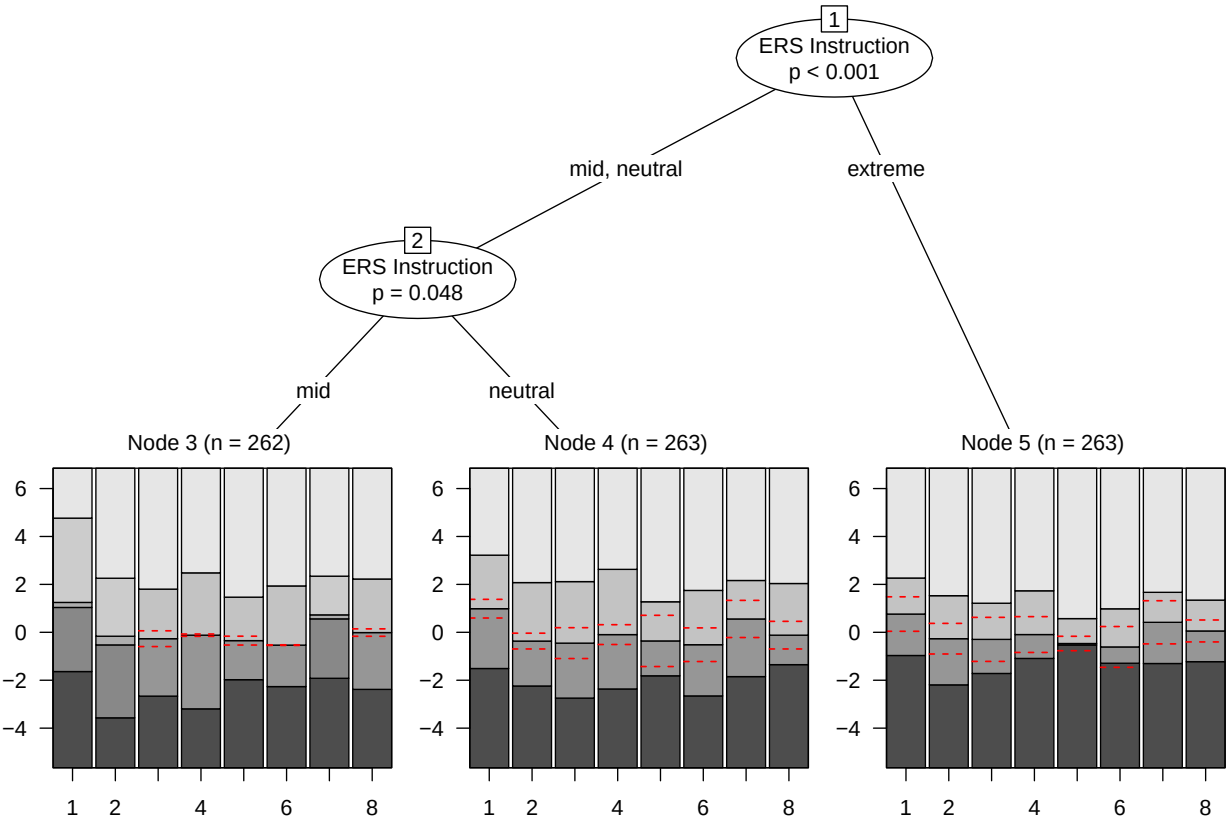

Figure D.9: PC tree of the Order (ORD) scale from part C of the questionnaire, with the type of ERS instruction as single covariate. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

D.7 Item Responses in the Control Group for Parts A, B, and C

Item responses of the control group in the Impulsivity and Order scales from parts A, B, and C of the questionnaire.

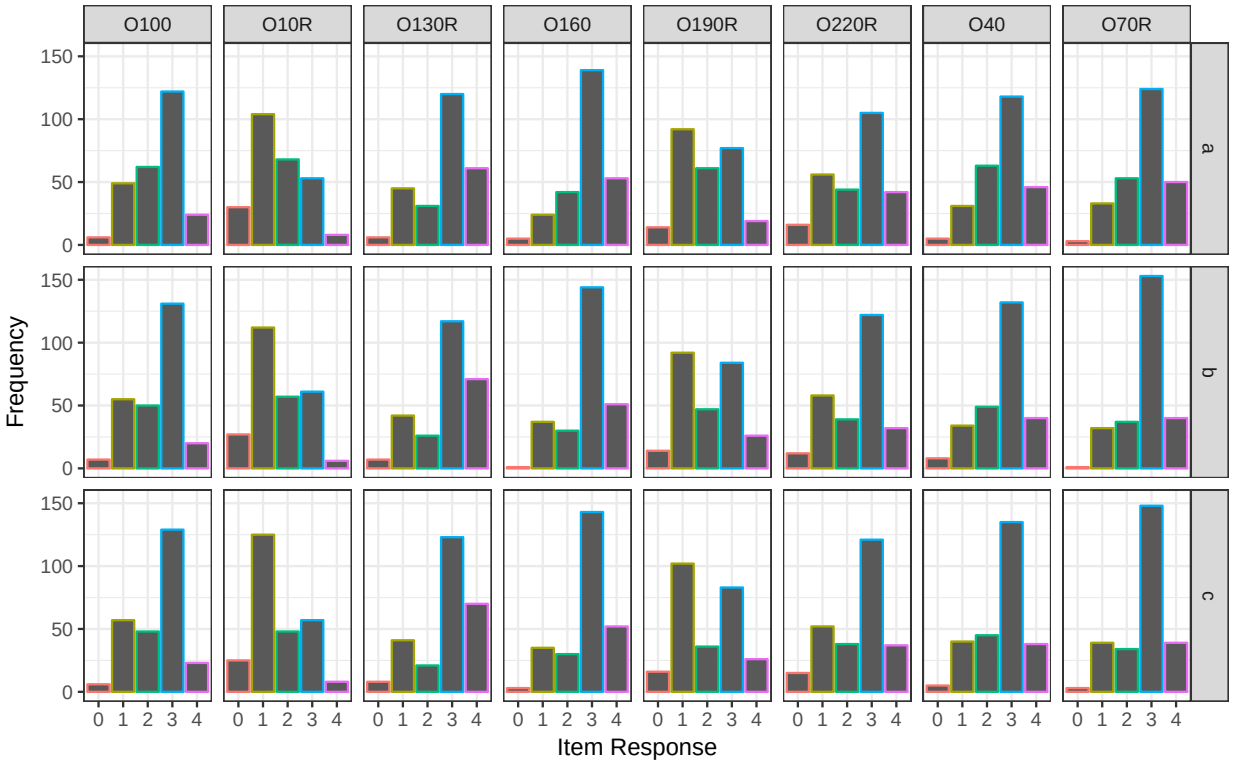

Figure D.10: Histograms for item responses of the Order (ORD) scale in the control group, for all three parts of the questionnaire. Items are labeled with O for Order, the position of the item in the German NEOI-PI-R and the letter R if the item has been recoded.

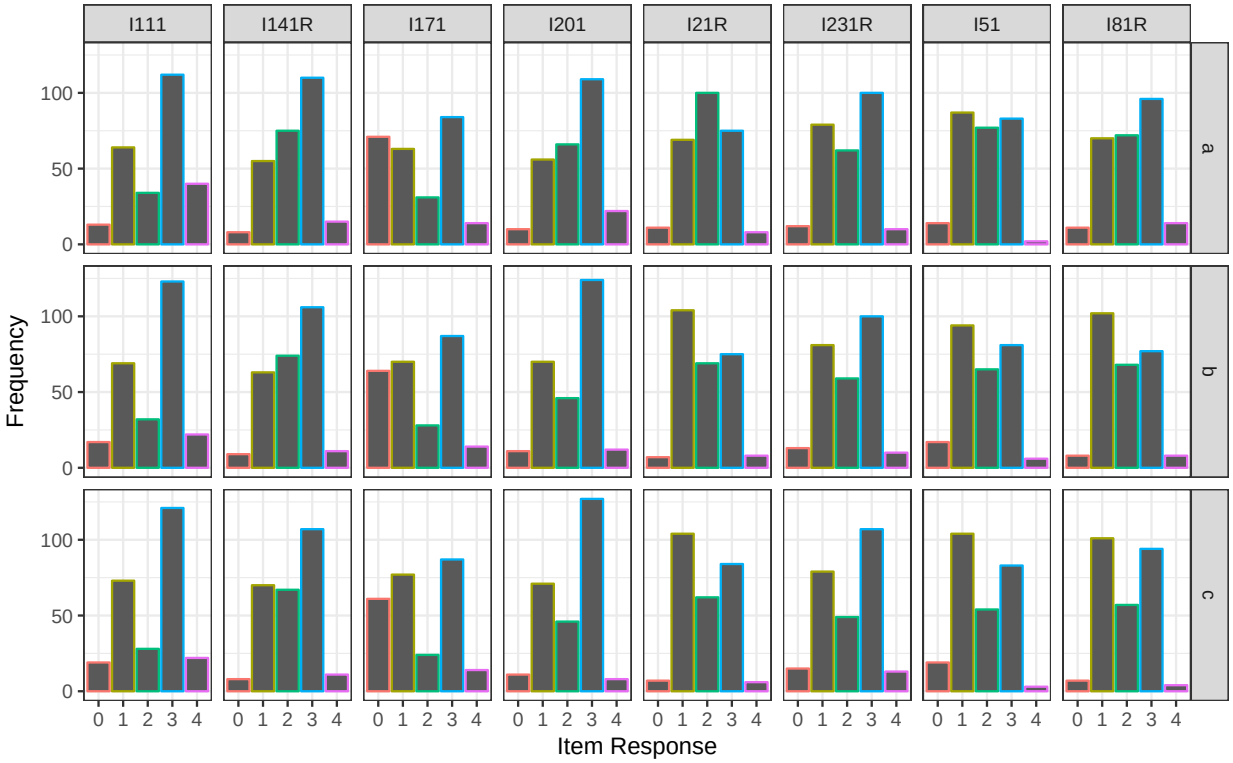

Figure D.11: Histograms for item responses of the Impulsivity (IMP) scale in the control group, for all three parts of the questionnaire. Items are labeled with I for Impulsivity, the position of the item in the German NEOI-PI-R and the letter R if the item has been recoded.

D.8 Target Correlations in the Control Group for Parts A, B, and C

Correlations in the control group between the sum scores of the Impulsivity and Order scales with the target variables, for parts A, B, and C of the questionnaire.

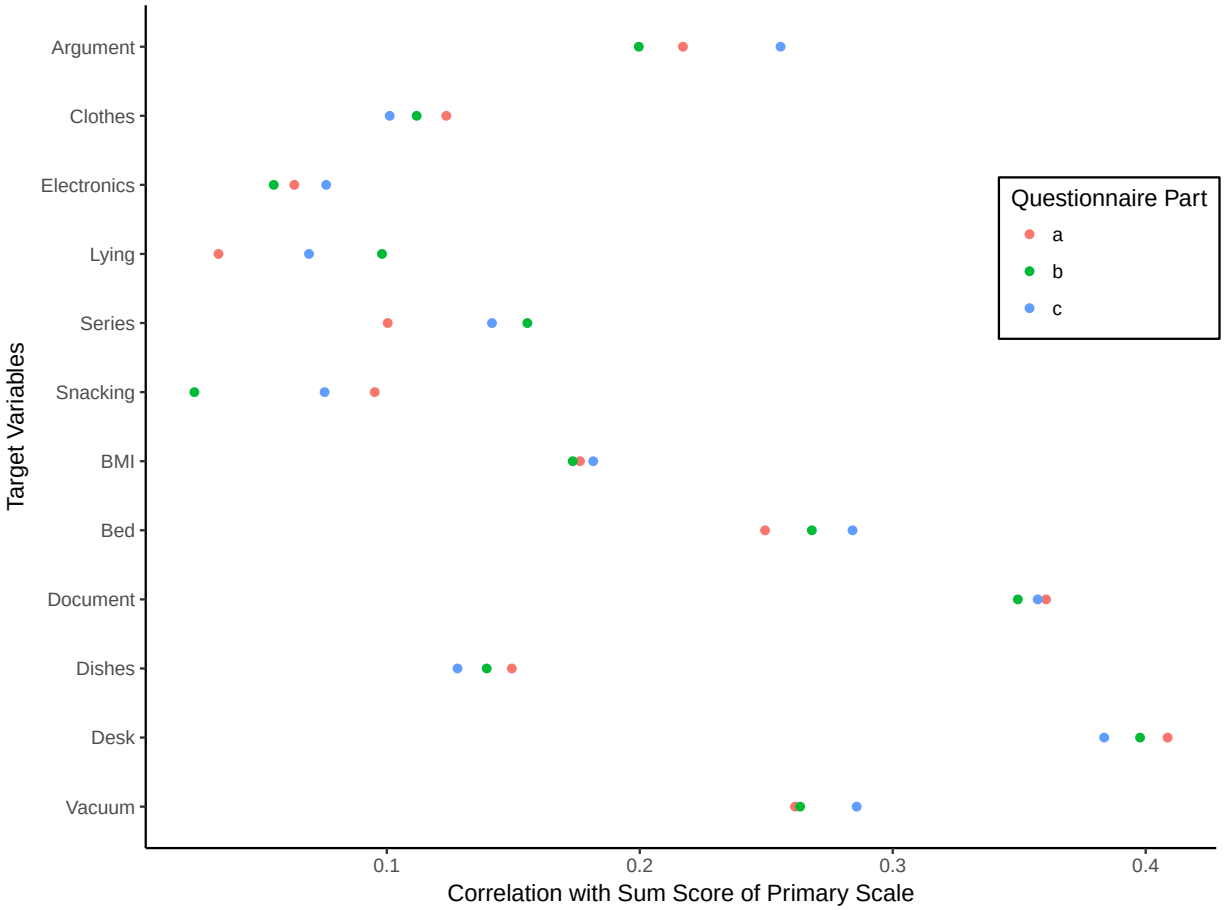

Figure D.12: Correlation of the target variables in the control group, with the sum score of their primary scales. Impulsivity (IMP) is considered the primary scale for variables Argument to BMI, while Bed to Vacuum are target variables for Order (ORD). Sum scores were computed for all three parts of the questionnaire.

D.9 PC Trees of Impulsivity and Order in the Aggravation Setting, with the ERS Index as Single Covariate

PC trees of the Impulsivity and Order scales in the aggravation setting, with the ERS index from heterogeneous items as single covariate. In the aggravation setting, the instruction which should increase the impact of ERS was chosen for each participant in the experimental group.

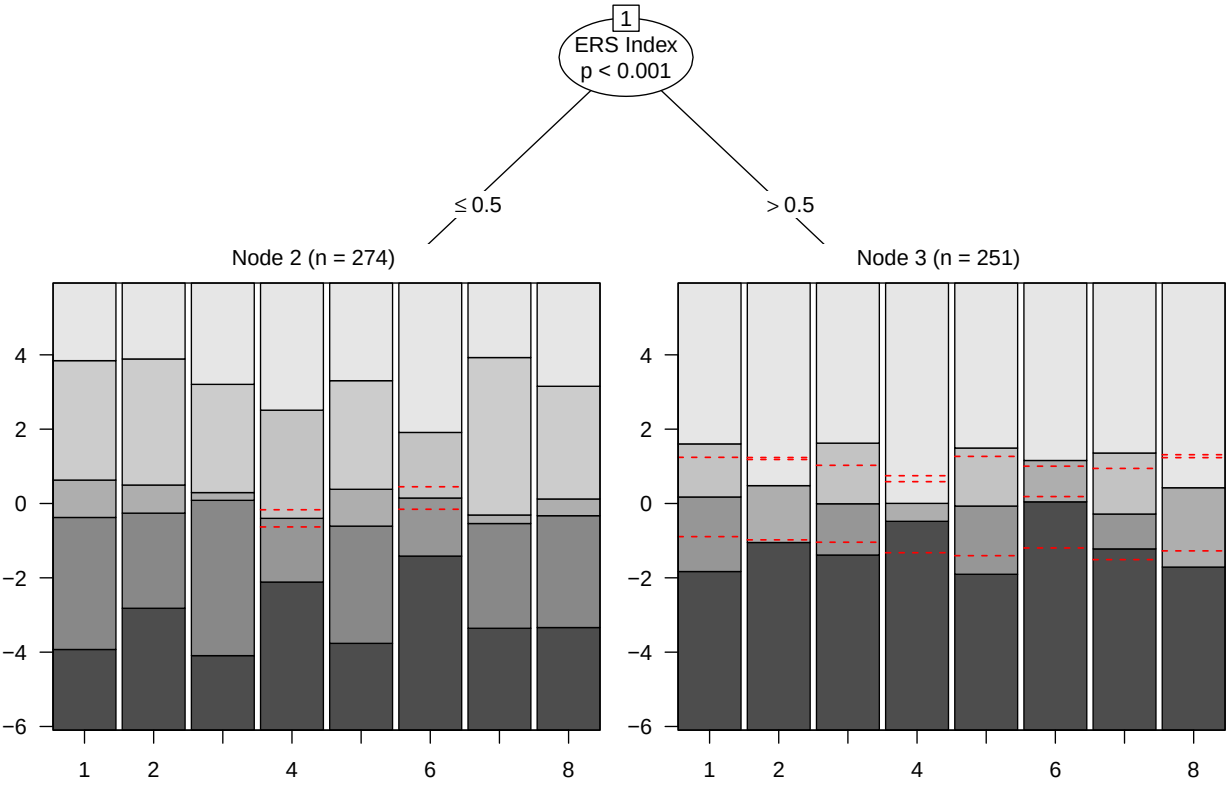

Figure D.13: PC tree of the Impulsivity (IMP) scale in the aggravation setting, with the ERS index from heterogeneous items as single covariate. Matching the ERS instruction to individual response style was based on a median split of the ERS index from heterogeneous items. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

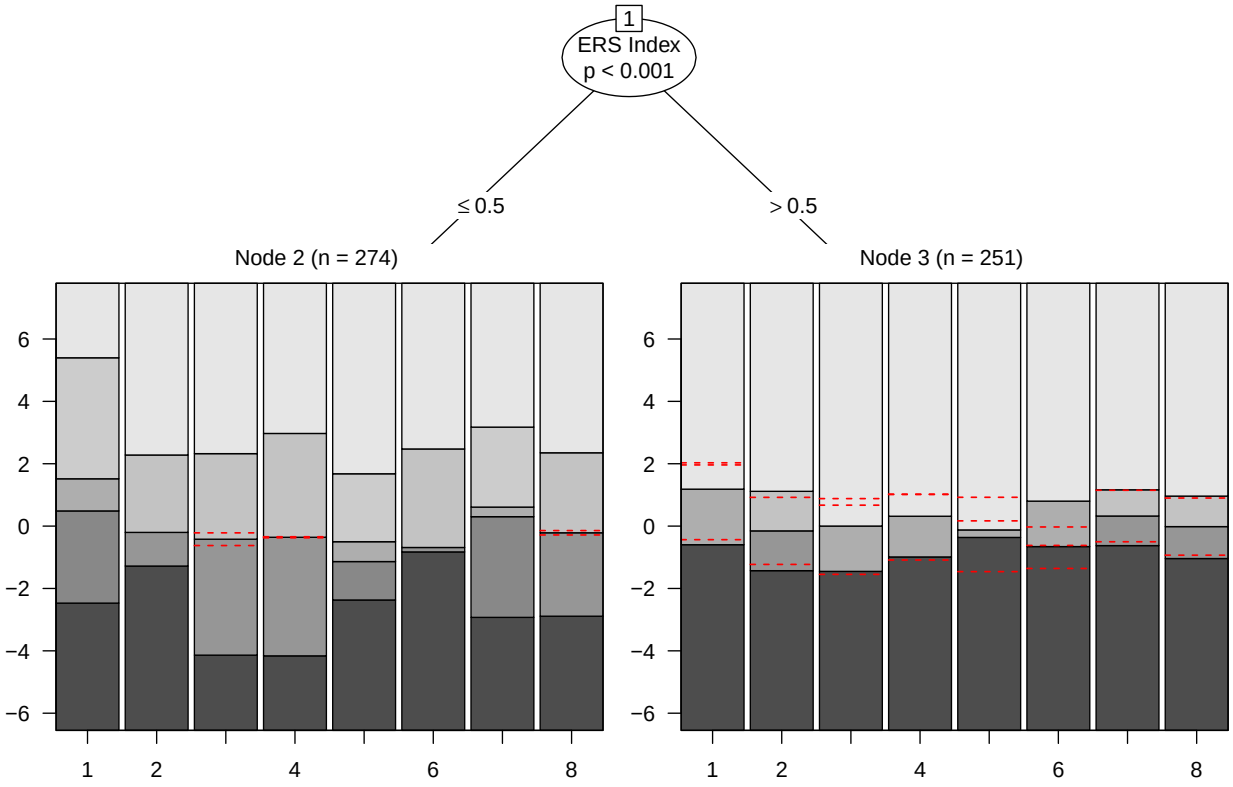

Figure D.14: PC tree of the Order (ORD) scale in the aggravation setting, with the ERS index from heterogeneous items as single covariate. Matching the ERS instruction to individual response style was based on a median split of the ERS index from heterogeneous items. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

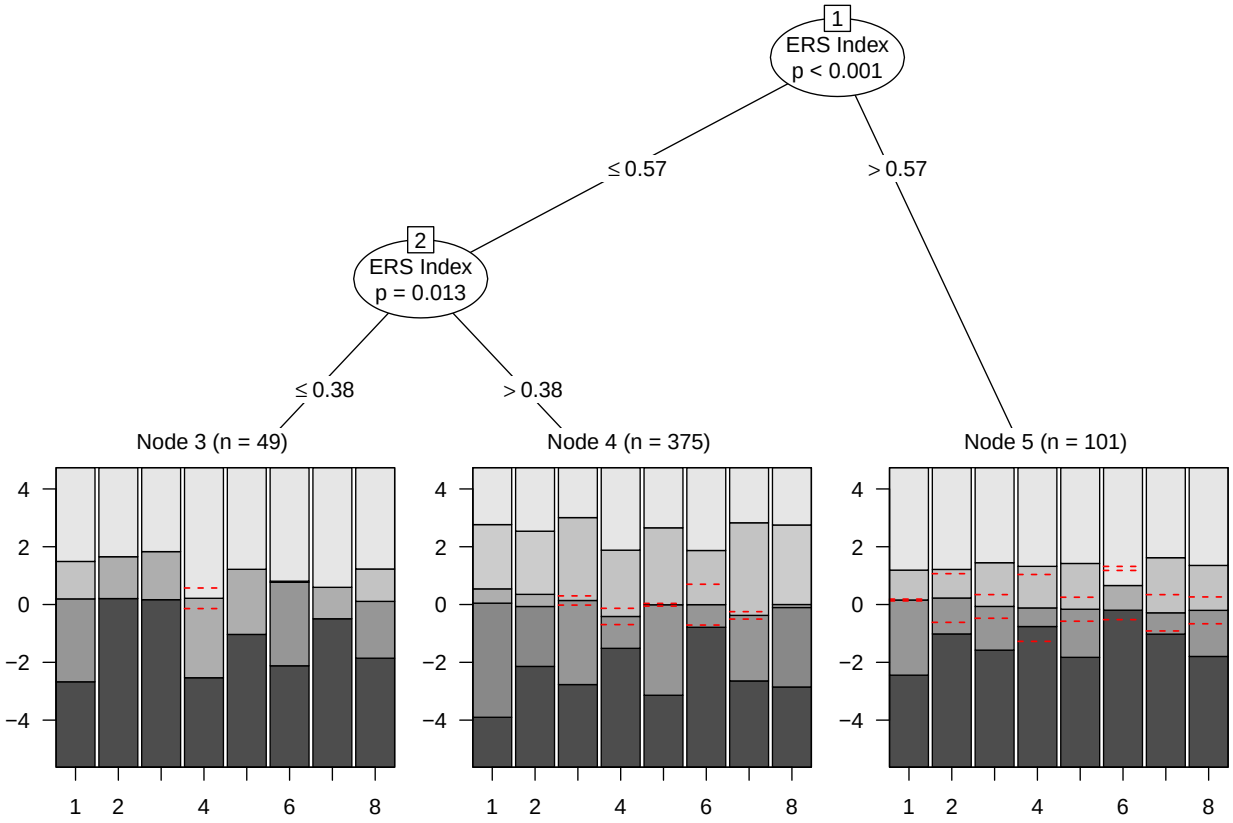

Figure D.15: PC tree of the Impulsivity (IMP) scale in the aggravation setting, with the ERS index from heterogeneous items as single covariate. Matching the ERS instruction to individual response style was based on SRERS. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. SRERS = self-reported extreme response style.

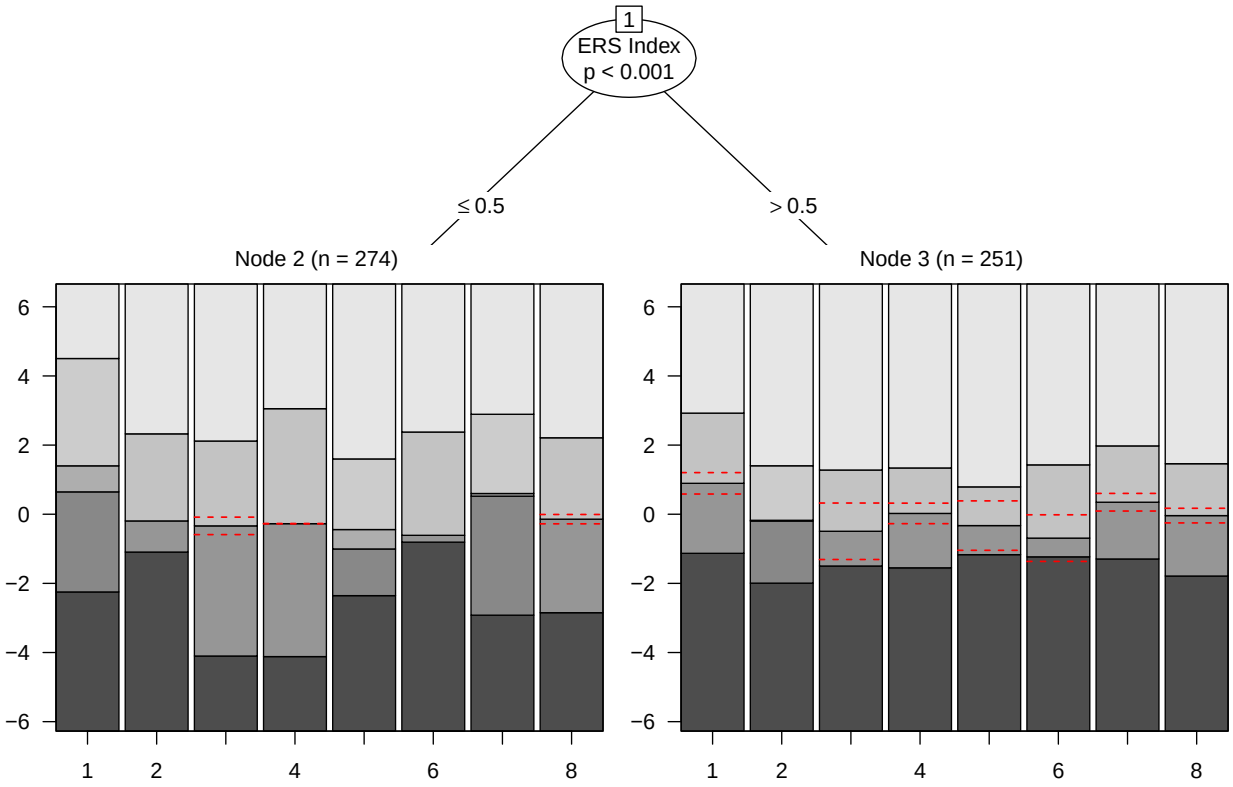

Figure D.16: PC tree of the Order (ORD) scale in the aggravation setting, with the ERS index from heterogeneous items as single covariate. Matching the ERS instruction to individual response style was based on SRERS. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped. SRERS = self-reported extreme response style.

D.10 PC Trees of Impulsivity and Order in the Control Setting, with the ERS Index as Single Covariate

PC trees of the Impulsivity and Order scales in the control setting, with the ERS index from heterogeneous items as single covariate. In the control setting, item responses for each participant in the control group were randomly chosen form either part B, or part C of the questionnaire.

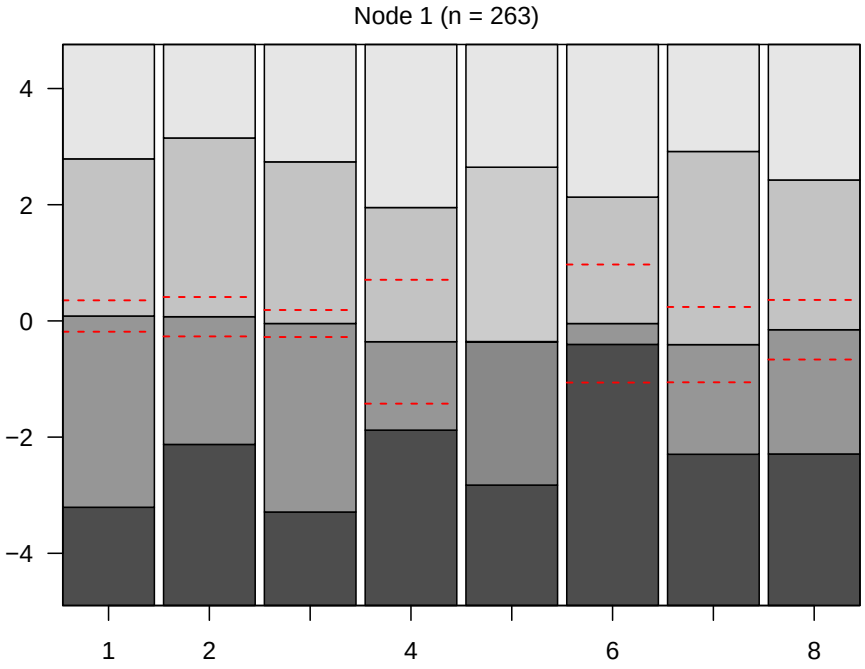

Figure D.17: PC tree of the Impulsivity (IMP) scale in the control setting, with the ERS index from heterogeneous items as single covariate. All item responses for each participant were randomly chosen either from part B or part C of the questionnaire. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

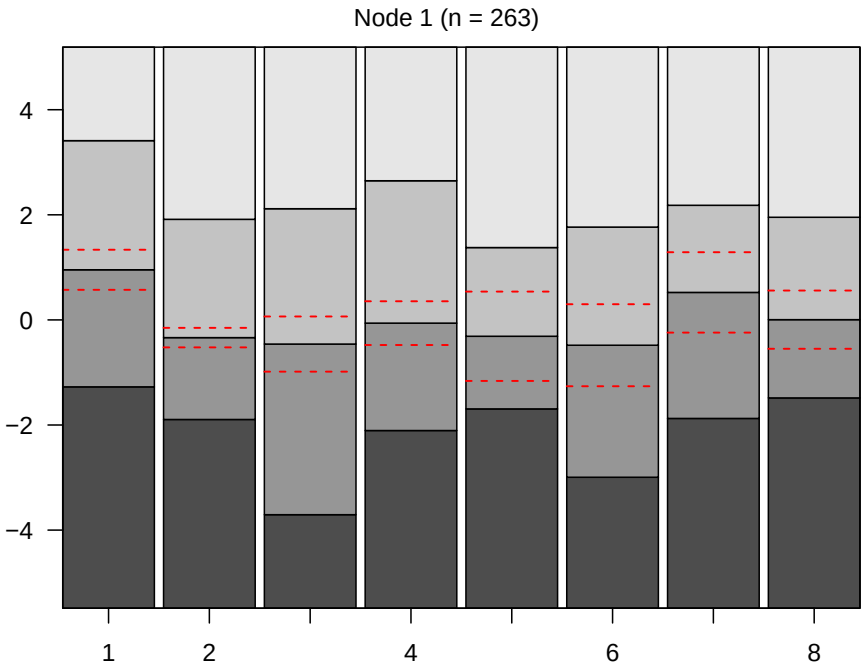

Figure D.18: PC tree of the Order (ORD) scale in the control setting, with the ERS index from heterogeneous items as single covariate. All item responses for each participant were randomly chosen either from part B or part C of the questionnaire. Regions of highest response category probability are shown in greyscale. Red lines indicate unordered thresholds in the partial credit model. Missing categories were dropped.

D.11 Additional Performance Measures for Predictive Models from Part A

Aggregated performance measures for the predictive models in which each target variable was predicted by item responses of the Impulsivity and Order scales from part A of the questionnaire, sex, and age. The ERS index from heterogeneous items and SRERS were either included of excluded. Cohen's κ, MMCE, SENS and SPEC are reported for binary target variables. R^2, MSE, and RMSE are reported for metric target variables.

Table D.3: Additional Performance Estimates from Part A for Binary Targets

Target	ERS Measures	κ	MMCE	SENS	SPEC
Argument	excluded	0.038	0.173	0.029	0.995
Argument	included	0.003	0.176	0.005	0.997
Clothes	excluded	0.019	0.249	0.028	0.986
Clothes	included	0.019	0.245	0.019	0.994
Electronics	excluded	0.000	0.068	0.000	1.000
Electronics	included	0.000	0.068	0.000	1.000
Lying	excluded	0.046	0.349	0.122	0.915
Lying	included	0.054	0.345	0.123	0.922
Series	excluded	0.128	0.410	0.779	0.343
Series	included	0.103	0.418	0.794	0.304
Bed	excluded	0.199	0.369	0.413	0.777
Bed	included	0.181	0.374	0.387	0.785
Document	excluded	0.189	0.341	0.338	0.837
Document	included	0.180	0.345	0.330	0.835
Dishes	excluded	0.099	0.366	0.863	0.225
Dishes	included	0.095	0.363	0.880	0.203
Desk	excluded	0.325	0.336	0.690	0.635
Desk	included	0.340	0.328	0.710	0.629

Note. Estimated performance for predicting binary target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale from part A of the questionnaire, sex, age, the ERS index from heterogeneous items and SRERS. Predictive models were estimated once with and once without the ERS measures. SRERS = self-reported extreme response style, κ = Cohen's κ, MMCE = mean misclassification error, SENS = sensitivity, SPEC = specificity.

Table D.4: Additional Performance Estimates from Part A for Metric Targets

Target	ERS Measures	R^2	MSE	RMSE
Snacking	excluded	0.009	11.405	3.377
Snacking	included	-0.006	11.527	3.395
BMI	excluded	0.121	16.172	4.021
BMI	included	0.128	16.013	4.002
Vacuum	excluded	0.043	3.185	1.785
Vacuum	included	0.043	3.190	1.786

Note. Estimated performance for predicting metric target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale from part A of the questionnaire, sex, age, the ERS index from heterogeneous items and SRERS. Predictive models were estimated once with and once without the ERS measures. SRERS = self-reported extreme response style, R^2 = coefficient of determination, MSE = mean squared error, RMSE = root mean squared error.

D.12 Additional Predictive Performance Measures when Matching ERS Instructions Based on a Median Split of the ERS Index

Aggregated performance measures for the predictive models in which each target variable was predicted by item responses of the Impulsivity and Order scales from part B or C of the questionnaire. We compare predictive models in which item responses were given under either the aggravation, compensation, or control setting. A median split of the ERS index from heterogeneous items was used in the matching process. Cohen's κ, MMCE, SENS and SPEC are reported for binary target variables. R^2, MSE, and RMSE are reported for metric target variables.

Table D.5: Performance Estimates for Binary Targets and Matching ERS Instructions Based on a Median Split of the ERS Index

Target	ERS Setting	κ	MMCE	SENS	SPEC
Argument	aggravation	0.089	0.159	0.075	0.986
Argument	compensation	0.057	0.161	0.048	0.990
Argument	control	0.037	0.210	0.050	0.976
Clothes	aggravation	0.012	0.248	0.024	0.985
Clothes	compensation	-0.003	0.250	0.013	0.985
Clothes	control	-0.007	0.273	0.036	0.958
Electronics	aggravation	0.000	0.061	0.000	1.000
Electronics	compensation	0.000	0.061	0.000	1.000
Electronics	control	-0.002	0.085	0.000	0.999
Lying	aggravation	0.127	0.329	0.218	0.890
Lying	compensation	0.126	0.321	0.197	0.911
Lying	control	0.013	0.389	0.156	0.855
Series	aggravation	0.041	0.460	0.697	0.343
Series	compensation	0.144	0.411	0.738	0.402
Series	control	0.168	0.390	0.739	0.425
Bed	aggravation	0.066	0.411	0.253	0.808
Bed	compensation	0.144	0.384	0.336	0.799
Bed	control	0.198	0.374	0.417	0.773
Document	aggravation	0.133	0.366	0.288	0.833
Document	compensation	0.143	0.358	0.283	0.849
Document	control	0.122	0.353	0.286	0.828
Dishes	aggravation	0.066	0.364	0.886	0.173
Dishes	compensation	0.079	0.364	0.877	0.191
Dishes	control	0.012	0.409	0.853	0.158
Desk	aggravation	0.278	0.360	0.674	0.603
Desk	compensation	0.308	0.345	0.688	0.620
Desk	control	0.369	0.303	0.811	0.549

Note. Estimated performance for predicting binary target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale under the compensation, aggravation, and control settings. ERS instructions were matched to individual response tendencies based on a median split of the ERS index from heterogeneous items. κ = Cohen's κ, MMCE = mean misclassification error, SENS = sensitivity, SPEC = specificity.

Table D.6: Performance Estimates for Metric Targets and Matching ERS Instructions Based on a Median Split of the ERS Index

Target	ERS Setting	R^2	MSE	RMSE
Snacking	aggravation	-0.004	11.994	3.463
Snacking	compensation	-0.001	11.806	3.436
Snacking	control	-0.134	11.135	3.337
BMI	aggravation	-0.031	19.370	4.401
BMI	compensation	-0.009	19.164	4.378
BMI	control	-0.107	17.642	4.200
Vacuum	aggravation	0.016	3.082	1.756
Vacuum	compensation	-0.007	3.136	1.771
Vacuum	control	-0.015	3.416	1.848

Note. Estimated performance for predicting metric target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale under the compensation, aggravation, and control settings. ERS instructions were matched to individual response tendencies based on a median split of the ERS index from heterogeneous items. R^2 = coefficient of determination, MSE = mean squared error, RMSE = root mean squared error.

D.13 Predictive Performance when Matching ERS Instructions Based on SRERS

Visual display of performance measures for the predictive models in which each target variable was predicted by item responses of the Impulsivity and Order scales from part B or C of the questionnaire. We compare predictive models in which item responses were given under either the aggravation, compensation, or control setting. SRERS was used in the matching process. Cohen's $\kappa$is depicted for binary target variables, while R^2 is shown for metric target variables.

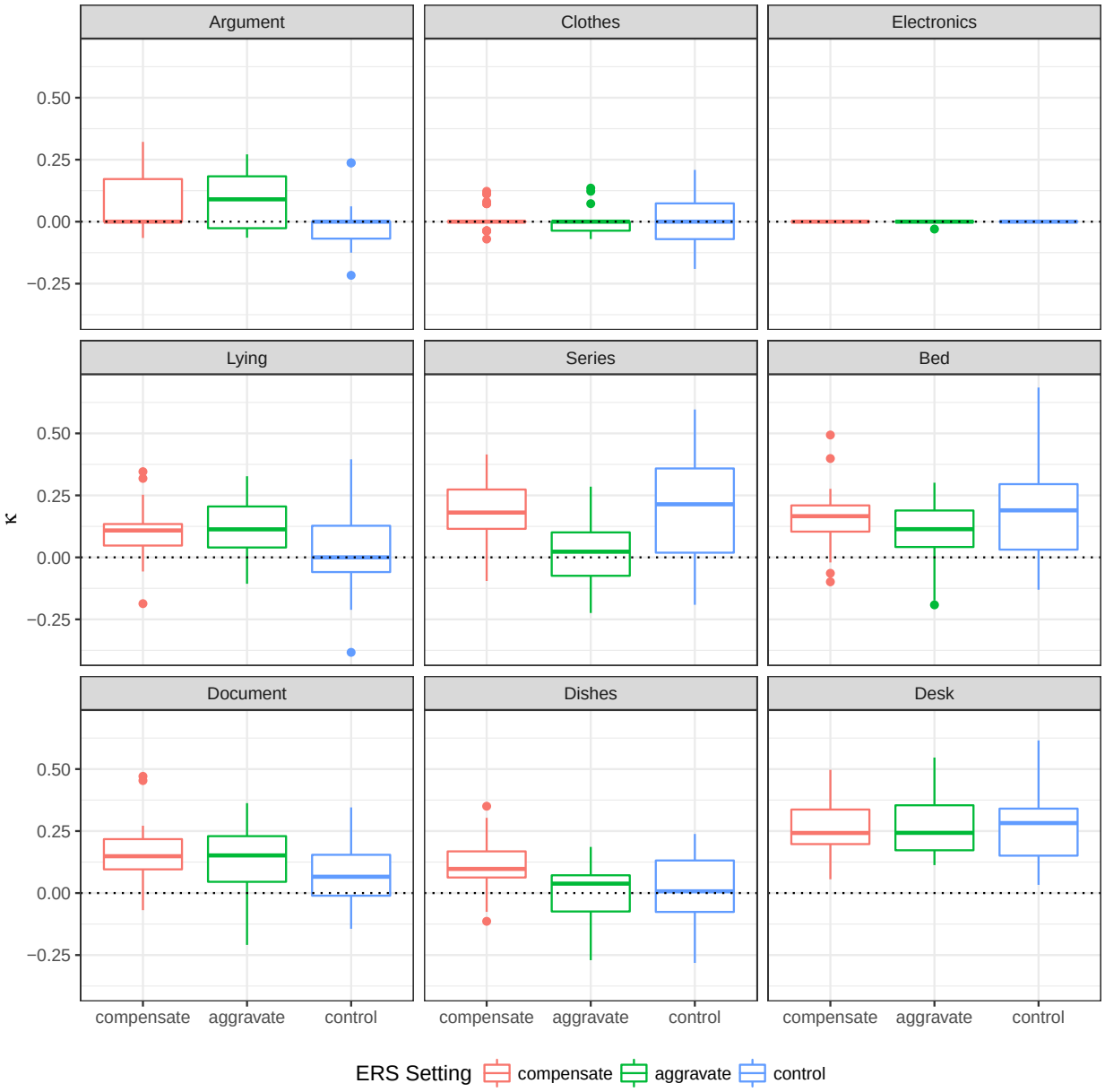

Figure D.19: Estimated performance for predicting binary target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale under the compensation, aggravation, and control settings. ERS instructions were matched to individual response tendencies based on SRERS. Performance was measured by Cohen's κ. SRERS = self-reported extreme response style.

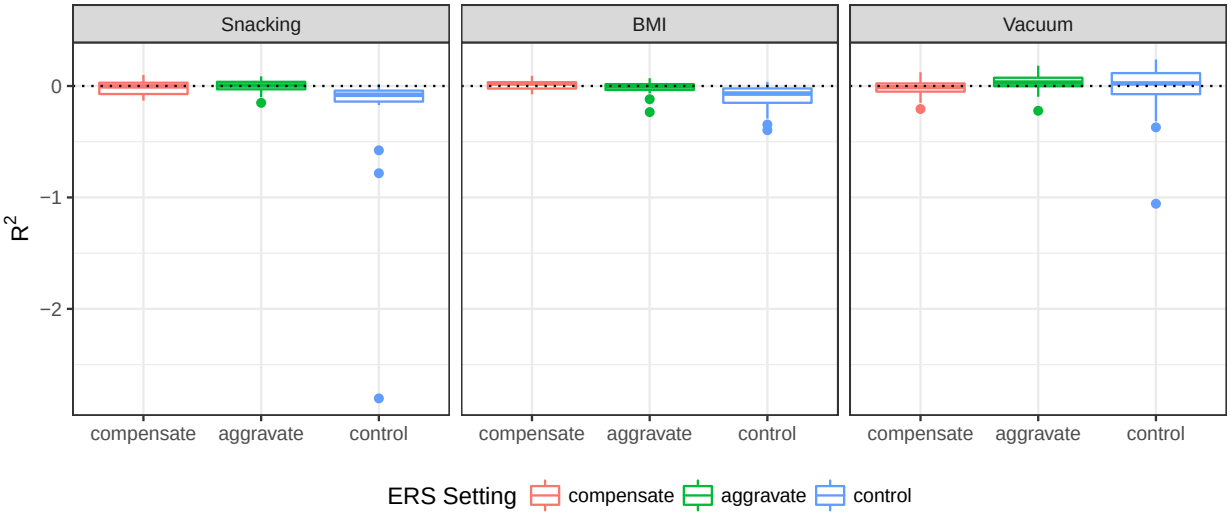

Figure D.20: Estimated performance for predicting metric target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale under the compensation, aggravation, and control settings. ERS instructions were matched to individual response tendencies based on SRERS. Performance was measured by R^2. R^2 = coefficient of determination, SRERS = self-reported extreme response style.

D.14 Additional Predictive Performance Measures when Matching ERS Instructions Based on SRERS

Aggregated performance measures for the predictive models in which each target variable was predicted by item responses of the Impulsivity and Order scales from part B or C of the questionnaire. We compare predictive models in which item responses were given under either the aggravation, compensation, or control setting. SRERS was used in the matching process. Cohen's κ, MMCE, SENS and SPEC are reported for binary target variables. R^2, MSE, and RMSE are reported for metric target variables.

Table D.7: Performance Estimates for Binary Targets and Matching ERS Instructions Based on SRERS

Target	ERS Setting	κ	MMCE	SENS	SPEC
Argument	aggravation	0.085	0.163	0.079	0.981
Argument	compensation	0.084	0.157	0.067	0.990
Argument	control	-0.015	0.224	0.023	0.965
Clothes	aggravation	-0.003	0.249	0.010	0.987
Clothes	compensation	0.014	0.244	0.018	0.992
Clothes	control	0.005	0.271	0.049	0.956
Electronics	aggravation	-0.001	0.062	0.000	0.999
Electronics	compensation	0.000	0.061	0.000	1.000
Electronics	control	0.000	0.083	0.000	1.000
Lying	aggravation	0.111	0.330	0.193	0.901
Lying	compensation	0.106	0.331	0.187	0.902
Lying	control	0.032	0.378	0.149	0.877
Series	aggravation	0.015	0.476	0.659	0.356
Series	compensation	0.186	0.392	0.739	0.443
Series	control	0.204	0.374	0.741	0.459
Bed	aggravation	0.098	0.399	0.286	0.806
Bed	compensation	0.154	0.376	0.330	0.815
Bed	control	0.176	0.389	0.445	0.727
Document	aggravation	0.129	0.368	0.289	0.831
Document	compensation	0.160	0.359	0.323	0.825
Document	control	0.078	0.364	0.231	0.839
Dishes	aggravation	0.001	0.385	0.885	0.115
Dishes	compensation	0.113	0.350	0.889	0.209
Dishes	control	0.012	0.410	0.843	0.170
Desk	aggravation	0.274	0.361	0.704	0.569
Desk	compensation	0.266	0.366	0.669	0.596
Desk	control	0.285	0.344	0.766	0.514

Note. Estimated performance for predicting binary target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale under the compensation, aggravation, and control settings. ERS instructions were matched to individual response tendencies based on SRERS. κ = Cohen's κ, MMCE = mean misclassification error, SENS = sensitivity, SPEC = specificity, SRERS = self-reported extreme response style.

Table D.8: Performance Estimates for Metric Targets and Matching ERS Instructions Based on SRERS

Target	ERS Setting	R^2	MSE	RMSE
Snacking	aggravation	-0.001	12.006	3.465
Snacking	compensation	-0.014	12.013	3.466
Snacking	control	-0.208	11.028	3.321
BMI	aggravation	-0.016	19.219	4.384
BMI	compensation	0.011	18.907	4.348
BMI	control	-0.098	18.081	4.252
Vacuum	aggravation	0.029	3.065	1.751
Vacuum	compensation	-0.018	3.149	1.774
Vacuum	control	-0.028	3.487	1.867

Note. Estimated performance for predicting metric target variables with item responses of the Impulsivity (IMP) and Order (ORD) scale under the compensation, aggravation, and control settings. ERS instructions were matched to individual response tendencies based on SRERS. R^2 = coefficient of determination, MSE = mean squared error, RMSE = root mean squared error, SRERS = self-reported extreme response style.

References

Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social science research*, *42*(3), 957–970.

Alarcon, G., Eschleman, K. J., & Bowling, N. A. (2009). Relationships between personality variables and burnout: A meta-analysis. *Work & stress*, *23*(3), 244–263.

Alexeev, N., Templin, J., & Cohen, A. S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, *48*(3), 313–332.

Ashton, M. C. & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and social psychology review*, *11*(2), 150–166.

Auguie, B. (2016). *Gridextra: miscellaneous functions for "grid" graphics*. R package version 2.2.1.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, *40*(6), 1235–1245.

Bachman, J. G. & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, *48*(2), 491–509.

Bäckström, M. & Björklund, F. (2014). Social Desirability in Personality Inventories. *Journal of Individual Differences*, *35*, 144–157.

Bartlett, M. S. (1951). The Effect of Standardization on a Chi Square Approximation in Factor Analysis. *Biometrika*, *38*(3), 337–344.

Baumgartner, H. & Steenkamp, J.-B. E. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, *38*(2), 143–156.

Ben-Porath, Y. S., Hostetler, K., Butcher, J. N., & Graham, J. R. (1989). New subscales for the MMPI-2 Social Introversion (Si) scale. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *1*(3), 169–174.

Bentler, P. M. (2016). Covariate-free and Covariate-dependent Reliability. *Psychometrika*, *81*(4), 907–920.

Beuckelaer, A. D., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: a cautionary note. *Quality & Quantity*, *44*(4), 761–775.

Biau, G. & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Birattari, M., Yuan, Z., Balaprakash, P., & Stützle, T. (2010). F-Race and Iterated F-Race: An Overview. In T. Bartz-Beielstein, M. Chiarandini, L. Paquete, & M. Preuss

(Eds.), *Experimental Methods for the Analysis of Optimization Algorithms* (pp. 311–336). Springer Berlin Heidelberg.

Bischl, B. & Lang, M. (2015). *Parallelmap: unified interface to parallelization back-ends*. R package version 1.3.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Jones, Z., ... Gallo, M. (2017). *Mlr: machine learning in r*. R package version 2.11.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., ... Jones, Z. M. (2016). Mlr Machine Learning in R. *Journal of Machine Learning Research*, *17*(170), 1–5.

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation*, *20*(2), 249–275.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678.

Böckenholt, U. & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181.

Bogg, T. & Roberts, B. W. (2004). Conscientiousness and health-related behaviors: a meta-analysis of the leading behavioral contributors to mortality. *Psychological bulletin*, *130*(6), 887.

Bolker, B. (2016). *Emdbook: support functions and data for "ecological models and data"*. R package version 1.3.9.

Bolt, D. M. & Johnson, T. R. (2009). Addressing Score Bias and Differential Item Functioning Due to Individual Differences in Response Style. *Applied Psychological Measurement*, *33*(5), 335–352.

Borkenau, P. & Ostendorf, F. (2008). *NEO-FFI: NEO-Fünf-Faktoren-Inventar nach Costa und McCrae* (2. Aufl.). Göttingen: Hogrefe.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199–231.

Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2015). *Randomforest: breiman and cutler's random forests for classification and regression*. R package version 4.6-12.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brown, A. & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502.

Brown, A. & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36.

Bühner, M. (2011). *Einführung in die Test-und Fragebogenkonstruktion*. Pearson Deutschland GmbH.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

Cabooter, E. (2010). *The impact of situational and dispositional variables on response styles with respect to attitude measures*. (Doctoral Dissertation). Faculty of Economics and Business Administration, Ghent University.

Cabooter, E., Millet, K., Weijters, B., & Pandelaere, M. (2016). The 'I'in extreme responding. *Journal of Consumer Psychology*, *26*(4), 510–523.

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological methods*, *21*(4), 603.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Churchill, S., Jessop, D., & Sparks, P. (2008). Impulsive and/or planned behaviour: Can impulsivity contribute to the predictive utility of the theory of planned behaviour? *British Journal of Social Psychology*, *47*(4), 631–646.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Costa, P. T. & MacCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*. Psychological Assessment Resources.

Crawford, J. R. & Henry, J. D. (2004). The positive and negative affect schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample. *The British Journal of Clinical Psychology / the British Psychological Society*, *43*(3), 245–265.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and psychological measurement*, *6*(4), 475–494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and psychological measurement*, *10*(1), 3–31.

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

Csárdi, G. & FitzJohn, R. (2016). *Progress: terminal progress bars*. R package version 1.1.2.

Dahl, D. B. (2016). *Xtable: export tables to latex or html*. R package version 1.8-2.

De Boeck, P. & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28.

Dolnicar, S. & Grün, B. (2007). How constrained a response: A comparison of binary, ordinal and metric answer formats. *Journal of Retailing and Consumer Services*, *14*(2), 108–122.

Dolnicar, S., Grün, B., & Leisch, F. (2011). Quick, simple and reliable: Forced binary survey questions. *International Journal of Market Research*, *53*(2), 231.

Eid, M. & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*(1), 20.

Engel, R. R. & Hathaway, S. R. (2003). *Minnesota multiphasic personality inventory-2: MMPI-2*. Huber.

Fahrenberg, J., Hampel, R., & Selg, H. (2001). FPI-R. *Das Freiburger Persönlichkeitsinventar. Manual*.

Faragher, E. B., Cass, M., & Cooper, C. L. (2005). The relationship between job satisfaction and health: a meta-analysis. *Occupational and environmental medicine*, *62*(2), 105–112.

Ferrando, P. J. & Condon, L. (2006). Assessing acquiescence in binary responses: IRT-related item-factor-analytic procedures. *Structural Equation Modeling*, *13*(3), 420–439.

Finn, J. A., Ben-Porath, Y. S., & Tellegen, A. (2015). Dichotomous versus polytomous response options in psychopathology assessment: Method or meaningful variance? *Psychological assessment*, *27*(1), 184.

Fox, J. & Hong, J. (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software*, *32*(1), 1–24.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Genz, A., Bretz, F., Miwa, T., Mi, X., & Hothorn, T. (2016). *Mvtnorm: multivariate normal and t distributions*. R package version 1.0-5.

GESIS. (2015). GESIS Panel Standard Edition (Version 6.0.0, Datenfile ZA5665). GESIS Datenarchiv: Köln.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Gollwitzer, M., Eid, M., & Jürgensen, R. (2005). Response styles in the assessment of anger expression. *Psychological assessment*, *17*(1), 56.

Greenleaf, E. A. (1992). Measuring Extreme Response Style. *Public Opinion Quarterly*, *56*(3), 328–351.

Gu, F., Little, T. D., & Kingston, N. M. (2013). Misestimation of reliability using coefficient alpha and structural equation modeling when assumptions of tau-equivalence and uncorrelated errors are violated. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(1), 30.

Harlow, L. L. & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*(4), 447.

Harrell, F. E., Jr. (2016). *Hmisc: harrell miscellaneous*. R package version 4.0-2.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning 2nd edition*. New York: Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity*. CRC press.

Jackson, D. N. & Messick, S. (1958). Content and style in personality assessment. *Psychological bulletin*, *55*(4), 243.

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the Behavioral

Indicators of Conscientiousness (BIC). *Journal of Research in Personality*, *44*(4), 501–511.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Jin, K.-Y. & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 116–138.

Kieruj, N. D. & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International journal of public opinion research*, *22*(3), 320–342.

Kieruj, N. D. & Moors, G. (2013). Response style behavior: question format dependent or personal style? *Quality & Quantity*, *47*(1), 193–211.

Komboz, B., Strobl, C., & Zeileis, A. (2016). Tree-Based Global Model Tests for Polytomous Rasch Models. *Educational and Psychological Measurement*.

Krohne, H. W., Egloff, B., Kohlmann, C.-W., & Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). *Diagnostica*, *42*(2), 139–156.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, *5*(3), 213–236.

Kuhn, M. & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kulas, J. T. & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43*(3), 489–493.

López-Ibáñez, M., Dubois-Lacoste, J., Cáceres, L. P., Birattari, M., & Stützle, T. (2016). The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, *3*, 43–58.

López-Ibáñez, M., Dubois-Lacoste, J., Cáceres, L. P., Stützle, T., Birattari, M., Yuan, E., & Balaprakash, P. (2015). *Irace: iterated racing procedures*. R package version 1.07.

MacKenzie, S. B. & Podsakoff, P. M. (2012). Common method bias in marketing: causes, mechanisms, and procedural remedies. *Journal of Retailing*, *88*(4), 542–555.

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

McEachan, R. R. C., Conner, M., Taylor, N. J., & Lawton, R. J. (2011). Prospective prediction of health-related behaviours with the theory of planned behaviour: A meta-analysis. *Health Psychology Review*, *5*(2), 97–144.

Meiser, T. & Machunsky, M. (2008). The personal structure of personal need for structure. *European Journal of Psychological Assessment*, *24*(1), 27–34.

Milborrow, S. (2017). *Rpart.plot: plot 'rpart' models: an enhanced version of 'plot.rpart'*. R package version 2.1.1.

Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, *21*(2), 271–298.

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., ... Barry, O., et al. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin*, *38*(11), 1423–1436.

Münzer, S. & Hölscher, C. (2011). Entwicklung und Validierung eines Fragebogens zu räumlichen Strategien. *Diagnostica*, *57*(3), 111–125.

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality Predictors of Extreme Response Style. *Journal of Personality*, *77*(1), 261–286.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Ostendorf, F. & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R)*. Göttingen, Germany: Hogrefe.

Peabody, D. (1962). Two components in bipolar scales: Direction and extremeness. *Psychological Review*, *69*(2), 65.

Plieger, T., Montag, C., Felten, A., & Reuter, M. (2014). The serotonin transporter polymorphism (5-HTTLPR) and personality: response style as a new endophenotype for anxiety. *International Journal of Neuropsychopharmacology*, *17*(6), 851–858.

Plieninger, H. (2016). Mountain or Molehill? A Simulation Study on the Impact of Response Styles. *Educational and Psychological Measurement*.

Plieninger, H. & Meiser, T. (2014). Validity of Multiprocess IRT Models for Separating Content and Response Styles. *Educational and Psychological Measurement*, *74*(5), 875–899.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2011). Sources of Method Bias in Social Science Research and Recommendations on How to Control It. *Annual Review of Psychology*, *63*(1), 539–569.

Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, *104*(1), 1–15.

R Core Team. (2016a). *Foreign: read data stored by minitab, s, sas, spss, stata, systat, weka, dbase, ...* R package version 0.8-67.

R Core Team. (2016b). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality*, *41*(1), 203–212.

Raykov, T. (2001). Bias of coefficient afor fixed congeneric measures with correlated errors. *Applied psychological measurement*, *25*(1), 69–76.

Revelle, W. (2017). *Psych: procedures for psychological, psychometric, and personality research*. R package version 1.6.12.

Rizopoulos, D. (2013). *Ltm: latent trait models under irt*. R package version 1.0-0.

Rorer, L. G. (1965). The great response-style myth. *Psychological bulletin*, *63*(3), 129.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*(1), 75–92.

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). Münster: Waxmann.

Schauberger, G. (2016). *Diflasso: a penalty approach to differential item functioning in rasch models*. R package version 1.0-2.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, *124*(2), 262.

Schmukle, S. C., Egloff, B., & Burns, L. R. (2002). The relationship between positive and negative affect in the Positive and Negative Affect Schedule. *Journal of Research in Personality*, *36*(5), 463–475.

Sieber, K. O. & Meyers, L. S. (1992). Validation of the MMPI—2 Social Introversion subscales. *Psychological assessment*, *4*(2), 185.

Simon, R. (2007). Resampling strategies for model assessment and selection. In *Fundamentals of data mining in genomics and proteomics* (pp. 173–186). Springer.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, *9*(1), 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 25.

Strobl, C., Kopf, J., & Zeileis, A. (2013). Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model. *Psychometrika*, *80*(2), 289–316.

Sutin, A. R., Ferrucci, L., Zonderman, A. B., & Terracciano, A. (2011). Personality and obesity across the adult life span. *Journal of personality and social psychology*, *101*(3), 579.

Terracciano, A., Sutin, A. R., McCrae, R. R., Deiana, B., Ferrucci, L., Schlessinger, D., ... Costa Jr, P. T. (2009). Facets of personality linked to underweight and overweight. *Psychosomatic medicine*, *71*(6), 682.

Therneau, T., Atkinson, B., & Ripley, B. (2015). *Rpart: recursive partitioning and regression trees*. R package version 4.1-10.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Tutz, G. (2011). *Regression for categorical data*. Cambridge University Press.

Tutz, G. & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21–43.

Tutz, G., Schauberger, G., & Berger, M. (2016). Response Styles in the Partial Credit Model. Technical Report Number 196, Department of Statistics, University of Munich.

Vaerenbergh, Y. V. & Thomas, T. D. (2013). Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217.

Van der Linden, W. J. (2016). *Handbook of item response theory*. CRC Press.

Van Der Linden, W. J. & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory* (pp. 1–28). Springer.

Verplanken, B. & Herabadi, A. (2001). Individual differences in impulse buying tendency: Feeling and no thinking. *European Journal of personality*, *15*(1), 71–83.

Verplanken, B., Herabadi, A. G., Perry, J. A., & Silvera, D. H. (2005). Consumer style and health: The role of impulsive buying in unhealthy eating. *Psychology & Health*, *20*(4), 429–441.

Von Davier, M. (2001). WINMIRA 2001. *St. Paul, MN: Assessment Systems Corporation*.

Warnes, G. R., Bolker, B., & Lumley, T. (2015). *Gtools: various r programming tools*. R package version 3.5.0.

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

Weijters, B. & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737–747.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247.

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The Individual Consistency of Acquiescence and Extreme Response Style in Self-Report Questionnaires. *Applied Psychological Measurement*, *34*(2), 105–121.

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological methods*, *15*(1), 96.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*(3), 409–422.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*(6), 956–972.

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences*, *34*(2), 69.

Wetzel, E., Böhnke, J. R., & Rose, N. (2016). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement*, *76*(2), 304–324.

Wetzel, E. & Carstensen, C. H. (2015). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*.

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*(2), 178–189.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*(3), 279–291.

Whiteside, S. P. & Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and individual differences*, *30*(4), 669–689.

Wickham, H. (2016a). *Plyr: tools for splitting, applying and combining data*. R package version 1.8.4.

Wickham, H. (2016b). *Reshape2: flexibly reshape data: a reboot of the reshape package*. R package version 1.4.2.

Wickham, H. (2017). *Tidyr: easily tidy data with 'spread()' and 'gather()' functions*. R package version 0.6.1.

Wickham, H. & Chang, W. (2016). *Ggplot2: create elegant data visualisations using the grammar of graphics*. R package version 2.2.1.

Wickham, H. & Francois, R. (2016). *Dplyr: a grammar of data manipulation*. R package version 0.5.0.

Wickham, H. & Miller, E. (2016). *Haven: import and export 'spss', 'stata' and 'sas' files*. R package version 1.0.0.

Xie, Y. (2016). *Knitr: a general-purpose package for dynamic report generation in r*. R package version 1.15.1.

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Manuscript submitted for publication.

Yin, P. & Fan, X. (2001). Estimating R 2 shrinkage in multiple regression: a comparison of different analytical methods. *The Journal of Experimental Education*, *69*(2), 203–224.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., & Kopf, J. (2016). *Psychotree: recursive partitioning based on psychometric models*. R package version 0.15-1.

Zettler, I., Lang, J. W., Hülsheger, U. R., & Hilbig, B. E. (2015). Dissociating Indifferent, Directional, and Extreme Responding in Personality Data: Applying the Three-Process Model to Self-and Observer Reports. *Journal of personality*, *84*(4), 461–472.

Ziegler, M., Dietl, E., Danay, E., Vogel, M., & Bühner, M. (2011). Predicting training success with general mental ability, specific ability tests, and (Un) structured interviews: A meta-analysis with unique samples. *International Journal of Selection and Assessment*, *19*(2), 170–182.

Ziegler, M. & Kemper, C. (2013). Extreme response style and faking: Two sides of the same coin. In *Interviewers deviations in surveys–impact, reasons, detection and prevention* (pp. 217–233). Frankfurt: Lang.

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – Five misunderstandings and ways to overcome them. *Journal of Individual Differences*. Measuring Psychological Constructs with Short Scales: Positive Outlooks and Caveats, *35*(4), 185–189.