On Measuring Some of the People Some of the Time with Some of the Items: The Search for Stability and Variation in Item Sets

Kumulative Dissertation zur Erlangung des akademischen Grades

Doktor der Philosophie (Dr. phil.)

Vorgelegt im Fachbereich Humanwissenschaften der Universität Kassel

Von Gabriel Olaru, M.Sc.

Kassel, Februar 2019

Erstgutachter:	Prof. Dr. Ulrich Schroeders
Zweitgutachter:	Prof. Dr. Oliver Wilhelm
Drittgutachter:	Prof. Dr. Johannes Zimmermann
Tag der Disputation:	17.07.2019

Table of Content

List of Tables

List of Figures

Abstract

Psychological assessment is shaped by the items used and the persons assessed. Both items and persons typically represent a random or representative sample of a much larger item and person pool. However, most of the focus on psychological measurement rests on the person sampling side. Item sampling from larger item pools is still a black box. In this dissertation, I present the advantages of new state-of-the-art item and person sampling procedures in the context of personality development research (manuscript 1). Measurement in personality development faces many psychometric problems. First, the theoretically assumed measurement models do not fit the data when tested with confirmatory factor analysis. Second, measurement invariance across age, which is necessary for a meaningful interpretation of age-associated personality differences, is rarely accomplished. And third, the continuous moderator variable age is often artificially categorized. I show how Ant Colony Optimization can be used to select indicators that provide adequate model fit and measurement invariance across age (manuscript 2). I also apply a combination of the item sampling approach Genetic Algorithm and the person sampling approach Local Structural Equation Modeling to identify the items that provide the most prototypical measurement of personality within restricted age samples (manuscript 3). These manuscripts address two sides of the measurement invariance problem (i.e., the DIF paradox): If normative age-associated differences should be studied, measurement invariant indicators across age need to be selected. If the measurement within restricted age ranges should be optimized, indicators that maximize model fit and *measurement variance* across age need to be sampled. The novel item sampling procedures can be applied in any assessment context to optimize psychometric requirements (e.g., model fit, reliability, difficulty). The person sampling method Local Structural Equation Modeling can also be applied to any measurement to study the robustness across continuous moderator variables (e.g., cognitive abilities, SES). In the epilogue, I discuss implications for personality measurement and provide an outlook on future research.

Zusammenfassung

Psychologische Messungen sind geprägt von den verwendeten Items und Personen. Beide stellen in der Regel eine zufällige oder repräsentative Stichprobe einer viel größeren Item- oder Personenpopulation dar. In der psychologischen Forschung konzentriert man sich jedoch oft nur auf die Personenziehung. Der Einfluss und die Vorteile von Itemsampling werden oft nicht beachtet. In dieser Dissertation stelle ich neue Item- und Personensamplingverfahren für die Persönlichkeitsentwicklungsforschung vor (Manuskript 1). Messungen in der Forschung zur Persönlichkeitsentwicklung sind aus mehreren Gründen problematisch: 1) Konfirmatorische Faktorenanalysen lehnen die theoretisch fundierten Messmodelle ab. 2) Messinvarianz übers Alter, die für eine sinnvolle Interpretation altersbedingter Persönlichkeitsunterschiede notwendig ist, wird selten erreicht. 3) Die kontinuierliche Variable Alter wird oft künstlich kategorisiert. In dieser Dissertation verwende ich Ant Colony Optimization, um Persönlichkeitsitems auszuwählen, die eine adäquate Modellanpassung und Messinvarianz über das Alter hinweg bieten (Manuskript 2). Ich wende eine Kombination aus dem Itemziehungsverfahren Genetischer Algorithmus und der Personengewichtungsmethode Lokale Strukturgleichungsmodellierung an, um Items zu identifizieren, die Alter-prototypische Messungen der ermöglichen (Manuskript 3). Beide Manuskripte behandeln zwei Aspekte des Messinvarianzproblems: Wenn Mittelwertsverläufe untersucht werden sollen, müssen messinvariante Items über das Alter hinweg ausgewählt werden. Wenn die Messung in eingeschränkten Altersbereichen optimiert werden soll, müssen Items zur Maximierung der Modellanpassung und *Messvarianz* übers Alter gezogen werden. Diese neuartigen Item- und Personensamplingverfahren können auf jede Art von psychologischer Messung angewendet werden um psychometrische Eigenschaften der Messung zu optimieren (z. B., Modelfit, Reliabilität, Schwierigkeit) und um die Robustheit über kontinuierliche Moderatoren zu untersuchen (z. B. Intelligenz, sozi-ökonomischer

Status). Im Epilog diskutiere ich Implikationen für die Persönlichkeitsmessung und gebe einen Ausblick auf zukünftige Forschungsvorhaben.

I

Prologue

The reason why some persons prefer to spend their evenings alone or with a small group of close friends, whereas others would rather go to a party or other social events can be attributed to differences in personality, in this case Extraversion. Extraversion and the other personality traits – Neuroticism, Openness, Agreeableness and Conscientiousness based on the Big Five (Goldberg, 1990) or Five Factor Model of Personality (Costa & McCrae, 1995) – are typically understood as dispositional traits that influence – among others – our behavior, preferences and attitudes (for the sake of simplicity, I will only refer to behaviors in the rest of the prologue). These traits are not categorical in nature (e.g., introverted vs. extraverted), but are normally distributed among a continuum. As such, most people will have average levels of the personality traits with extreme tendencies being rare or caused by underlying psychological illnesses (e.g., extreme introversion as a symptom of depression). The relation between personality and the corresponding behaviors is not deterministic, but probabilistic: People with higher levels of Extraversion will be more likely to go to social events, whereas people with low levels will prefer to spend time alone under otherwise similar circumstances. However, this doesn't mean that people with low Extraversion levels will never go to social events, the likelihood of doing so is just comparatively low.

Similar to other latent constructs, personality cannot be measured directly, but only through observable related behaviors. This is typically done by asking people to rate their agreement with adjectives (e.g., "I am gregarious"), statements that describe personality related behaviors (e.g., "I often go to parties"), emotions (e.g., "I am often sad"), interests (e.g., "I like to go to the ballet"), attitudes (e.g., "You cannot trust anyone"), or similar. Typically, some form of aggregate (i.e., sum or mean value) across these questions is then used as an indicator for the underlying personality trait (of course this notion is only correct if these indicators are unidimensional measures of the personality trait, for more details see manuscripts). If Sarah reports higher agreement with the items "I often go to parties", "I like meeting new people" and "I have many friends" than for instance Michael does, we infer that

her Extraversion level is higher than Michael's. In the following, I provide an uncritical overview of the current conceptions regarding personality assessment and development, before challenging these notions and presenting new methodological advances that can help improve the examination of personality and personality development across age.

Personality Assessment

Modern personality research originated in the lexical analysis of the trait descriptive language (Allport & Odbert, 1936; Goldberg, 1990; John, Angleitner, & Ostendorf, 1988). This approach rests on the assumption that language provides an exhaustive list of all relevant inter-individual differences (i.e., relevant enough to be named, and as such descriptions of differences in personality. *Webster's Unabridged Dictionary of the English Language* was thus searched for all terms (nouns and adjectives) capable of describing human temperament (Allport & Odbert, 1936; Norman, 1967). Over 18,000 terms were gathered and reduced to around 4,500 "stable traits". Later factor analyses of a subset of the adjective terms revealed five underlying factors that were assumed to exhaustively describe individual differences in personality (Digman & Takemoto-Chock, 1981; Fiske, 1949; Norman, 1963; Thurstone, 1934; Tupes & Christal, 1961). These recurring five factors were later termed the Big Five factors of personality (Goldberg, 1990).

The lexical analysis of personality descriptive language represents a purely inductive approach to personality assessment: A large and exhaustive set of personality descriptive indicators are initially gathered and subsequently reduced to a smaller number based on a wide variety of criteria, such as frequency of use, redundancy or centrality in factor analytic procedures. Modern measures of personality, such as the NEO-PI-R (Costa & McCrae, 1992) or the Big Five Inventory (John et al., 1991), left this purely inductive approach and used a combination of deduction and induction: First the traits to be measured were defined by the researchers. Based on this delineation of the traits and relevant behaviors, a large set of items with varying item types (e.g., adjectives, emotions, behaviors, interests) were created

(deduction). This large set of items was subsequently reduced based on a combination of expert judgment and statistical analysis, most commonly based on correlations with other personality inventories or principal component analysis with the goal of identifying the most central items to the assumed factors (inductive; I will disscuss this approach in more detail in a later section of this dissertation). Apart from the development process, newer personality inventories differ most strongly from earlier adjective marker questionnaires in their use of a broad set of indicators (e.g., behaviors, emotions, cognitions). There is no clear consensus on how broad this set of item types should be, with some researchers arguing that homogenous item sets containing only adjectives, emotional/cognitive patterns and behavioral habits (e.g., BFI-2 and Eyseck Personality Questionnaire; Eysenck & Eysenck, 1975; Soto & John, 2017) are best suited to measure personality, whereas the very popular NEO-PI-R (Costa & McCrae, 1992) and HEXACO inventories (Ashton & Lee, 2009; Lee & Ashton, 2004) apply a much broader and more heterogeneous set of item types (including e.g., interests, evaluations, world views).

As the Big Five factors represent very broad trait domains, newer conceptualizations of personality propose an additional level of more specific facet traits below the broad trait domains (e.g., Extraversion facets: Warmth, Gregariousness, Assertiveness, Activity, Positive Feelings; Costa & McCrae, 1995). The facets (e.g., Gregariousness) are much more specific than the broad trait domains (e.g., Extraversion), and as such, delineating relevant behaviors for the facet traits is somewhat easier than for the broad higher-order factors (e.g., Extraversion). However, no consensus on a common facet structure exists across personality inventories (see e.g., Costa & McCrae, 1995; Soto & John, 2009, 2017), and the construct coverage of personality inventories is still very much determined by the construct definitions of the questionnaire developer (Angleitner, John, & Löhr, 1986). The unfortunate consequence is that the comparability of findings across different personality questionnaires is still questionable. The number of facets assessed by the questionnaire are dependent on the

underlying conception of personality applied by the developer, but also by considerations regarding the length of the inventory, as more facets result in a higher number of items needed to measure these reliably.

The length of measures of the Big Five or Five Factor Model also differs strongly across inventories. Item numbers per factor range from one or two (Five and Ten Item Personality Inventory; Gosling, Rentfrow, & Swann, 2003) to 60 (IPIP-NEO-300; Goldberg et al., 2006). The shortest Big Five measure also capturing the facet level is the BFI-2 (Soto & John, 2017), using twelve items per factor (four per facet; but also see TSDI-42 with six to nine items per factor; Olaru, Witthöft, & Wilhelm, 2015). The decision on which inventory to use typically depends on the researchers' considerations regarding time constraints or participant fatigue (e.g., large scale panel studies, such as SOEP, will generally use very brief inventories due to the large number of measurements applied; Gerlitz & Schupp, 2005), as well as measurement precision or construct coverage (with longer inventories providing a more precise and broader measure of personality). Very brief inventories are typically developed with the goal of maintaining as much construct coverage as possible with the restricted number of indicators, thus relying on expert judgments on construct coverage and external correlations with longer inventories during the item selection process (Gosling et al., 2003). After creating broad item pools based on expert judgments, longer inventories are typically derived with the goal of maximizing the reliability of the scales, thus using Principal Component Analysis to select the indicators that provide the most central and homogenous measures of the extracted principal components (Costa & McCrae, 1992). Inventories with medium length will typically apply a combination of all the aforementioned criteria to maintain comparable construct coverage to longer scales, while also improving the reliability of the scale (Soto & John, 2017). But I want to point out that construct coverage is typically addressed by selecting items with high correlation to scale scores of longer inventories, which is arguably similar to the highest main loading criterion used to improve reliability (i.e., both

focus on the centrality of the items). A better indicator of construct coverage would be to maximize the correlation between the scale (but ideally factor) scores of the short and long instrument (Yarkoni, 2010). Unfortunately, none of these development processes address the issue of model fit in classical confirmatory analysis testing (i.e., whether the theoretical model of personality fits the empirical data). The exceptions that do test the models using confirmatory factor analysis typically dismiss problematic model fit, which is often encountered when testing personality models (Costa & McCrae, 1995; Donnellan, Oswald, Baird, & Lucas, 2006). I will address this issue and consequences thereof in a later section of the prologue, as well as within the manuscripts.

Personality Development

Comparisons of personality levels across age showed some considerable change in personality over the lifespan (Roberts & DelVecchio, 2000; Roberts, Walton, & Viechtbauer, 2006). Across several different questionnaires, cross-sectional and longitudinal studies, people have shown to generally become more Conscientious, Agreeable and Emotionally Stable across the course of life (Roberts et al., 2006), which is generally interpreted as people becoming more mature with age. These trends of personality development are normative, as they affect the entire population (e.g., Helson & Moane, 1987; McCrae et al., 2000) and are independent of sex (Helson, Jones, & Kwan, 2002; Roberts et al., 2006). Personality seems to show the highest plasticity in young age, but change can also be found in old age (Roberts & DelVecchio, 2000). Even though these normative trends of "maturation" have been consistently found in examinations of the general population, there are also considerable differences between the intra-individual developmental trajectories: The rank-order consistency of personality typically found in longitudinal studies ranges from .31 in childhood to .74 in old age (Roberts & DelVecchio, 2000). Reason for inter-individual differences in the developmental trajectories are manifold: Normative trends are understood to be driven by (biological) maturation (Costa & McCrae, 2000) and common social roles associated with

different expectations and obligations, such as becoming a parent or transitioning from school to work life (Roberts & Mroczek, 2008; Specht, Egloff, & Schmukle, 2011; Wrzus & Roberts, 2017). Differences in intra-individual age trajectories are caused by inter-individual differences in the selection of and reaction to situations or life events (Bleidorn, Hopwood, & Lucas, 2018; Löckenhoff, Terracciano, Patriciu, Eaton, & Costa Jr, 2009; Roberts & Mroczek, 2008; Specht et al., 2011; Wrzus & Roberts, 2017) as well as interventions (e.g., therapy, training; Roberts et al., 2017).

Personality development is usually studied by comparing average scale scores across different age points (typically age groups). Differences in the scores between age points are then attributed to meaningful normative development trends. However, little attention is paid to other types of age-associated personality differences across age: among mean-levels of the personality factors, differences can also be found in the structure and variance of the personality factors across age (Allemand, Zimprich, & Hertzog, 2007; Caspi & Roberts, 2001). In the following, I describe how the different types of age-associated personality differences (in a cross-sectional setting) can be identified with current psychometric methods and how these variations can be interpreted. Figure 1 presents a higher-order factor model with corresponding model parameters that are prone to change. In a cross-sectional context, age-associated differences can be categorized in three categories: a) absolute differences, which result in factor mean variations across age, b) structural differences, which affect factor loadings, factor covariance, factor and item intercepts, as well as c) divergence, which can be observed in as an increase or decrease in factor variance across age.

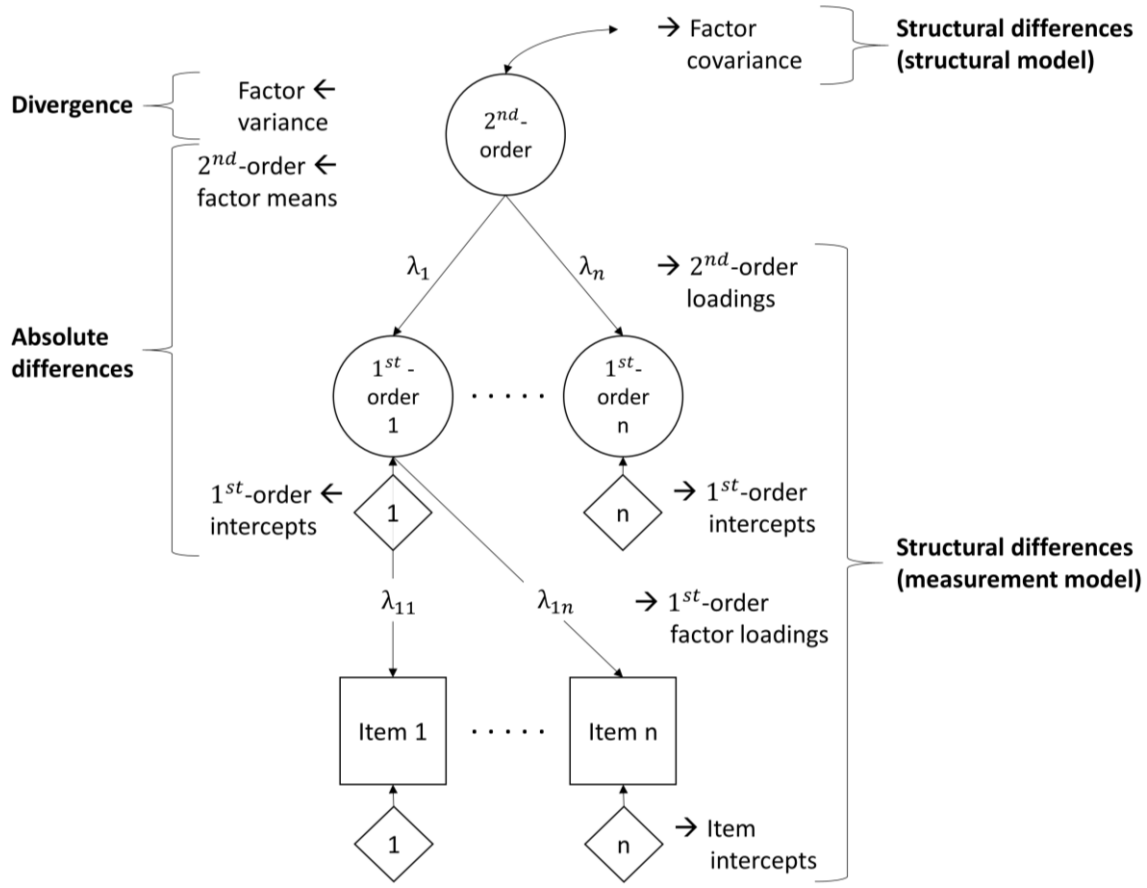


Figure 1. Age-associated differences in a higher-order factor model. Latent variables are depicted as circles. Manifest variables are represented by squares. Intercepts are presented as diamonds. Residual variances of the items and first-order factors are not depicted. Note that absolute and structural differences in the first-order factors intercepts are not equivalent, as the first refers to mean-level patterns across age, whereas the latter refers to the homogeneity/ heterogeneity of the mean-level patterns across first-order factors (see Table 1 for more details).

Table 1 provides an overview of all types of age-associated differences relevant in cross-sectional studies, where these differences can be observed in personality models, and how variations can be interpreted. As can be seen, a wide variety of age-associated differences can be studied, all with different implications to the understanding of personality development. Theoretically, these variations can occur independently of each other (e.g., factor means can remain stable despite an increase in factor variance; factor-loadings can change without affecting the mean level of the overarching factor). However, the interpretation of age-associated differences always relies on a comprehensive evaluation of all types of change, with structural stability being the prerequisite of meaningful interpretations

of other types of change: If the composition of the factors/facets is not equivalent across age, factors and corresponding mean scores may represent different traits (e.g., Gregariousness being central in young age; Activity in middle age; and Positive Feelings in old age). It is thus paramount to first ensure that structural stability is given before examining other types of change. Table 2 provides a representative overview of the literature on age-associated differences, as well as the type of difference examined.

Table 1. *Types of Personality Differences and Model Parameters Affected*

Type of difference	Parameter affected	Psychometric reason An increase decrease in the …	Interpretation
Absolute	Second-order factor mean	… mean-value of **all items** of a common factor	Normative developmental trends that affect the entire population. *Example*: Increasing Emotional Stability, Conscientiousness and Agreeableness mean-values across age (Roberts, 2006)
Structural	First-order factor loading	… covariance between **an item** and the rest of the facet scale	The relationship between the behavior and other trait-related behaviors (and consequently the trait) changes. Behaviors can become more or less central to the trait (higher or lower loadings). *Example*: The item "I love the thrill of roller coasters" is a prototypical Excitement-Seeking item in young age, but irrelevant in old age (manuscript 3)
	Second-order factor loading	… covariance between **a facet** and the other facets of the factor	The facet becomes more or less central of the overarching factor. *Example*: Excitement-Seeking is the most relevant facet of Extraversion in young age, but becomes less central (i.e., less related) to the factor with increasing age (hypothetical)
	Item intercept	… mean-value of **an item** of a facet scale, independent of the other facet scale items	Variations in the propensity of an item that cannot be explained by normative differences in the underlying personality trait. *Example*: The mean-level of Gregariousness and all related items remains stable across age, however the propensity to "go to parties" decreases due to a lack of "parties" in higher age (manuscript 2)
	First-order factor intercept	… responses to **a facet**, independent of the other facets of the factor	Normative differences in the facet traits that cannot be found in the other facets. *Example*: An increase in the Assertiveness facet (Extraversion) mean value across age, but not in the Sociability facet (Roberts, 2006)
	Factor covariance	… covariance between **different factors scales**	The relationship between personality traits changes across age. *Example*: Extraversion and Agreeableness are independent traits in young years, but become more social intertwined in older age as friendly seniors also maintain a larger social network (hypothetical)
Divergence	Factor variance	… **inter-individual differences** in the responses to items of a common factor	Inter-individual differences in the levels of a trait change across age. *Example*: While strong inter-individual variations exist in Emotional Stability in young age, older individuals experience much less negative emotions and inter-individual differences decrease as a result of this (Charles & Carstensen, 2010)

Table 2. *Overview of Personality Development Studies*

Study	Sample size	Age in years (Duration in years)	Countries	Inventory (Number of items)	Facets	*N*	*E*	*O*	*A*	*C*	Types of differences examined*
Cross sectional without measurement invariance testing											
McCrae et al., 1999	7,363	18 – 84	Germany, Italy, Portugal, Croatia, South Korea	NEO-PI-R (240)	NEO	--	---/0	---	+	+++	Absolute
Roberts, Walton, & Viechtbauer, 2006	92 studies	10 – 101	International	Meta-analysis	Two for *Extraversion*	---	+++/0	+	+++	+++	Absolute
Soto, John, Gosling, & Potter, 2011	1,267,218	10 – 65	English speaking countries	BFI (44)	Two per BFF (Soto, John, 2009)	--	0	0	++	++/0	Absolute
Specht, Egloff, & Schmuckle, 2011*	14,718	16 – 96	Germany	BFI-S (15)	none	0	-	---	+	+++	Absolute
Srivastava, John, Gosling, & Potter, 2003	132,515	21 – 60	USA, Canada	BFI (44)	none	--	0	-	++	++	Absolute
Cross-sectional with measurement invariance testing											
Allemand, Hertzog, & Zimprich, 2007	865	42 – 64	Germany	NEO-FFI (60)	none	0	-	-	+	0	Absolute; structural; divergence
Allemand, Zimprich, & Hendriks, 2008	2,494	16 – 91	Netherlands	FFPI (50)	none	0	-		+	++	Absolute; structural; divergence
Nye, Allemand, Gosling, Potter, & Roberts, 2015	31,452	20 – 50	English speaking countries	BFI (44)	data driven	--	0	+	+	++	Absolute; structural
Brandt, Becker, Tetzner, Brunner, Kuhl, & Maaz, 2018	19,879	11 – 84	Germany	BFI-S (16)	none						Structural, divergence

Study	Sample size	Age in years (Duration in years)	Countries	Inventory (Number of items)	Facets	*N*	*E*	*O*	*A*	*C*	Types of differences examined*
Longitudinal											
Bleidorn, Kandler, Riemann, Angleitner, & Spinath, 2009	187 twins	18 – 59 (10)	Germany	NEO-PI-R (240)	NEO	--	+/-	-/0	+	++	Absolute; structural; divergence
Donnellan, Conger, & Burzette, 2007	432	18 – 27 (10)	USA	MPQ (155)	MPQ	---	-			++	Absolute; divergence
Helson, Jones, & Kwan, 2002	368	21 – 75 (40)	USA	CPI (468)	CPI		--/0			+	Absolute; divergence
Robins, Fraley, Roberts, & Trzesniewski, 2001	270	18 – 23 (4)	USA	NEO-FFI (60)	none	--	0	+	++	+	Absolute; structural; divergence
Specht, Egloff, & Schmuckle, 2011*	14,718	16 – 96 (4)	Germany	BFI-S (15)	none	0	0	-	-	0	Absolute; structural; divergence
Terracciano, McCrae, Brant, & Costa, 2005	1,944	20 – 96 (5)	USA	NEO-FFI (240)	NEO	--	--/+	--/0	++	+++ /0	Absolute; divergence

Note. N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness. Columns N through C represent findings on absolute differences, based on effect size Cohen's *d* (1988): - = small negative effect; -- = medium negative effect; --- = strong negative effect; + = small positive effect; ++ = medium positive effect; +++ = strong positive effect. If first-order factors followed different mean-level trajectories, all varying effects are listed and separated by a dash. Longitudinal types of personality change (i.e., intra-individual) are not listed, but note that divergence is examined via rank-order stability instead of factor variance in this context. BFI = Big Five Inventory (John, Donahue, & Kentle, 1991); BFI-S = Big Five Inventory – SOEP (Schupp, Gerlitz, 2014).; CPI = California Psychological Inventory (Gough, 1956); FFPI = Five Factor Personality Inventory (Hendriks, Hofstee, & De Raad, 1999); MPQ-BF = Multidimensional Personality Questionnaire Brief Form (Patrick, Curtin, & Tellegen, 2002); NEO-PI-R = Revised NEO Personality Inventory (Costa, & McCrae, 2008); NEO-FFI = Five Factor NEO Personality Inventory (Costa & McCrae, 1989); * Specht , Egloff, & Schmuckle, 2011 examined both cross-sectional and longitudinal age-differences.

Table 2 only provides a small overview of all studies on personality development across age, and numerous more studies exist that only examine absolute age-associated differences in a cross-sectional setting. In general, structural differences in personality across age are rarely studied (but see Specht, Luhmann, & Geiser, 2014; Tackett et al., 2012 for profile and hierarchical factor analysis of personality). As mentioned earlier, the structural stability is a prerequisite for the unbiased comparison of normative personality levels across age. In other words, the measurement of the personality factors has to be equivalent across age to ensure that the same traits are measured at all age points (Borsboom, 2006b; Guenole & Brown, 2014).

Testing whether the measurement is equivalent across age is also known as measurement invariance testing (Cheung & Rensvold, 1999; Meredith, 1993; Vandenberg & Lance, 2000). In measurement invariance testing, the equivalence of model parameters across a moderator (age in this case) is tested. In a cross-sectional context, this is typically done using Multi-group Confirmatory Factor Analysis (MGCFA; e.g., Allemand, Zimprich, & Hendriks, 2008; Allemand et al., 2007; Brandt et al., 2018; Nye, Allemand, Gosling, Potter, & Roberts, 2016). Measurement equivalence is tested in MGCFA by estimating and comparing models with increasing equality constraints across age groups (Schroeders & Gnambs, 2018). If model fit decreases due to the additional parameter constraints, measurement invariance is only supported for the less restrictive model (given that the less restrictive model yields adequate model fit). Typically, a model with the same model structure but no additional constraints across groups is estimated as a baseline model (= configural measurement invariance). The only requirement at this point is sufficient overall model fit. Subsequently, factor loadings are constrained to equality across groups (= metric measurement invariance). Model fit is then compared to the configural model. If the decrease in model fit is sufficiently low and overall model fit is still satisfactory, metric measurement invariance is supported. This procedure is repeated with additional item intercept constraints (= scalar measurement

invariance) and equal item residual variances across groups (= strict measurement invariance). Depending on the research questions, other model parameters such as factor variances, correlations and second-order factor loadings can be constrained to equality to examine structural personality differences at the latent level.

The DIF-Paradox

From a psychometric perspective, it is desirable to have little to no age effects on the measurement and structure of the personality traits. This is the prerequisite for mean-level comparisons across age. However, as mentioned earlier, normative change is only one aspect of personality development. A lack of measurement invariance can also be seen as a sign of meaningful differences in the personality structure or relevant behaviors across age (e.g., Huang, Church, & Katigbak, 1997). Church and colleagues (2011) labeled these concurring – but potentially complementing – perspectives on measurement invariance the DIF-paradox: On the one hand, measurement invariant indicators are needed for mean-level comparisons, and as such it is desirable to eliminate non-invariant indicators from the model. On the other hand, these problematic indicators can indicate meaningful differences in personality related behaviors across age. By removing non-invariant indicators, this information is lost. However, by retaining these items, normative differences cannot be meaningfully examined across age. Both these perspectives are meaningful perspectives on personality development. Consequently, one can decide to develop measures that are applicable and thus comparable across broad age ranges or instead create assessments that maximize the measurement precision and construct coverage at specific age points by including corresponding cues (e.g., school, work, interests) and age-appropriate situational demands (Rauthmann, Sherman, & Funder, 2015) – however at the cost of comparability across age.

Issues in Personality Development Research

Due to the predominant focus on normative developmental trends, the first perspective of the DIF-paradox is dominant in personality development research. When measurement

invariance testing is applied, it is often done to support examinations on the normative level (e.g., Allemand et al., 2008; Nye et al., 2016). For such mean level comparisons, scalar measurement invariance (i.e., equal factor loadings and item intercepts across age) and adequate model fit is required. Unfortunately, neither sufficient model fit or measurement invariance is typically achieved using broad self-report measures of personality. A number of data driven modifications are thus typically applied to improve the psychometric properties of the model. These include parceling of items into aggregates (Allemand et al., 2008, 2007; Small, Hertzog, Hultsch, & Dixon, 2003), data driven modifications to the measurement models (Nye et al., 2016; Small et al., 2003) or freeing parameters for partial measurement invariance (Brandt et al., 2018). Some researchers also recommend using less restrictive testing procedures, such as *Exploratory Structural Equation Modeling* (ESEM; Asparouhov & Muthén, 2009; Brandt et al., 2018; Morin, Marsh, & Nagengast, 2013), which allows for cross-loadings between all items and factors. While all these procedures will improve model fit, they do not eliminate model misfit, but incorporate it into the model: Parceling will increase model fit by masking violations of unidimensionality and measurement invariance at the item level (Little, Cunningham, Shahar, & Widaman, 2002; Little, Rhemtulla, Gibson, & Schoemann, 2013). Partial measurement invariance or other data driven freeing of model parameters in specific groups will also increase overall model fit and decrease model fit differences between measurement invariance levels, but are often theoretically hard to justify and include misfit into the model instead of removing it. The added model misfit will then severely bias the resulting estimates at the factor level (Guenole & Brown, 2014). ESEM suffers from similar downsides: The number and magnitude of cross-loadings can be very high (see e.g., Brandt et al., 2018) and theoretical support for the additional parameters may be lacking.

Item and Person Sampling

One of the reasons less restrictive model testing procedures are so often used is in my opinion related to the dominant perspective on item sampling in personality research: The item sets presented by many popular personality inventories are seen as a fixed gold standard of personality measurement instead of a (arbitrary selection) of items from the item universe of all personality items (Loevinger, 1957). Consequently, if the model does not fit the items as expected, then the model must be wrong (Borkenau & Ostendorf, 1990; Costa & McCrae, 1995; Marsh et al., 2010; Vassend & Skrondal, 1997). In this dissertation, I want to dismiss this notion of "gold standard" item sets and argue that items should be considered as samples from a larger item population similar to how people are sampled. From the perspective of the Generalizability Theory (Brennan, 2001), responses to personality measurements (or psychological measurement in general) are the product of the items used, the persons assessed, and the measurement occasion (which is most relevant in longitudinal settings), as well as interactions between these sources of variance. In cross-sectional personality development research, fixed item sets (i.e., personality questionnaires) are answered by person samples of varying age. The results of these analyses (typically mean-level differences across age groups) are assumed to arise due to age differences between participants (e.g., age groups). Item effects, as well as the interaction effects between items and participants of different age, are typically neglected in this context. As such, the studies rely on the two assumptions that a) the items used are representative and unidimensional indicators of the underlying traits (i.e., the personality model fits the data independent of age) and b) the measurement is equivalent across participants of different age (i.e., is measurement invariant). However, there is no reason to believe that items originally selected based on main loadings in Principal Component Analysis – often applied on student samples – represent the ideal selection of items for every single research question on personality and personality development. Instead, these items can be seen as an item pool from which to select the

indicators best suited to address the issues of model fit and measurement invariance for subsequent comparison of normative age differences. In this dissertation, I will present two metaheuristic item sampling procedures – Ant Colony Optimization (ACO; Leite, Huang, & Marcoulides, 2008; Olaru et al., 2015; Schroeders, Wilhelm, & Olaru, 2016) and a Genetic Algorithm (GA; Eisenbarth, Lilienfeld, & Yarkoni, 2015; Schroeders et al., 2016; Yarkoni, 2010) – that can be used to select items that optimize a wide range of prespecified psychometric criteria, thus being able to eliminate undesired item and item x age interaction effects. Alternatively, items can be sampled to maximize item x age interaction effects to maximize the sensitivity of the personality measurement within restricted age ranges (see second perspective of the DIF paradox; Church et al., 2011)

Issues can also be found on the person sampling side of personality development studies. Typically, personality differences are studied across age groups, despite the continuous nature of age. All cross-sectional studies presented in Table 2 (except for; Soto, John, Gosling, & Potter, 2011; Srivastava, John, Gosling, & Potter, 2003) examined personality differences across artificially categorized age groups (arguably, even using single years of age also represents some form of categorization, as persons differing only one day in age may be assigned to different age year groups). In the case of studies applying MGCFA, this often results in a low number of very broad age groups, which ensures that sample size requirements are met. This artificial categorization of age will inevitably influence the findings (Hildebrandt, Lüdtke, Robitzsch, Sommer, & Wilhelm, 2016; MacCallum, Zhang, Preacher, & Rucker, 2002), as a low number of groups makes it difficult to examine non-linear developmental trends and find potential onsets of change (Hildebrandt et al., 2016). In addition, broad age groups result in a loss of information within group differences (MacCallum et al., 2002). As such, the generalizability of the findings to the more abstract level of personality development across age is questionable. To address this issue, I will use Local Structural Equation Modeling (LSEM; Hildebrandt et al., 2016, 2009) to weight

participants by their age instead of allocating them to separate age groups. This allows for the examination of personality differences across a continuous age moderator. By also including participants from neighboring age points with reduced weights, LSEM also reduces the effect of potential person sampling artifacts (e.g., higher cognitive ability in younger age groups due to an oversampling of participants), making LSEM particularly suited for an unbiased examination of single moderator variables.

Overview of the Dissertation Manuscripts

In summary, the methods applied to study personality development, particularly in cross-sectional settings, are inappropriate to do so. In this dissertation, I will present novel item and person sampling procedures as an alternative to the currently predominant – but flawed – approaches. More specifically, I will demonstrate how the metaheuristic item selection procedures ACO (Leite et al., 2008; Olaru et al., 2015; Schroeders et al., 2016) and GA (Eisenbarth et al., 2015; Schroeders et al., 2016; Yarkoni, 2010) can be used to improve a wide variety of psychometric properties, such as model fit, reliability and measurement invariance. These procedures can optimize model misfit by eliminating problematic items instead of modifying the model to incorporate model misfit. Based on these cleaned models (i.e., with adequate model fit and measurement invariance), I will examine normative differences in personality across age. I will also expand on the examination of structural change, which is typically done using measurement invariance testing, by using item selection algorithms to identify the most representative personality items for specific age points, allowing for a more profound examination of structural differences at the item level. And finally, I will show how the person sampling procedure LSEM (Hildebrandt et al., 2016, 2009) can be used to study personality development across a continuous age variable instead of categorical age groups. I will also combine both procedures to maximize the interaction between items and persons sampled, thus creating measures that maximize the validity of the assessment in specific age ranges. In the following, I will outline the research questions and

methods applied in each manuscript of this dissertation and compare novel with traditional procedures in personality development research.

Manuscript 1: A Tutorial on Novel Item and Person Sampling Procedures for Personality Research.

The first article in this dissertation presents a tutorial on the item sampling procedure *Ant Colony Optimization* and person sampling procedures *Local Structural Equation Modeling*. In this article, I show how ACO can be used to select short-scale items that optimize user-defined psychometric properties (e.g., model fit) beyond the full scale. I also illustrate how LSEM can be used to study age-associated differences (absolute, structural and divergent) across a continuous age variable (in contrast to age groups in MGCFA). By combining both methods, a wide variety of personality development research questions can be examined in a meaningful manner. This manuscript provides suggestions on how items and persons can be sampled to investigate both perspectives on the DIF-paradox: A) how to identify measurement invariant items to compare mean values across age and B) how to create age specific measurements for higher precision and representativeness within restricted age ranges.

Manuscript 2: A Confirmatory Examination of Age-Associated Personality Differences: Deriving Age-Related Measurement Invariant Solutions using Ant Colony Optimization.

The second manuscript examines questions on normative and structural personality factor differences across age. Normative change is typically examined by comparing scale or factor scores across age or age groups respectively. Structural differences are usually studied by testing measurement invariance of the model parameters. Many studies that focus on normative change do not account for possible structural differences, and studies that do so, only achieve partial invariance or use methodological tweaks to artificially increase model fit. In this study, I want to show how the item selection algorithm Ant Colony Optimization can

be used to derive unidimensional and measurement invariant models of personality that can be subsequently used to compare mean-levels across age groups. By modeling personality as a higher-order model with trait domains atop of more specific facet factors, I want to show how structural changes can be examined both at the facet and factor level of personality. In addition, I want to demonstrate the importance of also examining normative differences at the facet level, which may deviate from the overarching factor level.

Manuscript 3: "Grandpa, do you like roller coasters?": Identifying Age-Appropriate Personality Indicators.

The third manuscript examines structural differences in the measurement of personality across age. This is typically done by testing for measurement invariance of the model across age, as demonstrated in the second manuscript. However, this is often done with the goal of supporting measurement invariance for a subsequent comparison of factor means across age groups. As a result of this, non-invariance of indicators is often neglected or not considered in greater detail. In this article, I want to show how independent item sampling at different age points can be used to identify structural differences in the measurement of personality across age. More specifically, I combined the item sampling approach Genetic Algorithm and person sampling approach LSEM to identify item x age interaction effects on the validity of personality measurement. As measurement invariance across broad age spans is rarely achieved, it is assumed that these effects are quite substantial. In addition, modern deductively developed personality inventories, such as the NEO-PI-R (Costa & McCrae, 1992; Ostendorf & Angleitner, 2004) apply a wide range of different item types (e.g., behaviors, emotions, attitudes, interests) to provide a somewhat representative measure of the underlying traits. The traits to be measured and used item types can be somewhat confounded (e.g., Neuroticism is measured using a large number of emotion-type items). Thus, the effect on item-types used on potential age effects on the measurement of personality are also examined in this manuscript.

In the following, I will present all three manuscripts and summarize the major findings in the epilogue. I will also link them to existing research in personality development and provide suggestions for further research on this topic.

References

Allemand, M., Zimprich, D., & Hendriks, A. A. J. (2008). Age differences in five personality domains across the life span. *Developmental Psychology*, *44*, 758–770. DOI: 10.1037/0012-1649.44.3.758

Allemand, M., Zimprich, D., & Hertzog, C. (2007). Cross-sectional age differences and longitudinal age changes of personality in middle adulthood and old age. *Journal of Personality*, *75*, 323–358. DOI: 10.1111/j.1467-6494.2006.00441.x

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*.

Angleitner, A., John, O. P., & Löhr, F.-J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. In *Personality Assessment via Questionnaires* (pp. 61–108). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-70751-3_5

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 397–438. DOI: 10.1080/10705510903008204

Bleidorn, W., Hopwood, C. J., & Lucas, R. E. (2018). Life events and personality trait change: Life events and trait change. *Journal of Personality*, *86*, 83–96. DOI: 10.1111/jopy.12286

Bleidorn, W., Kandler, C., Riemann, R., Angleitner, A., & Spinath, F. M. (2009). Patterns and sources of adult personality development: Growth curve analyses of the NEO PI-R scales in a longitudinal twin study. *Journal of Personality and Social Psychology*, *97*, 142. DOI: 10.1037/a0015434

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences*, *11*, 515–524. DOI: 10.1016/0191-8869(90)90065-Y

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*, 176–181. DOI: 10.1097/01.mlr.0000245143.08679.cc

Brandt, N. D., Becker, M., Tetzner, J., Brunner, M., Kuhl, P., & Maaz, K. (2018). Personality across the lifespan. *European Journal of Psychological Assessment*, 1–12. DOI: 10.1027/1015-5759/a000490

Caspi, A., & Roberts, B. W. (2001). Personality development across the life course: The argument for change and continuity. *Psychological Inquiry*, *12*, 49–66. DOI: 10.1207/S15327965PLI1202_01

Charles, S. T., & Carstensen, L. L. (2010). Social and emotional aging. *Annual Review of Psychology*, *61*, 383–409. DOI: 10.1146/annurev.psych.093008.100448

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1–27. DOI: 10.1016/S0149-2063(99)80001-4

Costa, P. T., & McCrae, R. R. (1992). *Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo-ffi)*. Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the revised NEO personality inventory. *Journal of Personality Assessment*, *64*, 21–50. DOI: 10.1207/s15327752jpa6401_2

Donnellan, M. B., Conger, R. D., & Burzette, R. G. (2007). Personality development from late adolescence to young adulthood: Differential stability, normative maturity, and evidence for the maturity-stability hypothesis. *Journal of Personality*, *75*, 237–264. DOI: 10.1111/j.1467-6494.2007.00438.x

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP Scales: Tiny-yet-effective measures of the Big Five Factors of personality. *Psychological Assessment*, *18*, 192–203. DOI: 10.1037/1040-3590.18.2.192

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory–Revised (PPI-R). *Psychological Assessment*, *27*, 194–202. DOI: 10.1037/pas0000032

Gerlitz, J.-Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *DIW Research Notes*, *4*, 2005.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*, 14. DOI: 10.1037//0022-3514.59.6.1216

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

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504–528. DOI: 10.1016/S0092-6566(03)00046-1

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*. DOI: 10.3389/fpsyg.2014.00980

Helson, R., Jones, C., & Kwan, V. S. (2002). Personality change over 40 years of adulthood: Hierarchical linear modeling analyses of two longitudinal samples. *Journal of Personality and Social Psychology*, *83*, 752. DOI: 10.1037/0022-3514.83.3.752

Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., & Wilhelm, O. (2016). Exploring factor model parameters across continuous variables with Local Structural Equation Models. *Multivariate Behavioral Research*, *51*, 257–258. DOI: 10.1080/00273171.2016.1142856

Hildebrandt, A., Wilhelm, O., & Robitzsch, A. (2009). Complementary and competing factor analytic approaches for the investigation of measurement invariance. *Review of Psychology*, *16*, 87–102.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying Cultural Differences in Items and Traits: Differential Item Functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *28*, 192–218. DOI: 10.1177/0022022197282004

John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, *2*, 171–203. DOI: 10.1002/per.2410020302

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*, 411–431. DOI: 10.1080/00273170802285743

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 151–173. DOI: 10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*, 285–300. DOI: 10.1037/a0033266

Löckenhoff, C. E., Terracciano, A., Patriciu, N. S., Eaton, W. W., & Costa Jr, P. T. (2009). Self-reported extremely adverse life events and longitudinal changes in five-factor model personality traits in an urban sample. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies*, *22*, 53–59. DOI: 10.1002/jts.20385

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694. DOI: 0.2466/PR0.3.7.635-694

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40. DOI: 10.1037//1082-989X.7.1.19

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471–491. DOI: 10.1037/a0019227

McCrae, R. R., Costa, P. T., de Lima, M. P., Simões, A., Ostendorf, F., Angleitner, A., … Barbaranelli, C. (1999). Age differences in personality across the adult life span: parallels in five cultures. *Developmental Psychology*, *35*, 466. DOI: 10.1037//0012-1649.35.2.466

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. DOI: 10.1007/BF02294825

Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. In *Structural equation modeling: A second course, 2nd ed.* (pp. 395–436). Charlotte, NC, US: IAP Information Age Publishing.

Norman, W. T. (1967). *2800 personality trait descriptors - normative operating characteristics for a university population*. Ann Arbor, MI: Department of Psychology, University of Michigan.

Nye, C. D., Allemand, M., Gosling, S. D., Potter, J., & Roberts, B. W. (2016). Personality trait differences between young and middle-aged adults: Measurement artifacts or actual trends? *Journal of Personality*, *84*, 473–492. DOI: 10.1111/jopy.12173

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality*, *59*, 56–68. DOI: 10.1016/j.jrp.2015.09.001

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae: NEO-PI-R ; Manual*. Retrieved from https://pub.uni-bielefeld.de/publication/1878577

Rauthmann, J. F., Sherman, R. A., & Funder, D. C. (2015). Principles of situation research: Towards a better understanding of psychological situations: Principles of situation research. *European Journal of Personality*, *29*, 363–381. DOI: 10.1002/per.1994

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*, 3–25. DOI: 10.1037//0033-2909.126.1.3

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin*, *143*, 117. DOI: 10.1037/bul0000088

Roberts, B. W., & Mroczek, D. (2008). Personality trait change in adulthood. *Current Directions in Psychological Science*, *17*, 31–35. DOI: 10.1111/j.1467-8721.2008.00543.x

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1–25. DOI: 10.1037/0033-2909.132.1.1

Robins, R. W., Fraley, R. C., Roberts, B. W., & Trzesniewski, K. H. (2001). A longitudinal study of personality change in young adulthood. *Journal of Personality*, *69*, 617–640.

Schroeders, U., & Gnambs, T. (2018). Degrees of freedom in multigroup confirmatory factor analyses: Are models of measurement invariance testing correctly specified? *European Journal of Psychological Assessment*, 1–9. DOI: 10.1027/1015-5759/a000500

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant Colony Optimization and Genetic Algorithm. *PLOS ONE*, *11*, e0167110. DOI: 10.1371/journal.pone.0167110

Small, B. J., Hertzog, C., Hultsch, D. F., & Dixon, R. A. (2003). Stability and change in adult personality over 6 years: Findings from the Victoria Longitudinal Study. *The Journals of Gerontology: Series B*, *58*, P166–P176. DOI: 10.1093/geronb/58.3.P166

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*, 84–90. DOI: 10.1016/j.jrp.2008.10.002

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*, 117–143. DOI: 10.1037/pspp0000096

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, *100*, 330–348. DOI: 10.1037/a0021717

Specht, J., Egloff, B., & Schmukle, S. C. (2011). Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of Personality and Social Psychology*, *101*, 862. DOI: 10.1037/a0024950

Specht, J., Luhmann, M., & Geiser, C. (2014). On the consistency of personality types across adulthood: Latent profile analyses in two large-scale panel studies. *Journal of Personality and Social Psychology*, *107*, 540. DOI: 10.1037/a0036863

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*, 1041. DOI: 10.1037/0022-3514.84.5.1041

Tackett, J. L., Slobodskaya, H. R., Mar, R. A., Deal, J., Halverson, C. F., Baker, S. R., … Besevegis, E. (2012). The hierarchical structure of childhood personality in five countries: Continuity from early childhood to early adolescence: Child personality

structure. *Journal of Personality*, *80*, 847–879. DOI: 10.1111/j.1467-6494.2011.00748.x

Terracciano, A., McCrae, R. R., Brant, L. J., & Costa, P. T. (2005). Hierarchical linear modeling analyses of NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging*, *20*, 493–506. DOI: 10.1037/0882-7974.20.3.493

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. DOI: 10.1177/109442810031002

Vassend, O., & Skrondal, A. (1997). Validation of the NEO Personality Inventory and the five-factor model. Can findings from exploratory and confirmatory factor analysis be reconciled? *European Journal of Personality*, *11*, 147–166. DOI: 10.1002/(SICI)1099-0984(199706)11:2<147::AID-PER278>3.0.CO;2-E

Wrzus, C., & Roberts, B. W. (2017). Processes of personality development in adulthood: The TESSERA framework. *Personality and Social Psychology Review*, *21*, 253–277. DOI: 10.1177/1088868316652279

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*, 180–198. DOI: 10.1016/j.jrp.2010.01.002

II

A Tutorial on Novel Item and Person Sampling Procedures for Personality Research

Gabriel Olaru[1], Ulrich Schroeders[1], Johanna Hartung[2], & Oliver Wilhelm[2]

1: University of Kassel

2: Ulm University

Status – accepted

Olaru, G., Schroeders, U., Hartung, J., & Wilhelm, O. (2019). A Tutorial on Novel Item and Person Sampling Procedures for Personality Research. *European Journal of Personality, 33,* 400-419. DOI: 10.1002/per.2195

Abstract

Measurement in personality development faces many psychometric problems. First, theory-based measurement models do not fit the empirical data in terms of traditional confirmatory factor analysis. Second, measurement invariance across age, which is necessary for a meaningful interpretation of age-associated personality differences, is rarely accomplished. Finally, continuous moderator variables, such as age, are often artificially categorized. This categorization leads to bias when interpreting differences in personality across age. In this tutorial, we introduce methods to remedy these problems. We illustrate how *Ant Colony Optimization* can be used to sample indicators that meet prespecified demands such as model fit. Further, we use *Local Structural Equation Modeling* to resample and weight subjects to study differences in the measurement model across age as a continuous moderator variable. We also provide a detailed illustration for both tools with the Neuroticism scale of the openly available IPIP NEO inventory using data from the United Kingdom Sample (N = 15,827). Combined, both tools can remedy persistent problems in research on personality and its development. In addition to a step-by-step illustration, we provide commented syntax for both tools.

Keywords: Ant Colony Optimization, Local Structural Equation Modeling, item sampling, person sampling, personality development

The majority of findings in personality development research are based on the comparison of personality scale scores across age. Such an approach assumes that all items in the scale are valid representations of the underlying latent personality factors (Borsboom, 2006a, 2008), which is a prerequisite to build (manifest) scores that exhaust the information collected with the associated items. Fortunately, this assumption can be tested by fitting theory-driven models to empirical data using confirmatory factor analysis (CFA). Unfortunately, broad and general models of personality usually do not pass strict model testing (Borkenau & Ostendorf, 1990; Costa & McCrae, 1995; Olaru, Schroeders, Wilhelm, & Ostendorf, 2018; Olaru, Witthöft, & Wilhelm, 2015; Vassend & Skrondal, 1997).

Two salient and prevalent reasons behind this failure to pass strict model tests include the high number of cross-loadings and residual correlations of broad self-report scales, as well as a large number of manifest indicators when modeling broad personality domains at the item level (Moshagen, 2012). Personality researchers are thus commonly faced with four options: a) reject the model when considering all items of a scale; b) reject latent factor modeling and instead use manifest scale scores; c) change the measurement model within the CFA context (e.g., freeing parameters, parceling); or d) apply less restrictive methods, such as *Exploratory Structural Equation Modeling* (ESEM; Asparouhov & Muthén, 2009; Morin, Marsh, & Nagengast, 2013). Simply dumping measures with poor model fit will hardly be deemed an acceptable option because it would affect the vast majority of the instruments currently used in personality psychology. Representing personality as manifest scale scores is also hardly reconcilable with the overarching notion of personality as latent traits. ESEM and the parceling technique (Little, Cunningham, Shahar, & Widaman, 2002; Little, Rhemtulla, Gibson, & Schoemann, 2013) are often applied to minimize misfit due to cross-loadings and residual correlations respectively, both of which are pervasive issues in personality questionnaires. The downside of both procedures is that they conceal model misfit rather than resolve it.

Ensuring the fit of a confirmatory model that is theoretically meaningful (i.e. in line with the interpretation of latent variables we apply) is essential and indispensable when speaking about overarching and highly general personality traits. This standard notion of latent traits is usually assumed to hold across a range of relevant moderators – the most important probably being age. In other words, after establishing a measurement model, it is important to ensure that our interpretation of the trait is invariant across age and similar variables. Only then can researchers draw conclusions about age-associated differences in personality traits.

To illustrate such issues of invariance, consider how items such as, "I keep my workplace tidy" might bias the comparison of personality scores between respondents being in the workforce versus those who are already retired. Items such as, "I like to go to the opera" might yield higher scores for participants of a certain cohort or provenance, despite them having equal levels of the overarching trait Openness. Evidently, a wide range of moderators can affect the measurement of personality: differences in cognitive abilities (Gnambs & Schroeders, 2017), situational transitions across life stages (Bleidorn, Hopwood, & Lucas, 2018; Wrzus & Roberts, 2017; Wrzus, Wagner, & Riediger, 2016), or systematic differences in the social network (Wrzus, Hänel, Wagner, & Neyer, 2013).

The concept of measurement equivalence across a moderator (e.g., gender, age) is referred to as measurement invariance and can be examined with different statistical methods (Mellenbergh, 1989; Meredith, 1993; Millsap, 2012). Cross-sectional personality development studies that test for measurement invariance across age (Allemand, Zimprich, & Hendriks, 2008; Allemand, Zimprich, & Hertzog, 2007; Brandt, Becker, Tetzner, Brunner, Kuhl, & Maaz, 2018; Nye, Allemand, Gosling, Potter, & Roberts, 2016; Olaru et al., 2018) usually examine measurement invariance across age groups by artificially categorizing age in an arbitrary number of groups after separating them based on equally arbitrary thresholds, even though age is continuous in nature. This approach and the associated decisions

concerning number of groups, for example, will inevitably influence the results and can therefore provide us with a distorted picture of personality development (Hildebrandt, Lüdtke, Robitzsch, Sommer, & Wilhelm, 2016; Hildebrandt, Wilhelm, & Robitzsch, 2009; MacCallum, Zhang, Preacher, & Rucker, 2002).

In this tutorial, we present two recently developed methods: The *Ant Colony Optimization* item sampling procedure (ACO; Janssen, Schultze, & Grötsch, 2015; Leite, Huang, & Marcoulides, 2008; Olaru et al., 2018, 2015; Schroeders, Wilhelm, & Olaru, 2016b; Schroeders et al., 2016b) and the *Local Structural Equation Modeling* person sampling procedure (LSEM; Hildebrandt et al., 2016, 2009). On a more general stance, both item- and person-sampling procedures can also be seen as approaches to improve or study the construct validity of a scale (Cook, Campbell, & Shadish, 2002). For instance, the lack of model fit for many personality scales shows that the used personality scores are not unidimensional measures of the personality factors. ACO can be used to identify sets of items that fit the model and thus improve construct validity. LSEM as a person-sampling method can be used to examine differences in the model across observations. In the case of personality development, this refers to the question whether the structure of the personality models is affected by age. While these methods may seem to be very different – ACO is used to improve the model, whereas LSEM is used to identify differences in the model across persons – we argue that both examine under which item-person combinations our theoretical model of personality holds. Combined, these two tools can be used in personality development research to identify items that work across broad age spans (Olaru et al., 2018), or only do so for specific ages (Olaru, Schroeders, Wilhelm, & Ostendorf, 2019), thus indicating variations in personality-related behaviors across age that transcend simple mean differences. Applications of both tools are, of course, not limited to questionnaire data, but can be used to derive short-scales and examine structural differences on test data, behavioral ratings, etc. (Briley, Harden,

Bates, & Tucker-Drob, 2015; Eisenbarth, Lilienfeld, & Yarkoni, 2015; Hildebrandt et al., 2016, 2009; Janssen et al., 2015; Schroeders et al., 2016b).

We explain and illustrate the application of both tools. For both methods, we first describe the psychometric problem in more detail, followed by an application of both methods respectively in order to understand or alleviate psychometric issues. In form of a step-by-step guide, we show how these methods can be applied to your research question and data using R (R Core Team, 2018). To foster the utility of *Ant Colony Optimization* as an item-sampling method and *Local Structural Equation Modeling* as a person-sampling method, we provide the commented R-scripts used in this tutorial in an online repository on OSF (Nosek et al., 2015): https://osf.io/yx4km/.

In this tutorial, we applied both methods on the Neuroticism scale with the underlying facets (*Anxiety*, *Anger*, *Depression*, *Self-Consciousness*, *Immoderation*, and *Vulnerability*) of the IPIP NEO 300 personality inventory (Johnson, 2014). The analysis was based on the UK sample (N = 16,489) of the openly available IPIP NEO 300 data (https://osf.io/tbmh5/; Johnson, 2014). We removed test-takers with an age below 15 (n = 661) and an age above 75 (n = 1). The remaining 15,827 participants (8,545 female or 54%) had an average age of 25.46 years (SD = 9.87).

Item Sampling

Personality data are a product of the persons assessed and the items used (Brennan, 1992). While person sampling is often considered, for instance by matching experimental groups based on covariates (e.g., propensity score matching; Dehejia & Wahba, 2002), or by allocating regression weights to respondents to account for non-representative samples (Biemer & Christ, 2008; DuMouchel & Duncan, 1983), the sampling of items from measures is often a black box.

Ideally, the development of personality scales begins by delineating the domains to be measured and creating a broad item pool that encapsulates all relevant content related to the

trait to be measured (Buss & Craik, 1983; Loevinger, 1957). From this representation of the item universe of personality items, the most relevant indicators for the desired population can then be selected. Personality inventories often apply item-sampling procedures aiming for high internal consistency of scales and an underlying simple structure of principal components (Costa & McCrae, 1995; Donnellan, Oswald, Baird, & Lucas, 2006; Kruyen, Emons, & Sijtsma, 2012; Saucier, 1994; Soto & John, 2009, 2017). However, model fit of broad personality inventories in terms of confirmatory standards (Hu & Bentler, 1999) is usually poor (Borkenau & Ostendorf, 1990; Costa & McCrae, 1995; Olaru et al., 2018, 2015; Vassend & Skrondal, 1997). Therefore, interpreting such scales as a gold standard can be problematic, as such scales only represent one potential item sample from a hypothetical personality item universe (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), which is influenced by the item selection procedure (e.g., increase internal consistency) and underlying sample (e.g., student sample). Apart from issues related to the transition from the ever-prevalent data-reduction method Principal Components Analysis towards the latent modeling procedure CFA (Borsboom, 2006a, 2008), selecting items solely based on main loadings does not necessarily improve model fit (Olaru et al., 2015).

With respect to the goal of identifying psychometrically-sound item sets from a larger pool, three aspects deserve attention. First, item-level parameters (e.g., factor loadings, modification indices) can only vaguely serve as proxies for scale-level characteristics (e.g., reliability, model fit; see Mellenbergh, 1996). As such, item selection based on item-level characteristics will often be inferior to scale-level information selection (Olaru et al., 2015). Accordingly, an item selection procedure that evaluates scale-level instead of item-level information is desirable.

Second, model parameters will change when removing items. For instance, an item can have a high factor loading as long as it is included in the scale with similar items, but low after similar items have been removed. Sequentially removing items will ignore such effects

of items similarities and can lead to local optima with inferior solutions following the optimization process (Olaru et al., 2015; Schroeders et al., 2016b). To solve this issue, it is necessary to select items in a combinatorial rather than in a stepwise fashion. This strategy increases the length of the item selection process considerably. For instance, a stepwise reduction of a scale from 60 to 30 items only requires 31 model estimations, whereas comparing all potential models with 30 items results in 118,264,581,564,861,424 possible combinations. To reduce the computational load of the analysis, we need meta-heuristic procedures that allow us to search for promising item sets. Such procedures have been developed and used in computer science to solve similar combinatorial problems (e.g., Dorigo & Stützle, 2010).

Third, when selecting items based on more than one criterion—for instance, model fit and reliability—these criteria have to be considered simultaneously and must be balanced in a meaningful matter. For instance, removing items based on factor loadings first and model fit later will negatively affect the quality of the final solution due to the aforementioned sequence effects. We hence need an item selection procedure that takes into account several criteria simultaneously and weighs them based on the requirement to optimally answer the research question.

In our example, we used a broad personality inventory from which we selected an item set that adheres to a predefined standard such as minimally required model fit. Of course, the methods we present may also be applied to derive short scales that optimize any desired criteria (e.g., correlation with long form, predictive validity of the scale), depending on the research question. The derived item sets can be used to study a wide range of research questions of personality structure and its development.

Ant Colony Optimization

ACO (Leite et al., 2008; Marcoulides & Drezner, 2003) is an optimization procedure capable of tackling the aforementioned issues by finding an optimal (or near-optimal) solution

using a search heuristic inspired by the foraging behavior of ants (Deneubourg, Aron, Goss, & Pasteels, 1990). ACOs have been successfully applied in a number of studies to derive efficient short scales (Janssen et al., 2015; Leite et al., 2008; Olaru et al., 2018, 2015; Schroeders et al., 2016b; Schroeders, Wilhelm, & Olaru, 2016a).

Similar to the way in which ants use pheromones to attract other ants to the shortest route to a food source, ACO uses virtual pheromones to increase the attractiveness of item sets that yield better psychometric properties (e.g., model fit). Initially, ants will randomly explore the space between their nest and the food source. During their search, each ant leaves a pheromone trail. On shorter routes, more pheromones accumulate in a given time frame. Higher levels of pheromones attract more ants, and this in turn further increases pheromone levels until many or most ants will follow the shortest route.

In the context of scale construction, ants select item sets instead of routes. Each item has a corresponding pheromone level that determines the probability of the item of being selected by an ant. Through several iterations item sets are evaluated and pheromone levels are adjusted based on the quality of the solutions. Figure 1 provides a simplified illustration on how six items are selected and evaluated across 2 iterations based on the Comparative Fit Index (CFI). In Figure 1, three models (= ants) are estimated on each iteration. Items for these models are drawn based on the pheromone levels for each item. Each time a new best model is found (e.g., Model A in iteration 1; Model B in iteration 2), the pheromone levels for the corresponding items increase. This change in turn increases the items selection probability in subsequent iterations.

Iter.	Operation	Ant	Items 1	2	3	4	5	6	Criterion (CFI)	Pheromones					
1	Select and evaluate items based on pheromones	A	■		■		■		A: .91	1	1	1	1	1	1
		B		■	■			■	B: .85	1	1	1	1	1	1
		C	■			■		■	C: .80	1	1	1	1	1	1
1	Find best model and increase pheromones for corresponding items	A	■		■		■		A: .91	2	1	2	1	2	1
										2	1	2	1	2	1
										2	1	2	1	2	1
2	Select and evaluate items based on pheromones	A	■		■	■			A: .89	2	1	2	1	2	1
		B		■	■		■		B: .93	2	1	2	1	2	1
		C			■		■	■	C: .86	2	1	2	1	2	1
2	Find best model and increase pheromones for corresponding items	B		■	■		■		B: .93	2	2	3	1	3	1
										2	2	3	1	3	1
										2	2	3	1	3	1

Figure 1. Ant Colony Optimization illustration. Iter. = Iteration; Black = selected item; White = unselected item; Darker shades of grey indicate higher pheromone levels; Pheromones determine the probability of item selection. As such, item 1, 3 and 5 will be selected twice as likely as item 2, 4, and 6 in the second iteration. Over the course of several iterations, better performing items will be selected more frequently than worse performing items. The procedure is repeated until no improvement of the model can be found across several iterations.

Figure 2 shows how pheromones in the current analysis increased for the most promising items across several iterations. Pheromone levels can never reach a value of zero; thus, items with low pheromone levels can still be selected in later iterations. This feature ensures that the search process does not get stuck in local optima. Because ACO is a probabilistic approach, it will not necessarily find the optimal solution. In order to approach optimal solutions, results across several runs of the algorithm should be compared (Dorigo & Stützle, 2010; Leite et al., 2008). In the following, we discuss how ACO can be adjusted to best approach the underlying research question with the available data.

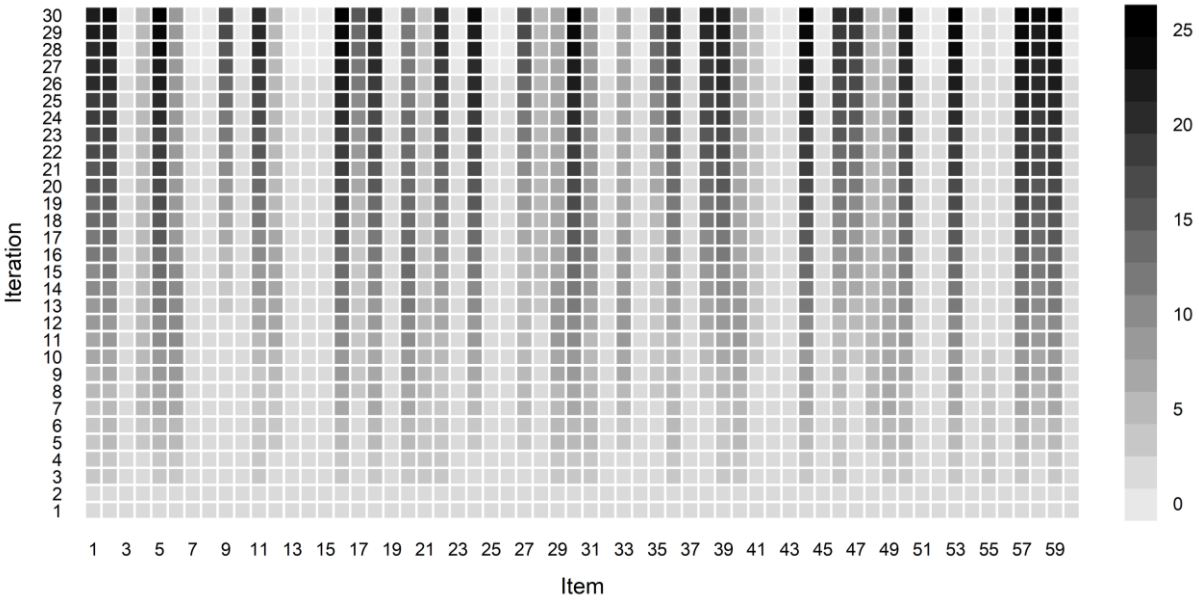


Figure 2. Pheromones across iterations. Darker shades represent higher pheromone levels. Depicted are pheromone values across the first 30 iterations in the current analysis.

Item Sampling with ACO

Before using ACO, the starting position of the search as a point of reference needs to be evaluated (i.e., running the full model). After this initial step, the core of ACO (i.e. the optimization function that is used to evaluate and select models) has to be defined. Finally, based on the research goals and the number of possible short models, further ACO parameters can be adjusted. These settings (i.e., number of ants, stopping criterion, and function of evaporation) affect the convergence of ACO. In the following sections, we give a detailed overview of each step and will provide some practical recommendations. At the end of each section, we present and discuss our decisions made in the current application based on the Neuroticism factor of the IPIP NEO 300.

What model do I want to optimize?

Item selection procedures are meant to identify one or more item sets that yield the best solution according to the optimization criterion given the *a priori*-set model. Obviously, model misspecifications can have severe adverse impact on the convergence behavior of the item selection process. ACO will evaluate short-scales based on the initially-defined model structure. In the case of personality research, discarding the facet level and trying to optimize

a one-factor model is an example of misspecification. If the optimization criterion includes reliability, the selection procedure will then gravitate to select highly-correlated items from a single facet, arguably increasing reliability at the expense of validity (Schroeders et al., 2016b). Before selecting items, it is thus important to make sure that the *a priori* measurement model is in line with theoretical assumptions. If no theoretical assumptions are available, exploratory factor analytical methods could be implemented to derive an initial model that is then subsequently optimized. We also want to stress the importance of choosing an adequate estimator in model estimation to avoid bias: personality items typically use categorical response scales and estimators for categorical data (e.g., *Weighted Least Squares*) should thus be used when estimating personality models in CFA. We want to point out that *Maximum Likelihood* estimation is also appropriate (and more efficient) for most normally distributed Big Five or FFM measures with at least five categories per item (Beauducel & Herzberg, 2006; Rhemtulla, Brosseau-Liard, & Savalei, 2012). However, in the case of more extreme (i.e., uncommon) personality traits, such as the Dark Triad of personality (Paulhus & Williams, 2002) or the DSM-5 maladaptive personality model (Griffin & Samuel, 2014; Krueger, Derringer, Markon, Watson, & Skodol, 2012), using estimators for continuous data (e.g., *Maximum Likelihood*) may severely impact model fit due to the skewness of the item distributions.

In the current application, we follow common practice in analyzing these personality factors and specified a correlated factor model with six factors[1] - each factor representing a facet of Neuroticism (Costa & McCrae, 1995). We want to demonstrate how several psychometric criteria can be optimized simultaneously: model fit and reliability are optimized

[1] Note that although a higher order factor model with Neuroticism as a second order factor is a more accurate representation of the theoretical structure of the personality factor (Olaru, Schroeders, Wilhelm, & Ostendorf, 2018), we decided to use a correlated factor model, as this type of model is a more prevalent representation of many psychological constructs. The script includes options to change the model to a higher order, bi-factor, or acquiescence model.

in this case, or more specifically factor saturation McDonald's ω (2013). McDonald's ω is an indicator of the amount of item variance explained by the underlying latent factor. In contrast to Cronbach's α (Cronbach et al., 1972), it does not suffer from a lack of tau-equivalence across items, which is often the norm in personality research. Both the relative (Comparative Fit Index; CFI = .789) and absolute (Root Mean Square Error of Approximation; RMSEA = .062) indicators of model fit of the full model were inadequate based on prevalent cut-off criteria (CFI ≥ .95; RMSEA ≤ .06; Hu & Bentler, 1999). Factor saturation of all facets in the full scale was sufficient (ω ≥ .70), ranging from .80 (*Immoderation*) to .92 (*Depression*).

What criteria do I want to optimize?

By evaluating the full model, we can identify problematic psychometric properties of the model that need to be addressed (e.g., model fit). In addition, a wide range of other desirable criteria can be included for optimization, for instance dimensionality of the measure, measurement invariance, reliability, or predictive validity. Any type of quantifiable criteria can be optimized (e.g., item difficulties, balance of positively and negatively coded items, etc.). Keep in mind that some properties will improve with the number of items discarded (e.g., CFI; Moshagen, 2012; Olaru et al., 2015), whereas others may decrease due to the reduced number of items (e.g., reliability; Kruyen et al., 2012). As such, even if reliability is acceptable in the full model, it can drop to a critical level due to the reduced number of items. In the current analysis, we want to optimize model fit (CFI and RMSEA) and also include factor saturation ω to ensure that it is not negatively affected by the item reduction.

Choosing a meaningful set of criteria to optimize is the most critical step in the item selection procedure, as the final optimization function and modeling approach (e.g., one-factor CFA model) both have a strong impact on the item selection process. Focusing on a *single optimization criterion* may result in *overfitting* this criterion, at the cost of other non-optimized but still relevant psychometric properties. For example, only optimizing the absolute model fit neglects questions of reliability, whereas only optimizing factor saturation

will neglect model fit. Scales are often shortened by selecting items with high factor loadings. This procedure will usually increase the internal consistency of the scale but the item selection procedure will result in a set of homogenous items. Therefore, the construct coverage of the reduced item set may be severely limited (see Schroeders, Wilhelm, & Olaru, 2016a). Note that the item sampling procedure can also be restricted to maintain construct coverage, for instance by retaining the facet level of the scale or maintaining the balance between positively and negatively worded items.

By optimizing more than one criterion, the resulting scale can be tailored to meet a mixture of demands but it is also possible to optimize *too many* criteria. It may be tempting to select a large number of optimization criteria, but the improvement will be limited by the number of items from which we select. The chance of meeting all criteria simultaneously is low. We thus recommend to first run ACO with a small set of essential optimization criteria and to study the convergence process carefully (this can be done using the monitor function and output files of the provided R script), before expanding the number of criteria. Which optimization criteria are used in which combination always depends on the research question, scale properties, sample, and researchers' preferences. Therefore, we can only provide general advice that needs adaptation to the application context.

ACO is a data-driven procedure that will optimize the model based on the specified criteria and the data. All else being equal, as sample size decreases, so does the likelihood that the model will fit in other samples. To test for *overfitting*, cross-validating the derived model on an independent sample is recommended. The ACO R script includes a cross-validation function that optimizes the model on a randomly selected subsample of participants and subsequently tests the robustness of model fit and parameter estimates on the remaining sample. Cross-validation and the provided function are discussed in more detail later.

How can I weight each criterion?

ACO will evaluate each item selection based on a single numerical value, which should be an aggregate of all relevant criteria. It is up to the user to find a suitable optimization function that weights and subsequently sums up (or averages) all optimization criteria. The issue faced at this stage is that the potential criteria candidates differ in the range of numerical values and direction (e.g., CFI and RMSEA). In addition, critical cutoff values for the criteria vary substantially (e.g., CFI ≥ .95; RMSEA ≤ .06; ω ≤ .70). Simply adding (or subtracting) these parameters will overemphasize criteria with larger value ranges (e.g., ω over RMSEA). We hence recommend transforming these values first to ensure that all criteria are weighted as intended (i.e. equally). One possible transformation is the logit transformation (see Equation 1 to 3 and Figure 3), which has the benefit of scaling the values to a range between 0 and 1 (Janssen et al., 2015; Schultze, 2017), resulting in comparable criteria. Furthermore, logit transformation will maximize the differentiation around a critical cutoff value (e.g., RMSEA ≤ .06). Due to the shape of the logistic function and a maximum transformed value of one, the optimization of several criteria will be balanced, as over-optimizing a single criterion will not be further rewarded. This result is particularly beneficial in cases with adverse starting criteria (e.g., in the current analysis: near-acceptable RMSEA, but critical CFI because of relatively low factor loadings; see also Moshagen & Auerswald, 2018). Weighting of the different criteria can be easily done by changing the desired cutoff value in the transformation (e.g., further emphasizing CFI by increasing the critical value from CFI ≥ .90 to CFI ≥ .95). The slope of the logistic transformation can be adjusted via the factor in the exponential function (in this case 100; see Equations 1 to 3). If, for instance, CFI values are particularly low compared to other criteria, the slope of the CFI transformation function can be decreased to stronger reward increases in the lower spectrum. In the current study, we transformed each optimization criterion (i.e. CFI, RMSEA and ω) around the critical cutoff values as suggested by the literature (see Equations 1 to 3). Figure 3 illustrates

how CFI and RMSEA were transformed. The final optimization criterion for ACO was the sum of the average of the two transformed fit values and the transformed ω (see Equation 4).

$$\varphi_{CFI} = \frac{1}{1+e^{100*(0.90-CFI)}} \quad (1)$$

$$\varphi_{RMSEA} = 1 - \frac{1}{1+e^{100*(0.06-RMSEA)}} \quad (2)$$

$$\varphi_{\omega} = \frac{1}{1+e^{100*(0.70-\omega)}} \quad (3)$$

$$\varphi_{overall} = \frac{\varphi_{CFI} + \varphi_{RMSEA}}{2} + \varphi_{\omega} \quad (4)$$

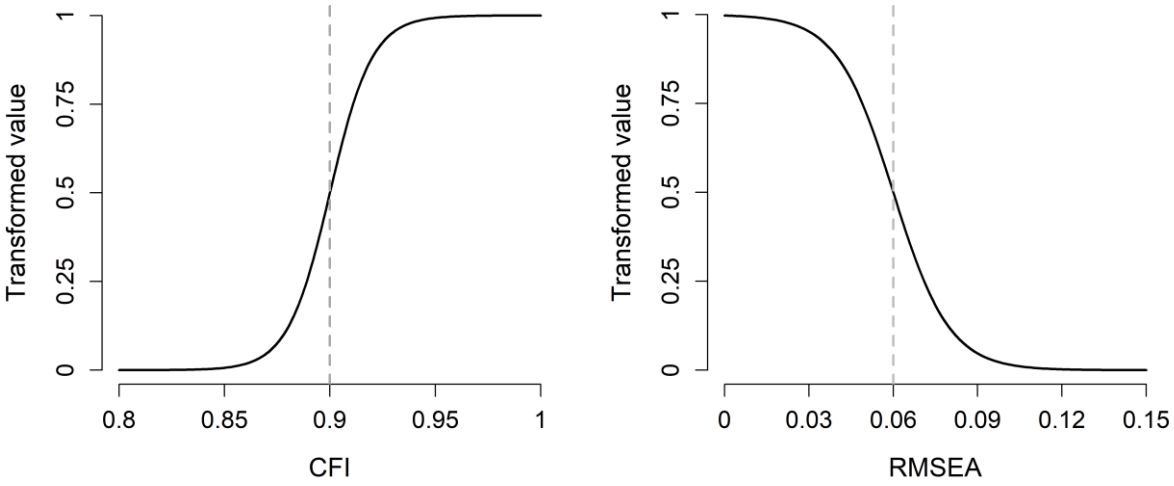


Figure 3a, b. Logistic transformation of CFI and RMSEA for optimization function.

Figure 4 shows how CFI was optimized in the current analysis across several iterations based on the current optimization criterion. Note that even though the item reduction alone resulted in acceptable CFI values in the first iteration (see also, Moshagen, 2012; Olaru et al., 2015), the average CFI of the randomly selected models was only .86. This value increased asymptotically towards to a value of .95 across the presented iterations. As can be seen, the logistic transformation resulted in a strong increase in CFI across the first iterations (around the critical cut-off), but only small improvements as acceptable levels were reached. At this point, ACO will focus on optimization criteria that are closer to the critical cut-off.

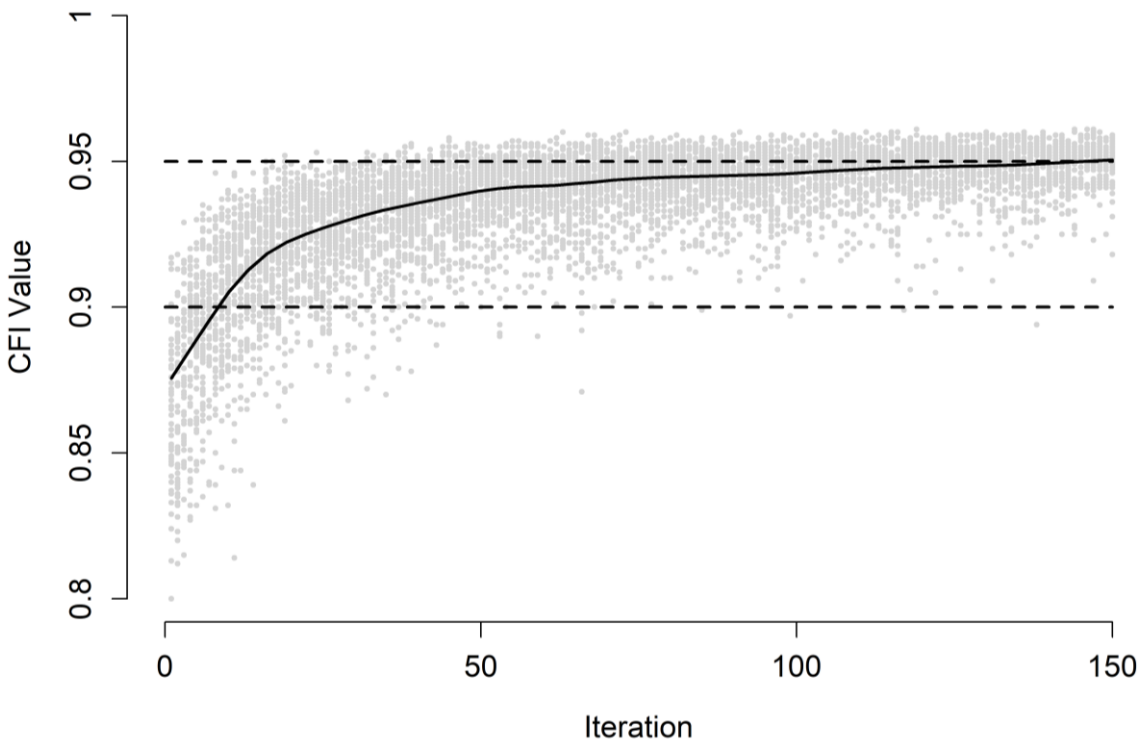


Figure 4. CFI convergence across several iterations. Every grey point represents the CFI value of an estimated 24-item model. The solid black line shows the (smoothed) average CFI across iterations. The dashed black lines represent critical CFI cut-off values.

How many items should I select?

The overall number of available items and the quality of the full model (e.g., model fit, reliability) determine the chance of finding an adequate short scale. The further desired properties are from acceptable levels, the more items might need to be discarded before finding adequate models. On the other hand, reducing the item number will negatively affect reliability and measurement precision at the individual level (Kruyen et al., 2012), as well as predictive validity (Soto & John, 2018). While reliability (i.e., measurement precision at the group level) is related to measurement precision at the individual level, the latter will suffer much more from shortening scales (Emons, Sijtsma, & Meijer, 2007; Kruyen et al., 2012). The number of items to select thus always depends among others on the current research question, considerations of the researcher, and quality of the full model. It is up to the user to find a suitable item number that provides the best compromise between strong optimization

and aspects such as reliability, measurement precision, or construct coverage (see, Schroeders et al., 2016a; Soto & John, 2018).

Should I use ACO at all?

This question might seem a little out of place at this point. However, at this point, one can determine whether the number of possible models with the given item number warrants the use of ACO or whether a brute force algorithm examining all possible models will do just as well (or even better, considering that it will find the optimal model with absolute certainty). This question can be answered by calculating the number of possible models or item combinations per factor with $\binom{n}{k}$ to gauge the computational load and time required. Metaheuristics such as ACO are made to solve problems with a high number of possible combinations (e.g., ten out of 30 items: $\binom{30}{10} = 30{,}045{,}015$ models) that cannot be solved in a reasonable amount of time by doing a full search. For instance, If we want to create a short-scale of the BFI-2 (three facets per factor with four items each; Soto & John, 2017) that retains the facet level, we can select three items per facet, for a total of $\binom{4}{3}^3 = 64$ combinations per factor. In this case, computing every possible model is much faster than the heuristic search procedure applied by ACO. In the current example, we selected four out of ten items per facet (six facets in total) for a total number of 24 items for Neuroticism. The possible number of item combinations per facet was thus 210 with a total of 210^6 = 85,766,121,000,000 combinations for the entire factor. Estimating every model is not computationally sensible, hence ACO is ideal in this case.

What do ants and pheromones have to do with all of this?

Before the search process can begin, a number of ACO parameters need to be set first. In principle, all parameters affect the convergence behavior and length of the search process. Some experimentation might be needed before adequate values can be found that balance search length with quality.

Ants and stopping criterion.

Ants represent the number of models estimated per iteration. A higher number of ants will lead to a more conservative search, as a larger number of models are evaluated before updating pheromone levels on the best solution. Decreasing the number of ants leads to a faster convergence as pheromone levels are updated more frequently at the expense of a higher risk of increasing pheromones on inferior solutions. After a specific number of iterations without improvement in the final optimization criterion, the search process is stopped. Increasing the number of iterations with no improvement will lead to a more conservative search, as more models are tested before the search is ended, but at the cost of a substantially prolonged runtime. In the R script, any amount of improvement in the overall optimization function will reset the iterations counter. If desired, a minimum value of improvement can be specified in the script to reduce the overall number of iterations in the search. However, depending on the transformation of the criteria, even a small improvement in the transformed values may be substantial in the original criterion if these changes are further from the critical cut-off value (in the case of the logistic transformation). The provided R script also contains a redundancy check to estimate each selected model only once and retaining the corresponding optimization criterion value, thus reducing the impact of a high number of iterations on overall computation time. Generally, with an increasing number of possible solutions, the number of ants and stopping criterion (i.e., maximum number of iterations without improvement in the final optimization criterion) should be increased accordingly. There are no guidelines for the stopping criterion and number of ants that should be applied, as these can vary between applications. Desired settings will be influenced by the number of possible models, the computational time or computational power available, as well as the anticipated consequences of a non-perfect solution. In the current analysis, we chose a maximum number of 40 iterations without improvement to the model and 60 ants per iteration. We generally suggest using a sufficiently high number of ants (generally, values

between 10 and 100 are used, see Schultze, 2017) to ensure comparison of a larger number of models prior to updating the pheromones, hence reducing the chance that pheromones are increased for problematic solutions.

Pheromones and Evaporation.

All items start with equal pheromone levels. Pheromone levels determine the selection probability of the corresponding items. The equal pheromone levels at the start of the search process result in a completely random item selection in the first iteration. Pheromone levels at the start of the search process can also be modified based on theoretical considerations or *a priori* assumptions on the item performance to speed up the search process. If some items are essential for the scale (e.g., linking items), the pheromone values of these items can be substantially increased to ensure that these items are selected (alternatively, the model writing syntax can be modified to always include these items). At the end of each iteration, pheromone levels of the items in the best solution found within the iteration are increased by the overall model quality (i.e., the optimization function φ_{overall}). Pheromones will hence increase depending on the quality of the solution. Increased pheromone levels then yield a higher selection probability for the corresponding items in subsequent iterations. Similar to the way natural pheromone levels evaporate over time, overall pheromone values also decrease over the course of the search process. In the current analysis, all pheromone levels are multiplied with an evaporation parameter of .99 after each iteration. Evaporation ensures that pheromone levels of items that received a pheromone boost in early iterations—in which the quality of the solutions is typically low—will be adjusted if these items do not perform well in subsequent iterations. With a stronger evaporation (i.e., with a lower evaporation multiplier), pheromones of rarely-selected items will be punished more strongly and the difference in selection rates will increase much faster than when using pheromone multipliers close to one (e.g., .99). The closer the evaporation parameter is to 1, the more extended and more precise the search procedure will generally be.

In summary, the following parameter settings increase computational time and the chance of finding a better solution: a) an increased number of ants per iteration, b) iterations necessary until the search is stopped, and c) evaporation multiplier. By decreasing these parameters, or speeding up pheromone accumulation, the search process will terminate faster. This acceleration comes at the risk of finding inferior solutions (i.e. there is a tradeoff between computational time and accuracy). Furthermore, the function used to transform the optimization criteria will also affect the search behavior (e.g., a logistic function in the example). Changing this transformation function will also affect the convergence behavior. For example, reducing the slope in the logistic function (or generally, increasing the range in which the criterion is optimized) results in an extended optimization of the criterion, at the risk of *over-optimizing* criteria with broader ranges (e.g., reliability compared to RMSEA). In contrast, a steeper slope results in a faster search, as the maximum value of the optimization function can be reached faster, at the risk of *capitalizing on chance* (i.e., *under*-optimization). If no (or a linear) transformation of the criteria is used, this problem is negligible but there is a risk that ACO *over-optimizes* criteria with broader ranges.

How can I ensure that my results are robust and replicable?

ACO is a probabilistic search procedure that will not necessarily find the best solution. It is hence recommended to run ACO several times with identical settings before accepting a final solution. If the goal is to find the single best solution possible, the search should only be ended if the solution has been replicated across several runs of ACO with identical settings. If the quality of solutions found across several runs of ACO varies strongly, the number of ants and iterations should be increased to reduce the impact of chance on the item selection. In the current example, we ran ACO ten times and selected the best solution across these runs. To ensure that the findings are replicable, it is possible to set the random seed in R to a specific value using `set.seed` before starting ACO.

Similar to other optimization or machine learning algorithms, the robustness of the item selection should be cross-validated to avoid overfitting (Yarkoni & Westfall, 2017). We recommend an approach commonly used in the machine learning context, in which the sample is initially split into a training and validation subsample. In a first step, ACO is run on the training sample and, in a second step, the recommended solution is evaluated in the so-called validation sample. By estimating the model on the independent validation sample, the robustness of model fit and other scale level criteria can be evaluated. Parameter equivalence (e.g., factor loadings) of the model can additionally be tested by specifying a MGCFA across the two subsamples (see also Schultze, 2018). In the machine learning context, around 80% of the total sample are typically allocated to the training sample (Yarkoni & Westfall, 2017), as the prediction algorithm requires a large number of cases to derive robust parameters (i.e., regression weights). The predictive validity of these parameters is subsequently evaluated on an individual person level. For this reason, only a relatively small validation sample is required. However, in the present case of item selection with CFA, we recommend allocating 50% of the sample to both subsamples (see also Schultze, 2018), to ensure that both samples are sufficiently large for model estimation. An imbalance between the sample sizes of the two samples (i.e., a too small validation sample) will also bias the resulting parameter estimates in a MGCFA context towards the larger sample (Yoon & Lai, 2018). A function for cross-validating the final solution on an independent subsample of the data is included in the R scripts.

The following code shows how ACO can be applied to select ten items for a unidimensional model with the goal of optimizing CFI and RMSEA values. Note that the ACO function is user-defined and not presented in the example due to its length (see https://osf.io/yx4km/ for the full ACO function).

```
# Create function for model evaluation
fit.function = function(model){

  fit <- cfa(model = model, data = my.data)
  CFI <- fitMeasures(fit, "cfi")
  RMSEA <- fitMeasures(fit, "rmsea")

  phi.CFI <- 1/(1+exp(100*(.90-CFI)))       # Transform CFI
  phi.RMSEA <- 1-1/(1+exp(100*(.06-RMSEA))) # Transform RMSEA

  phi.overall <- (phi.CFI + phi.RMSEA)/2    # Optimization criterion
  return(phi.overall)

}

# Run ACO with RNG seed for replicability
set.seed(12345)
short <- ACO(dat = my.data,          #data
         list.items = names(my.data),# item names by factor
         i.per.f = 10,               # N items to select per factor
         max.iter = 20,              # stopping criterion
         ants = 30,                  # ants per iterations
         evaporation = .99,          # evaporation multiplier
         summaryfile.all = "ACO_all.csv",   # save all models to…
         summaryfile.final = "ACO_best.csv",# save best model to…
         fit.function,               # user-defined fit function
         monitor = TRUE)   # print/plot fit estimates during search
```

ACO Application

After all these preparations, ACO can be started to find the desired item set. In the following section, we present the results of our analysis. The goal was to find a model that would yield adequate model fit and factor saturation. The full model only fulfilled the prespecified factor saturation requirements. We thus decided to use ACO to identify an optimal or near-optimal selection of four items per facet that would result in acceptable model fit. We additionally included factor saturation ω in the optimization function so that values wouldn't drop below critical values due to reduced item number. We ran ACO ten times with 40 iterations and 60 ants per run. The best selected model yielded a substantially better model fit (CFI = .961; RMSEA = .038) than the original model (CFI = .789; RMSEA = .062). Factor saturation decreased due to the reduced item number (average ω = .77; full scale average ω = .87). Factor saturation levels of the short scale ranged from ω = .68 (*Self-consciousness*) to ω = .90 (*Depression*).

All final models found across the ten ACO runs yielded good model fit (CFI = .954 - 961; RMSEA = .038 - .042) and adequate factor saturation (with one exception: *Self-consciousness* ω = .64 - .68 across all runs). No final model was reported more than once. With such a large search space, this outcome follows a common pattern. Generally, there is no single best model, but a large number of equivalent models to choose from. At this point, one can compare these final models based on additional criteria (e.g., external correlations) or restart the search process with additional optimization criteria or higher cut-offs. However, with an increasing number of optimization criteria, this pool of "adequate" solutions becomes smaller until no model can be found that satisfies all criteria.

With the current specifications, ACO estimated on average 8,546 unique models, ranging from 4,187 (seed = 6) to 11,040 (seed = 1). The full search lasted on average 3:24 hours (ranging from 1:30 to 6:36) on a standard laptop with i7700HQ processor (4 cores with 2.80 GHz and 16GB RAM). The ACO item selection and pheromone update functions contribute only marginally to the overall computation time. Computation time is primarily driven by the estimation of the CFA models, which is run on a single core. To reduce the overall computation time, parallelization of the ants could be implemented or several ACOs with different seeds could be started simultaneously. As the dispersion of pheromone values becomes more extreme in later iterations (i.e., items will typically either have very high or low pheromones in higher iterations), the item samples begin to overlap more strongly in successive iterations. This tendency will lead to estimation of identical models in later iterations of the ACO search process. To counteract this issue, we included a check for redundancy, which uses the previously-saved optimization criterion value instead of re-estimating redundant models. On average, this check reduced the number of models estimated in each ACO run by 1804, ranging from 313 (seed = 6) to 3949 (seed = 10). This modification reduced the runtime of each ACO by around one fifth in comparison to the non-optimized version. This redundancy check reduces the impact of choosing an inefficient stopping

criterion (i.e., too many iterations with redundant model estimations because of extreme pheromone values).

Discussion of ACO as an Item Sampling Procedure

In the first section of this tutorial, we demonstrated how item selection procedures can be used to optimize several psychometric criteria of personality scales or self-report scales in general. We improved model fit of the IPIP NEO Neuroticism scale, while also taking into account factor saturation. ACO is a very flexible method that can be applied to improve any given set of criteria and it is up to the user to find a reasonable set of optimization goals. Other criteria that we didn't apply in this tutorial may for instance be measurement invariance (Olaru et al., 2018), correlations with external criteria (e.g., prediction of job success), item difficulty distributions (e.g., maintaining the difficulty distribution of the full scale; see Schroeders et al., 2016a), or even the balance between positively- and negatively-worded items (e.g., reducing the effect of acquiescence tendencies and predictive validity bias; Soto & John, 2018). Note that the search process becomes increasingly difficult and longer with an increasing number of optimization goals. Some items might, for instance, be suited to increase reliability because of a high redundancy with other items, but can negatively impact model fit as a result of residual correlations with the other redundant items. The user must find and specify an adequate set of optimization criteria that are suited for the research goals.

ACO can be used to improve various psychometric criteria simultaneously, such as factor structure, coverage of facets, reliability, et cetera. Instead of evaluating items separately, the scale is evaluated as an aggregate. This approach is unaffected by sequence effects and can balance several criteria. Another interesting advantage of considering scale-level instead of item-level criteria has been demonstrated by Yarkoni (2010). For many abbreviated personality scales, items were (among other criteria) typically selected based on the correlations between the item and the full scale (Donnellan et al., 2006; Gosling, Rentfrow, & Swann, 2003; Rammstedt & John, 2007; Saucier, 1994). While such simplistic

item selection procedures result in item sets that may seem valid and central to the scale, they are also very homogenous. Yarkoni (2010) showed how meta-heuristic selection procedures can reduce this redundancy by creating a short scale that captured the full variance of several longer scales.

ACO is one of several meta-heuristics that can be applied to select items. Another approach that has been used in psychological research is the *Genetic Algorithm* (Eisenbarth et al., 2015; Schroeders et al., 2016a; Scrucca, 2013; Yarkoni, 2010). The *Genetic Algorithm* uses the Darwinian evolutionary principles of selection, cross-over, mutation, and survival of the fittest to derive optimal short scales. Two R packages – the GA (Scrucca, 2013) and the stuart package (Schultze, 2018) – provide implementations of the algorithm suited for item selection. New algorithms based on other natural phenomena are constantly being developed (e.g., based on the foraging behavior of bees; Karaboga & Basturk, 2007; Karaboga, Gorkemli, Ozturk, & Karaboga, 2014). Evaluating which algorithm is superior to others is impossible, as this inevitably depends on the context in which it is applied (i.e., the no-*free-lunch theorem;* Wolpert & Mcready, 1997). However, when choosing an optimization algorithm for item selection, we recommend using a selection procedure that can optimize several scale-level criteria simultaneously instead of relying on sequential item selection based on item-level information (Olaru et al., 2015).

The R script presented in this article allows for maximal flexibility in the ACO search process, as every parameter setting in the optimization, ACO and cross-validation function can be modified in order to optimize any type of factorial model such as higher-order, bi-factor, ESEM or multi-group models (e.g., Olaru et al., 2018; Schroeders et al., 2016b). The commented script also provides insight into the workings of the ACO algorithm. Users might also want to try the *stuart* package in R (Schultze, 2018), which provides a number of example datasets and default settings. The *stuart* package also includes options for measurement invariance testing (across groups and measurement occasions), parallelization of

the search and cross-validation of the final model. The *stuart* package also provides a "brute-force" search to estimate all possible models if the number of potential solutions is small.

In the second section of this tutorial, we present current methodological advances that allow for the examination of latent models across a continuous moderator variable. More specifically, we show how *Local Structural Equation Modeling* (LSEM; Briley et al., 2015; Hartung, Doebler, Schroeders, & Wilhelm, 2018; Hildebrandt et al., 2016, 2009) can be used to investigate whether the personality model holds across a broad age range by examining differences in the personality model structure and factor means across age. LSEM estimates the specified personality model at each age point, and as such only yields meaningful results if the model fulfills model fit requirements. As many full personality scales suffer from problematic model fit, using ACO to sample the best-fitting indicators before using LSEM to examine the model across age can be a viable approach to study personality development.

Person Sampling

To compare the mean levels of psychological constructs across a moderator, it is important to consider the structure of the construct in question across the moderator as well. Means can only be meaningfully compared when the structure of the measurement models is equivalent across the moderator (Borsboom, 2006b; Wicherts & Dolan, 2010). For instance, in personality development studies, work-related items (e.g., "I keep my workplace tidy") will only be relevant for participants that are part of the work force. The importance of measurement invariance across age in personality development research has received more attention in recent years (Allemand et al., 2008, 2007; Brandt et al., 2018; Nye et al., 2016; Olaru et al., 2018; Small, Hertzog, Hultsch, & Dixon, 2003). Because the number of observations per year of age is often too small to estimate multi-group confirmatory factor analysis (MGCFA) for singular age points, participants are often grouped to larger units (e.g., decades). Thus, researchers are faced with the problem of finding an adequate age range and cut-offs for the age groups, which can be derived theoretically (e.g., developmental stages) or methodologically (e.g., equally large sample sizes). In an extreme group design, two age groups that are disjoint and far apart (e.g., young and old adults) are often compared (Allemand et al., 2008, 2007; Nye et al., 2016). Even if the sample size is sufficient to examine participants by years of age, this design also represents a form of grouping, with participants born at the beginning of each year being closer in age to participants born at the end of the previous year than to respondents at the end of the same year group. Such an artificial categorization of the moderator age which is continuous in nature will inevitably influence the findings (Hildebrandt et al., 2016; MacCallum et al., 2002). Hence, it can be difficult to identify non-linear developmental processes or possible onsets of differences (Hildebrandt et al., 2009; see Figure 5). In addition, information of within group differences are lost (MacCallum et al., 2002; see Figure 5). As a possible solution to these problems, we present *Local Structural Equation Modeling* (Briley, Harden, Bates, & Tucker-Drob, 2015;

Hildebrandt et al., 2016, 2009) as a non-parametric method to study structural differences and personality development over age as a continuous variable on comparatively small samples. Treating age as a continuous variable allows investigating onsets of age-differences, modeling individual differences across the age range, and more easily comparing studies due to their independence from grouping.

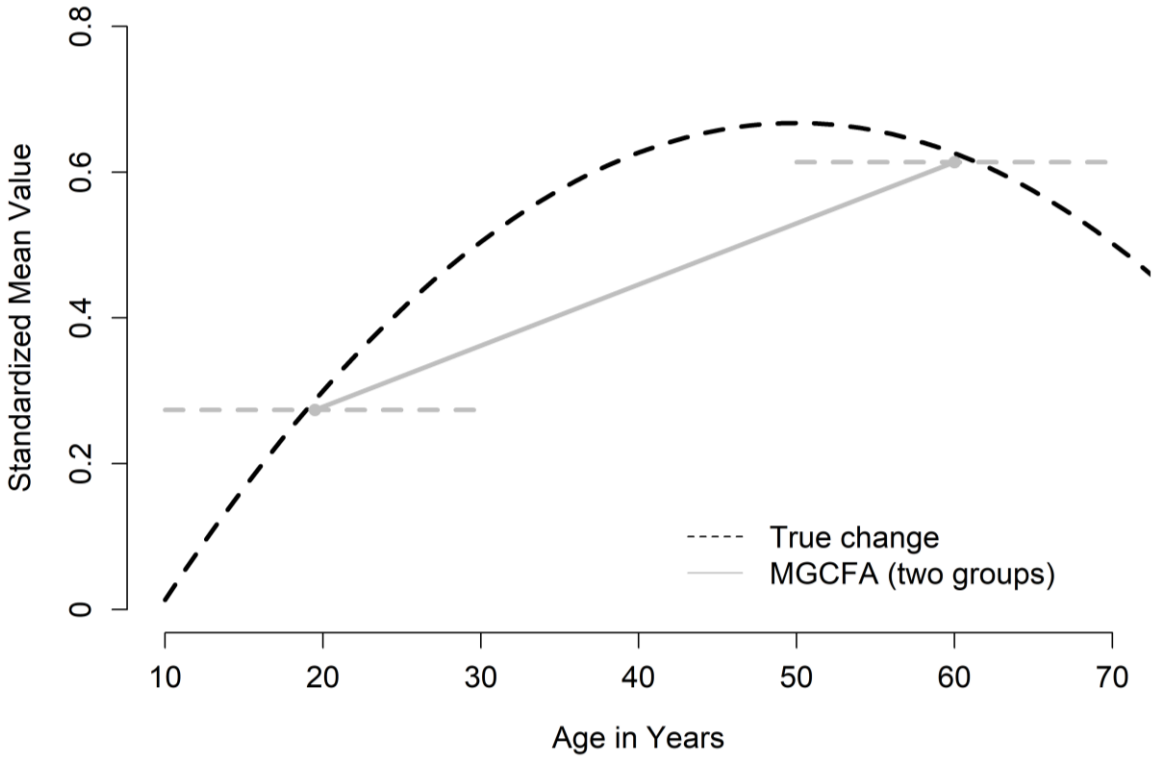


Figure 5. MGCFA findings on non-linear age differences. Illustrated are potential findings from a two-group MCGFA design (i.e., young vs. old). The "true change" line is based on the findings on Openness in Roberts et al. (2006). In addition to only identifying a linear increase the two-group MGCFA would underestimate the overall increase in the mean level of Openness.

Local Structural Equation Modeling

In the following section, we illustrate how LSEM can be used as a tool to examine personality development. LSEM are traditional structural equation models fitted along a moving weighting window across the moderator. Instead of separating the sample into distinct age groups and estimating models on each groups (i.e., MGCFA), models are estimated at each age point based on weighted samples (Wu & Zhang, 2006). Sample weights in LSEM follow a normal distribution around the desired focal point, with a full weight at the age point

and decreasing weights for participants further from this age point. For instance, estimating the measurement model at the age focal point 30 with LSEM will include participants with age 30 with the heaviest weights, participants aged 29 or 31 with slightly lower weights, participants of age 28 or 32 with even lower weights, and so on (see Figure 6 for weighting of participants of age 27, 30, and 33 in the current analysis). By weighting, and hence (partially) including observations near the focal points of the moderator variable, the effective sample size for each focal point is increased. The underlying rationale is that people close to each other on the moderator variable should be more decisive for parameter estimation compared to people far away on the moderator.

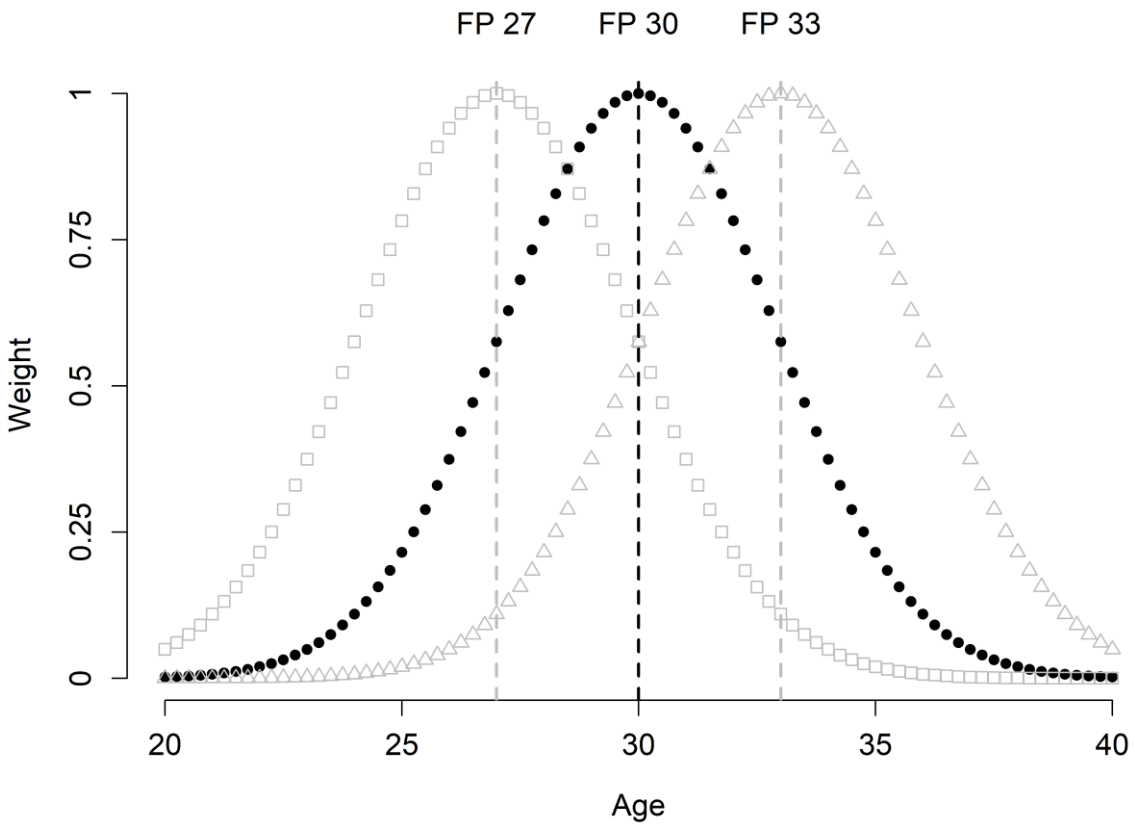


Figure 6. Gaussian sample weights in Local Structural Equation Modeling. Depicted are sample weights in the current analysis at the focal points 27, 30, and 33 (bandwidth parameter $h = 2$).

The increased sample size due to the inclusion of participants around the focal points allows for the SEM estimation across the entire range of the moderator and provides higher power and a more robust parameter estimation at each focal point, in comparison to MGCFA.

LSEM allows for the examination of all model parameters at each age point. Age-associated differences in the parameters (e.g., factor means, factor loadings, model fit etc.) can then be easily visualized.

Person Sampling using LSEM

In the following, we give a detailed overview of each step required to use LSEM and provide some basic recommendations. To illustrate these steps, we applied LSEM using the R-package *sirt* (Robitzsch, 2019) to study measurement invariance and normative personality differences across age. The baseline model we used for this application is the model we derived using ACO. This model had sufficient model fit and reliability in terms of factor saturation. Thus, in the current example, we have already improved all necessary psychometric properties of the model in the previous section of this tutorial for the total sample and we used the abbreviated scale to examine personality development across age. We think it is important to stress this point because many broad personality measures suffer from inadequate model fit. LSEM tests the given model across weighted subsamples and hence will probably suffer from the same model misspecifications as the SEM on the full sample. Thus, it might be wise to apply an item selection procedure such as ACO to optimize the model before examining across a continuous moderator variable.

How does LSEM weight participants?

Sample weights in LSEM follow a normal distribution around the desired focal point, with a maximum weight of 1 at the age point (i.e., full inclusion) and decreasing weights with increasing distance to the age point (see Figure 6). Because the weighted samples include observations around a focal point, the number of observations for that model is higher than it would be if only the observations with the exact focal age were included, which makes structural equation modeling feasible even if the sample size is small.

The weighting procedure can be adjusted via the bandwidth parameter *h* in the function `lsem.estimate` in the R package *sirt*. The parameter *h* can be interpreted as a

multiplier of the breadth of the weighting kernel. A higher bandwidth results in higher weights for participants who are distant from the focal point. This outcome further increases the resulting weighted sample size (N_{eff}). Please note that increasing h will also cause neighboring focal points to show stronger overlap in the weighted samples. As such, the bandwidth parameter can also be interpreted as a smoothing parameter. With increasing bandwidth, the standard error of parameter estimates will decrease since they are less prone to sampling effects. While a bandwidth factor of $h = 1.1$ corresponds to an approximation of the rule-of-thumb bandwidth estimator for a Gaussian function (Silverman, 2018), Hildebrandt and colleagues (2016) compared different bandwidth parameters on a sample with 30 to 60 observations per focal point and recommend a value of 2, which will reduce the effects of noise while still being accurate enough to detect differences in the model. Figure 7 shows how participants around the focal age point of 30 were weighted in the current data set using a bandwidth parameter of 1.1 and 2. Note that a bandwidth parameter of 0 will result in models being estimated solely based on participants at the focal point. This special case of LSEM corresponds to a MGCFA with each focal point representing a group with no overlap between samples.

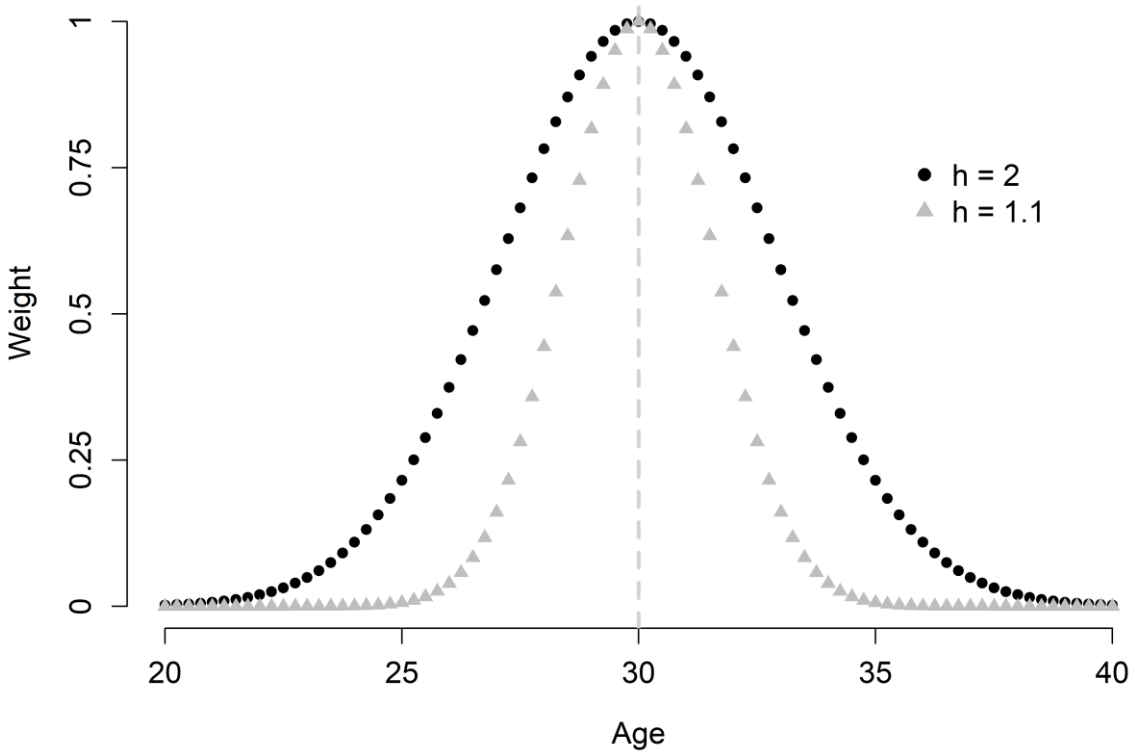


Figure 7. LSEM sample weights with a bandwidth parameter of 1.1 and 2.

What is the difference between observed and effective sample size?

The sample size resulting from the weighting procedure is referred to as effective sample size. The effective sample size can be computed for each focal point by summing up the corresponding LSEM weights across all participants. As shown in Figure 8, the effective sample size in the current analysis was much larger than the number of observations at each age.

In the current analysis, the weighting function resulted in an effective sample size 3 to 7 (bandwidth = 1.1) or 4 to 14 (bandwidth = 2) times larger than the original sample size at each focal point. With bandwidth parameters of 1.1 or 2 respectively, the weighted sample sizes was sufficiently large for the intended SEM up to age 53 or 57 (for sample size considerations in a SEM context, refer to Bentler & Chou, 1987; Boomsma, 1985; Wolf, Harrington, Clark, & Miller, 2013). In contrast, the observed sample size only allowed an estimation of models with $N > 200$ up to 38 years of age. The increased sample size also

results in more robust and precise parameter estimates, as well as higher power. The sample size can be examined using the `summary` function on the fit object resulting from `lsem.estimate`. While we both recommend and applied a bandwidth of 2, it can be increased if the effective sample sizes are too small. However, this modification will also increase the overlap and the dependency between weighted samples and blur differences across the moderator variable.

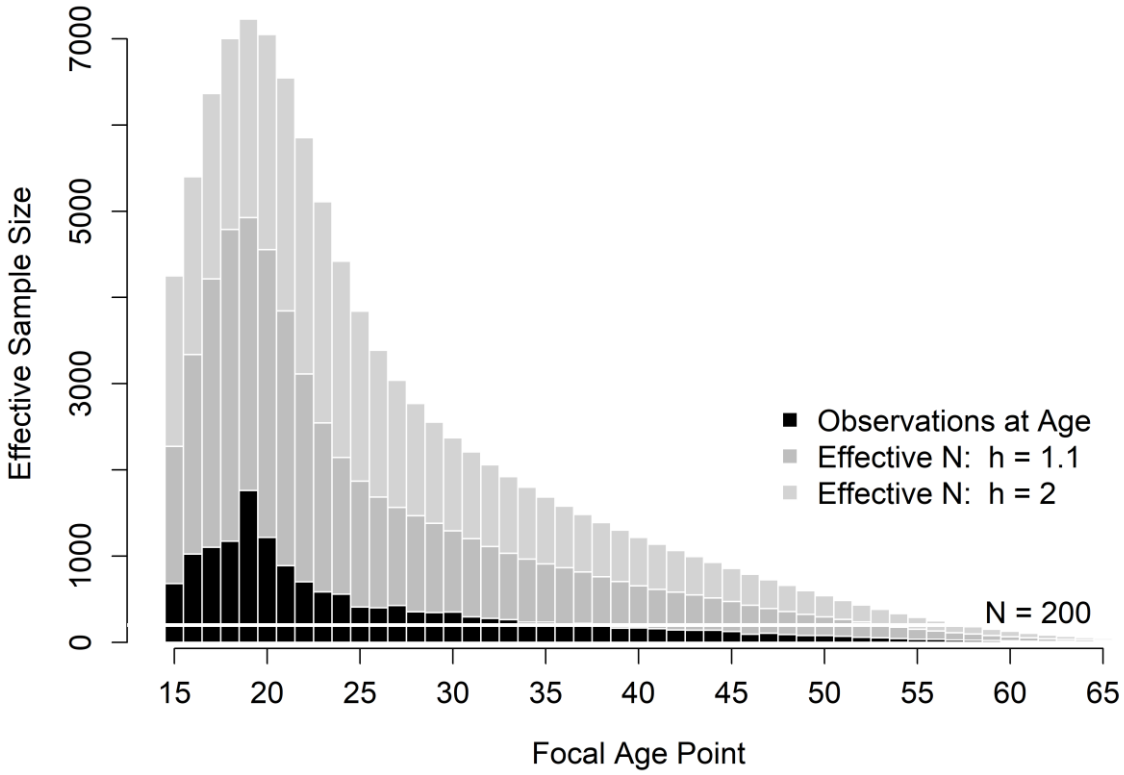


Figure 8. Number of Observations and Effective Sample Size Across Focal Points. *h* = bandwidth parameter. The sample size at each focal point is represented by the absolute height of the bars. For example, at focal point 15: *N* Observations = 680; Effective *N* with bandwidth parameter 1.1 = 2271.3; Effective *N* with bandwidth parameter 2 = 4251.1.

What do I have to consider when choosing focal points?

In the typical LSEM application, focal points correspond to the levels of the moderator variable, which allows for a fine-grained examination of parameter variations across the moderator. In personality development, participants will usually be weighted by years of age. If age, for instance, is measured in months or weeks, using these levels provides a more

detailed examination of personality differences across age. The number of focal points, however, also depends on the number of observations across the moderator. If the sample size is low, it can be useful to aggregate levels of the moderator variable (e.g., using years of age instead of months). There are no limits to the number of focal points, but depending on the research question, focal points can also be spaced further apart (e.g., when no differences are expected at a more fine-grained level, see Olaru et al., 2019). LSEM can also be applied on the moderator variables with a small number of levels (e.g., restricted age range or years of education, see Hartung, Doebler, Schroeders, & Wilhelm, 2018), as the weighting function can be adapted accordingly using the bandwidth parameter.

The distribution of the moderator variable does not only affect the choice of focal points, but also the resulting effective moderator variable value in the weighted samples. At and around each focal point, the symmetrical Gaussian weighting function is applied to increase the effective sample size for model estimation. Participants are solely weighted based on their distance to the focal point—independent of the direction of the difference. If the number of observations is not equally distributed across the moderator (e.g., a low number of older participants), the resulting weighted sample is skewed towards points with higher numbers of observations. This imbalance is often salient at the extreme ends of the moderator variable, because the lack of participants outside the boundaries results in an asymmetric weighting that will be skewed towards the middle of the moderator variable (Hildebrandt et al., 2016). Therefore, the value of the resulting moderator variable for the weighted samples – which will be referred to as the "effective moderator variable"– might deviate from the targeted focal point. This effect is often visible at the boundaries of the moderator variable or when the number of observations is not equally distributed across the moderator (see Figure 9 for difference between targeted age and resulting effective mean age in the current analysis). To bypass the issue of asymptotic weighting at the boundaries of the moderator variable, researchers can exclude focal points at the extremes (Hartung et al., 2018). In the current

analysis, we excluded focal points at the extremes with a deviation in effective age of more than a year or an effective sample size below 200. This resulted in an examinable age range from 16 to 53 (bandwidth = 1.1) or 18 to 52 (bandwidth = 2) respectively (see Figure 9).

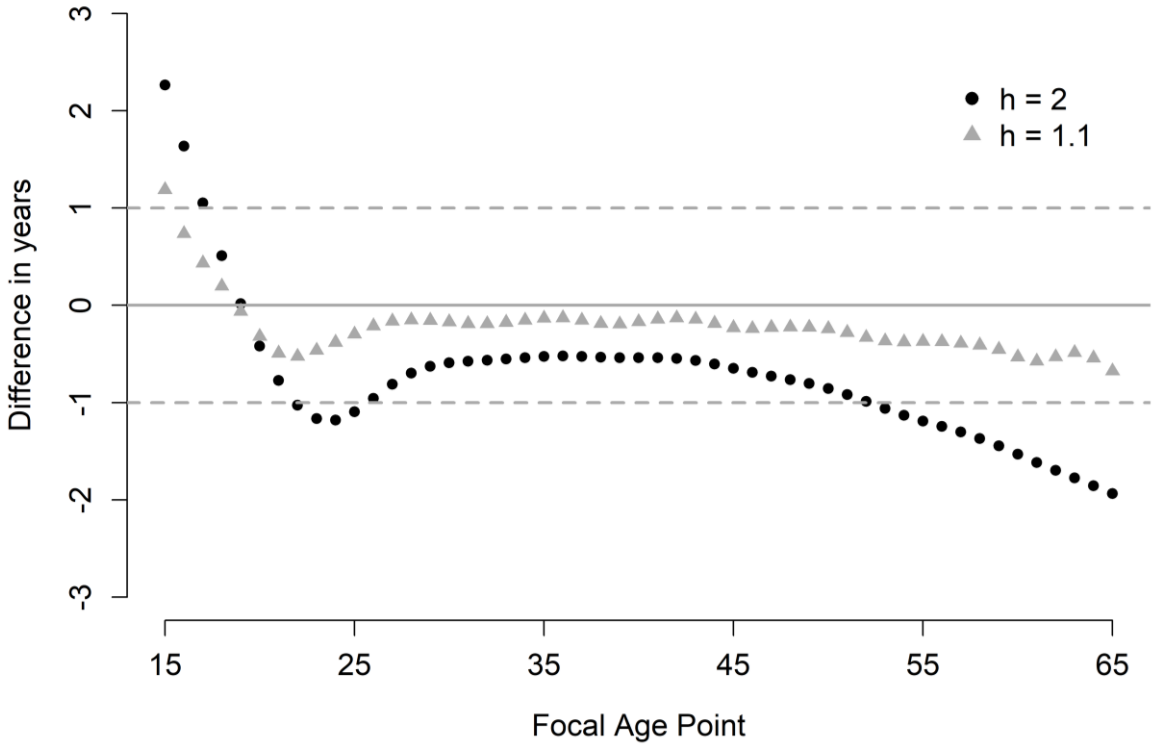


Figure 9. Difference between the effective moderator variable and focal point in years. *h* = bandwidth parameter. Deviations are depicted for a bandwidth parameter of 1.1 and 2. Deviations are particularly high at the age extremes and around the focal age point 24 due to a lack of participants outside the age range and an overrepresentation of 19-year-old participants respectively.

How can I constrain parameters to equality across the moderator variable?

Parameters can be constrained in the model specification of *lavaan* (as the `lsem.estimate` function relies on *lavaan* to estimate models; Rosseel, 2012). Model parameters can also be fixed to a value before running a model using the regular *lavaan* procedure for doing so. However, since all models are estimated independently at each focal point, `lsem.estimate` currently does not provide a way of automatically constraining parameters to equality across the moderator (as is typically done in MGCFA; e.g., `groups.equal=c("loadings", "intercepts")`). A procedure based on MGCFA estimation is currently under development. However, we want to point out that LSEM is often

used to study parameter variation across a continuous moderator variable. To determine where exactly violations of measurement invariance occur, variation of model parameters and their confidence intervals across the moderator variable need to be studied. For a thorough investigation of violations of measurement invariance, we recommend the resampling approach (Hartung et al., 2018), which tests parameter equivalence across each pair of focal points using classical MGCFA testing on non-overlapping samples. We discuss this approach in more detail in the discussion.

LSEM Application

After determining the range and number of focal points, LSEM can be used to study the differences in the model parameters across age as a continuous moderator variable. In the following section, we present the LSEM application to the previously derived short scale, show how to examine model parameters, and suggest how to interpret variations.

In order to freely estimate and compare the means of the latent variables across age, we used the effects-coding method (Little, Slegers, & Card, 2006). Hence, the means of all factor loadings are fixed to 1 and the mean of the item intercepts is fixed to 0. As a result, all measurement parameters including factor variances and means can be freely estimated and are represented in the same metric as the underlying indicators (Little et al., 2006, p. 63). Effect-coded factor means represent a weighted mean based on the indicators weighted by the factor loading – or, in other words, how much each indicator contributes to the factor. By *z*-standardizing variables before model estimation, differences in factor means across focal points can be interpreted similar to the effect size Cohen's *d*. Effect-coded factor variances represent the average indicator variance explained by the latent variable (Tucker-Drob & Salthouse, 2008). In addition, the variance of the latent variables can be interpreted as the proportion of explained variance (= R^2). This interpretation is particularly interesting when examining the structure of a model. Differences within the structure of the personality factors can manifest as differences in the factor loadings and item residuals, as well as factor

variances and covariance (Molenaar, Dolan, Wicherts, & van der Maas, 2010). These differences in the factor structure are typically studied in the context of factor differentiation-dedifferentiation, which is particularly prominent in intelligence research (for an overview see Hartung et al., 2018). Examinations of differentiation-dedifferentiation in the context of personality are rare (Murray, Booth, & Molenaar, 2016), but can also provide more insight into structural differences of personality across various moderators.

In the following, we show how parameter estimates across age can be examined in a LSEM context. More specifically, we present model fit, factor loadings, item intercepts, and factor means of the Neuroticism model across age and demonstrate how trends can be interpreted. All model parameters examined can be extracted using the generic R functions `summary` or `plot` on the fit object of `lsem.estimate`. The following R code shows how LSEM was applied in the current analysis and how the output object can be examined:

```
fit <- lsem.estimate(
       data = dat.scaled,
       moderator = "AGE",          # Moderator variable
       moderator.grid = 18:52,     # Focal points
       lavmodel = aco.little,      # Model
       h = 2,                      # Bandwidth parameter
       meanstructure = TRUE,       # Estimate factor means
       residualize = FALSE,        # across the moderator
       standardized = TRUE)        # Provide standardized estimates

summary(fit)
plot(fit)
```

Model Fit

Figure 10 Panel A and B show CFI and RMSEA trends of the full and shortened Neuroticism model across adulthood. In contrast to the full model, the optimized model fitted well (CFI ≥ .95 and RMSEA ≤ .06) across the majority of age samples (with still adequate fit in older age points; CFI ≥ .90). Both models showed a slight deterioration at the end of the considered age range, most likely due to the reduced sample size (see also Figure 8).

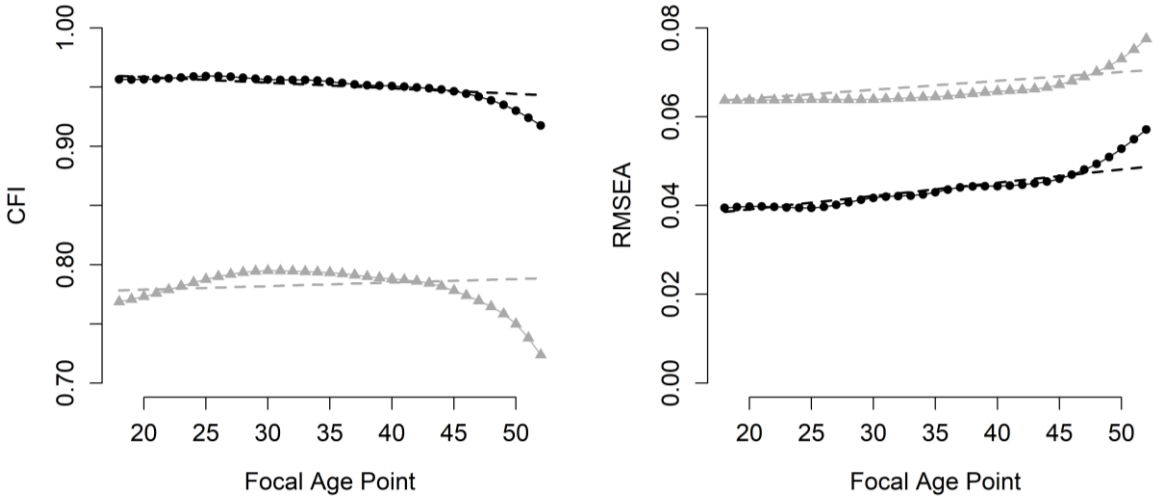


Figure 10a, b. CFI and RMSEA of the full and short model across focal points. Dots represent point estimates at each focal point. The dashed lines represent a linear approximation of the difference pattern across focal points. Fit values for the full model are presented in grey, for the short model in black.

Factor Loadings and Intercepts

After considering the overall model fit, the measurement parameters such as factor loadings and item intercepts should be examined in more detail to study measurement invariance. While currently there is no direct possibility to test for measurement invariance in the *sirt* package, confidence intervals are reported/plotted alongside the parameter estimates. Examining the values and confidence intervals across the moderator can help find potential measurement invariance violations. Figure 11 Panel A and B show the unstandardized factor loading and item intercept of item 251 (“Feel comfortable with myself”; Neuroticism facet *Depression*). The unstandardized factor loading was close to 1 and remained stable across focal points. Confidence intervals of the point estimates overlapped across the entire range and indicated equivalence of the factor loading over age. In contrast, the intercept decreased linearly across age until confidence intervals no longer overlapped. This indicates a violation of measurement invariance in the item intercept over age. Please note that these examinations are descriptive and do not substitute proper measurement invariance testing. As the weighted samples are overlapping and models from different focal points are partly based on the same participants, the equality of single indicators across the moderator variable cannot be tested

with classical inference testing procedures. We present a resampling approach for evaluation of parameter equivalence across the moderator variable in the discussion.

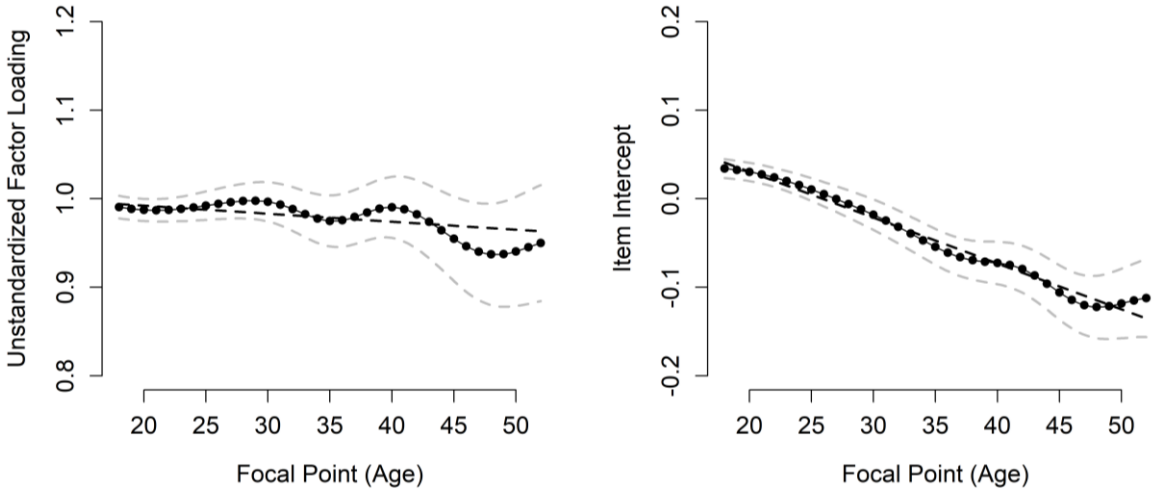


Figure 11a, b. Factor loading and intercept of the Depression item "Feel comfortable with myself" across focal points. Black dots represent point estimates at each focal point. The dashed black line represents a linear approximation of the pattern across focal points. Dashed grey lines represent the 95% confidence interval of the parameter estimation.

Factor Means

Factor means of the facets *Immoderation*, and *Vulnerability* are presented in Figure 12 (see https://osf.io/yx4km/ for factor means of all six facets across age). Both factor means showed a linear decline across most of the examined age range – which is in line with expectations (Roberts, Walton, & Viechtbauer, 2006). *Immoderation* means first increased up until age 27 before decreasing, whereas *Vulnerability* showed a positive trend after focal point 44. These findings demonstrate the strength of LSEM to examine non-linear trends of personality development. The factor means should be interpreted with caution, as measurement invariance across age has not been established.

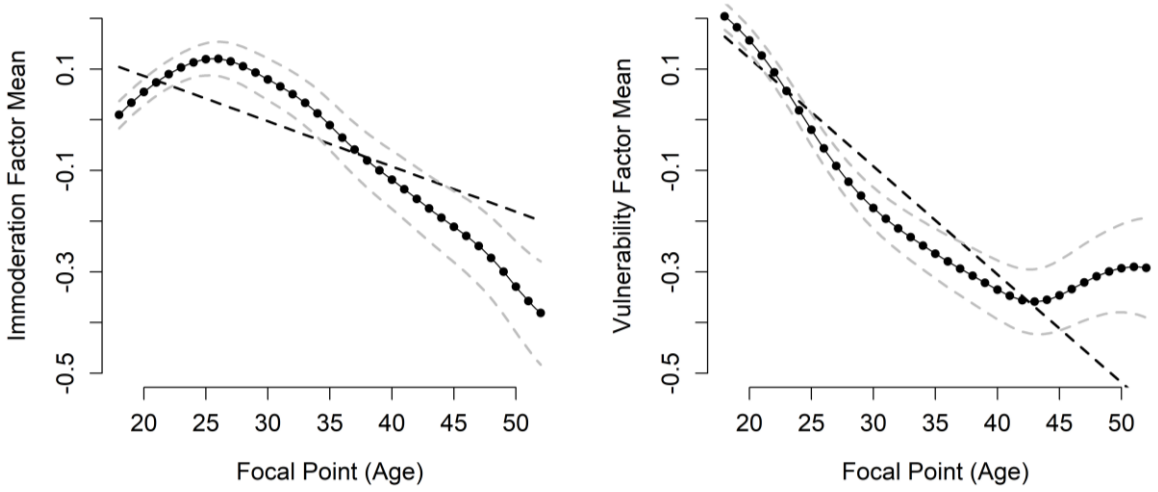


Figure 12a, b. Factor means of Immoderation and Vulnerability across focal points. Black dots represent point estimates at each focal point. The dashed black line represents a linear approximation of the difference pattern across focal points. Dashed grey lines represent the 95% confidence interval of the parameter estimation.

Discussion of LSEM as Person Sampling Approach

In the second section of this tutorial, we demonstrated how LSEM can be used to examine personality development across a continuous age variable. LSEM allows for a more nuanced examination of age-related differences than classical MGCFA applications. In comparison to traditional methods, LSEM has several advantages, such as detecting non-linear trends, examining sources of measurement invariance violations, and avoiding capitalizing on chance due to arbitrary age groups. In the current example, we found an initial increase in the mean of *Immoderation* up until age 27 and a subsequent decline across the examined age range. This trend might be due to the transition from a less-regulated life at post-secondary institutions to a more structured work life that then arguably becomes typical. The increase found for *Vulnerability* at the end of the age range available here might be attributed to an increased physical frailness in higher years of age (Hill & Roberts, 2016) or decreasing self-esteem (Orth, Robins, & Widaman, 2012). Whereas the focus of personality development research often rests on normative differences across age (Roberts et al., 2006; Soto et al., 2011; Terracciano, McCrae, Brant, & Costa, 2005), structural differences in the personality model can also indicate meaningful variations in personality across age. Age-

associated differences in the factor loadings or covariance across age can be related to a shift in the factor composition and thus differences in the interpretation of the personality factors across age. LSEM is a useful tool that can examine questions of structural differences of the personality across age, both at the facet (e.g., changing factor loadings or item residuals) or second-order factor level (e.g., changing facet covariance or second-order factor loadings; see Molenaar et al., 2010; Murray et al., 2016).

LSEM is a nonparametric approach for visualizing differences in factor model parameters across a moderator, without *a priori* specification of the shape (e.g., linear, quadratic, exponential) of the moderation effect. Another approach which allows for continuous moderation of factor model parameters is moderated factor analysis (MFA; Cheung, Harden & Tucker-Drob, 2015; Tucker-Drob, 2009; Molenaar, et al., 2010). In contrast to LSEM, model parameters such as factor loadings are moderated according to a parametric function (e.g., a linear function). When the chosen parametric function conforms to the shape of the true moderation function, MFA is advantageous for increasing power while decreasing false positives. However, to the degree that the parametric function deviates from the shape of the true moderation function, MFA has lower power to detect the interaction, and may give a false impression of the true shape. LSEM is advantageous for diagnosing the shape of a moderation function without imposing a specific class of moderation curves (e.g., linear, quadratic, exponential) in advance.

The weighting procedure of LSEM artificially increases the effective sample size for model estimation, which also reduces the standard errors of model parameters. In addition, the overlap of the weighted samples results in smoothed parameter estimations across the moderator, which will fluctuate less due to noise (i.e., sampling effects). The downside of the partly overlapping samples is that the measurement invariance of the model cannot be tested by traditional means of inference testing (e.g., χ^2-test). To test if the parameter differences across age are significant, Hülür, Wilhelm, and Robitzsch (2011) proposed a permutation test.

In this approach, several copies of the data set are created with randomly-assigned moderator values. LSEM models are then estimated on each of these data sets. Due to the randomly assigned moderator values, model estimates are independent of the moderator variable in these data sets. As such, the parameter distribution across the permutated datasets can be used to test whether that parameters in the original model deviate from the mean of the distribution. However, as this approach only provides invariance testing for single parameters and does not allow for increasingly strict constraints on parameters, Hartung and colleagues (2018) developed a resampling procedure that is similar to classical MGCFA testing and provides global measurement invariance testing. To eliminate the overlap between weighted samples, participants are assigned to two disjoint groups of participants for each focal point combination. Instead of splitting up participants by a fixed cut-off value, respondents are assigned based on the corresponding LSEM-weights. The weights are used as drawing probabilities to assign participants to two groups of equal size with a mean on the moderator variable that ideally equals the focal point. After creating two groups for each focal point combination, parameter equivalence is evaluated with fit statistics used in the traditional MGCFA invariance testing. The R script for the resampling procedure can be downloaded at https://osf.io/yx4km/ .

LSEM was only introduced recently and many extensions are currently under development. One such extension is the possibility of examining differences across more than one moderator. This approach accounts for possible interaction effects of moderator variables. Hartung and colleagues (2018) used *Multivariate Local Structural Equation Model* (MLSEM) to examine the differences in the structure of intelligence across both age and years of education simultaneously. From a methodological perspective, in the case of two moderators, a two-dimensional grid across the two moderators is used to weight the sample. Participants are weighted by their Mahalanobis distance to the corresponding grid coordinates. Model

parameters can then be plotted and examined across the two moderator variables (see Hartung et al., 2018), which allows for the identification of potential interaction effects.

Of course, applications of LSEM is not restricted to age as a moderator. LSEM can also be used to study differences in model parameters across other continuous moderator variables such as socio-economic status, cognitive abilities, or cultural background (e.g., individualism – collectivism). LSEM allows researchers to identify differences in the structure of personality and relevant behaviors and, thus helps in understanding of personality and personality assessment. This flexibility also applies to the latent trait investigated. LSEM (and ACO) can be applied in a similar fashion to study differences across a continuous moderator on any measure with an underlying reflective model (i.e., a latent trait influencing the indicators; Borsboom, 2006a), such as measures of cognitive abilities or motivation.

While we presented LSEM as a method of examining age differences in personality in a cross-sectional setting, applications on longitudinal data are also possible. LSEM as a person sampling method can for instance be applied to latent growth curve models (Preacher, Wichman, Briggs, & MacCallum, 2008) to study differences in model parameters across any relevant continuous moderator variable. For instance, LSEM can be used to study moderation effects of socio-economic status, cognitive abilities, or age on longitudinal change trajectories. Combining longitudinal models with LSEM allows researchers to examine the interaction between age and other moderator variables on the development of personality, or any other latent trait for that matter.

General Discussion

The nature of our world incites us to sample whenever we want to make observations. Testing implies presenting a sample of tasks to a sample of people. Item- and person-sampling have to be subject to careful considerations before meaningful deductions can be made. Items for personality inventories are often sampled from an unspecified item universe that often represents the researchers' interpretation of the personality traits (Loevinger, 1965). Results

are thus tied to the specific item sample that have been used. When sampling participants, it often remains unclear to what extent the results are representative for the general population. Findings are therefore always specific to the sample of collected persons. Whether the findings also hold on other samples with different characteristics concerning for instance age, educational, or cultural background has to be considered before findings can be generalized.

Poor model fit of broad personality measures can have several causes. One important source of these problems is the quality of the items/indicators that are used. Items might be sampled from a restricted item pool on restricted person samples (e.g., student samples). Often used and popular personality measures are still based on items that show simple structured loadings in *Varimax* rotated solutions of *Principal Component Analysis* in unrepresentative person samples. Another potential source of model misfit is that existing item sets might be tailored for use with subjects of a particular age, education, or cultural background. Studying the relevance of such moderators for the quality of personality measurement is still in its infancy, but both methods presented here can be used to examine the appropriateness of personality indicators based on characteristics of the person sample intended to assess (e.g., by sampling appropriate items depending on participants' age, see Olaru et al., 2019).

In this tutorial, we presented ACO as *one* specific method to optimize prespecified psychometric properties by sampling items. Other metaheuristic approaches, for instance, rely on evolutionary principles of cross-over and selection (Genetic Algorithm; Yarkoni, 2010), the foraging behavior of bees (Artificial Bee Colony; Karaboga & Basturk, 2007) or memory structures (Tabu Search; Glover, 1990) to solve such optimization problems. Typically, item sampling implies reducing an initial item set and using the short scale in subsequent studies. However, one can also sample items from a large item pool for each participant at the time of the assessment. By doing so, the issue of item specificity can be evaluated. One noteworthy example in personality research using broad item sampling is the *Synthetic Aperture*

Personality Assessment (SAPA) project, which is currently collecting data on approximately 7,000 items measuring temperament, abilities, and interests. Its scope is so vast that it covers items from 92 openly-available personality inventories (Condon & Revelle, 2015), which comes at the cost of massive missingness because participants only work on a small, randomly-drawn item set. However, with sufficiently large person samples, this approach allows for findings that can be generalized beyond specific item sets.

In this tutorial, we presented LSEM as a person-sampling method that can be used to generalize findings from a specific sample to more abstract levels of person characteristics instead. Various other person sampling methods – such as survey weights or propensity score matching – are used to ensure that results can be generalized beyond specific person samples. The advantage of LSEM is that it is a flexible way to analyze models across continuous moderator variables. In the current application, we used LSEM to study personality differences across age. However, LSEM can also be used across a wide range of other continuous moderators – such as socio-economic status, other personality factor or cognitive ability scores – to study inter-individual differences in mean-levels, structure, or developmental trajectories (in the case of longitudinal models).

While we presented ACO and LSEM separately in this tutorial, we think that item- and person-sampling issues should be studied in tandem, Accordingly, a combination of these methods can substantially expand perspectives on personality and personality development. ACO can, for example, be used to identify the most measurement-invariant items across age before examining normative differences across age (Olaru et al., 2018). ACO can also be used to challenge the notion that personality can be measured with the same set of indicators across the entire life starting in young adulthood. Given the large differences in tasks, roles (Roberts et al., 2006), and situations (Bleidorn et al., 2018; Wrzus & Roberts, 2017; Wrzus et al., 2016) across life, different personality measures might be needed to capture the factors at different stages in life. Using ACO and LSEM simultaneously, items can be sampled at each focal

point separately to identify prototypical indicators at specific life stages, which can then be compared across the entire age range (Olaru et al., 2019).

In this tutorial, we applied both tools in the context of cross-sectional questionnaire data. Both approaches can be similarly applied in other contexts such as cognitive tests and longitudinal settings. We used ACO and LSEM on the standard psychometric model of personality, which is a reflective one (i.e., a latent trait influencing the manifest variables; Borsboom, 2006a). ACO can also be applied to formative models (i.e., manifest variables determining the construct; Borsboom, 2006a) to identify indicators that maximize external correlations of the formative construct. It is also possible to combine network analysis (Costantini et al., 2015) with LSEM to examine structural differences in the network across a continuous moderator variable. This approach can be used to examine the centrality of behaviors across age and structure of personality across a wide variety of moderators. We wish to point out that ACO and LSEM (or similar procedures) are tools that can be applied to either improve the findings on already established research questions (e.g., normative differences across continuous moderators) or create new opportunities to study a wide variety of new research questions (e.g., whether prototypical items differ as a result of the respondents' age).

References

Allemand, M., Zimprich, D., & Hendriks, A. A. J. (2008). Age differences in five personality domains across the life span. *Developmental Psychology*, *44*, 758–770. DOI: 10.1037/0012-1649.44.3.758

Allemand, M., Zimprich, D., & Hertzog, C. (2007). Cross-sectional age differences and longitudinal age changes of personality in middle adulthood and old age. *Journal of Personality*, *75*, 323–358. DOI: 10.1111/j.1467-6494.2006.00441.x

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 397–438. DOI: 10.1080/10705510903008204

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*, 186–203. DOI: 10.1207/s15328007sem1302_2

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*, 78–117. DOI: 10.1177/0049124187016001004

Biemer, P. P., & Christ, S. L. (2008). Weighting survey data. In: de Leeuw, E., Hox, J., Dillman, & D. (Eds.) *International Handbook of Survey Methodology* (pp. 317–341). New York: Routledge.

Bleidorn, W., Hopwood, C. J., & Lucas, R. E. (2018). Life events and personality trait change. *Journal of Personality*, *86*, 83–96. DOI: 10.1111/jopy.12286

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *50*, 229–242. DOI:10.1007/BF02294248

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences*, *11*, 515–524DOI:10.1016/0191-8869(90)90065-Y

Borsboom, D. (2006a). The attack of the psychometricians. *Psychometrika*, *71*, 425-440. DOI:10.1007/s11336-006-1447-6

Borsboom, D. (2006b). When does measurement invariance matter? *Medical Care*, *44*, S176–S181. DOI: 10.1097/01.mlr.0000245143.08679.cc

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research & Perspective*, *6*, 25–53. DOI:10.1080/15366360802035497

Brandt, N. D., Becker, M., Tetzner, J., Brunner, M., Kuhl, P., & Maaz, K. (2018). Personality across the lifespan. *European Journal of Psychological Assessment*, 1–12. DOI:10.1027/1015-5759/a000490

Briley, D. A., Harden, K. P., Bates, T. C., & Tucker-Drob, E. M. (2015). Nonparametric estimates of Gene × Environment interaction using local structural equation modeling. *Behavior Genetics*, *45*, 581–596. DOI: 10.1007/s10519-015-9732-8

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review*, *90*, 105-126. DOI: 10.1037/0033-295X.90.2.105

Cheung, A. K., Harden, K. P., & Tucker-Drob, E. M. (2015). From specialist to generalist: Developmental transformations in the genetic structure of early child abilities. *Developmental Psychobiology*, *57*, 566-583. DOI: 10.1002/dev.21309

Condon, D., & Revelle, W. (2015). Selected personality data from the SAPA-Project: On the structure of phrased self-report items. *Journal of Open Psychology Data*, *3,* e6. DOI:10.5334/jopd.al

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin. .

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the revised NEO Personality Inventory. *Journal of Personality Assessment*, *64*, 21–50. DOI:10.1207/s15327752jpa6401_2

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13–29. DOI: 10.1016/j.jrp.2014.07.003

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley & Sons.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, *84*, 151–161. DOI: 10.1162/003465302317331982

Deneubourg, J.-L., Aron, S., Goss, S., & Pasteels, J. M. (1990). The self-organizing exploratory pattern of the argentine ant. *Journal of Insect Behavior*, *3*, 159–168. DOI : 10.1007/BF01417909

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP Scales: Tiny-yet-effective measures of the Big Five Factors of personality. *Psychological Assessment*, *18*, 192–203. DOI:10.1037/1040-3590.18.2.192

Dorigo, M., & Stützle, T. (2010). Ant Colony Optimization: Overview and Recent Advances. In M. Gendreau & J.-Y. Potvin (Eds.), *Handbook of Metaheuristics* (146th ed, pp. 227–263). Boston, MA: Springer US. DOI:10.1007/978-1-4419-1665-5_8

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, *78*, 535–543. DOI: 10.2307/2288115

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory–Revised (PPI-R). *Psychological Assessment*, *27*, 194–202. DOI:10.1037/pas0000032

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*, 105–120. DOI:10.1037/1082-989X.12.1.105

Glover, F. (1990). Tabu Search: A tutorial. *Interfaces*, *20*, 74–94. DOI: 10.1287/inte.20.4.74

Gnambs, T., & Schroeders, U. (2017). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment*, 1-15.DOI:10.1177/1073191117746503

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504–528. DOI:10.1016/S0092-6566(03)00046-1

Griffin, S. A., & Samuel, D. B. (2014). A closer look at the lower-order structure of the Personality Inventory for DSM-5: Comparison with the Five-Factor Model. *Personality Disorders: Theory, Research, and Treatment*, *5*, 406–412. DOI:10.1037/per0000074

Hartung, J., Doebler, P., Schroeders, U., & Wilhelm, O. (2018). Dedifferentiation and differentiation of intelligence in adults across age and years of education. *Intelligence*, *69*, 37–49. DOI:10.1016/j.intell.2018.04.003

Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., & Wilhelm, O. (2016). Exploring factor model parameters across continuous variables with local structural equation models. *Multivariate Behavioral Research*, *51*, 257–258. DOI:10.1080/00273171.2016.1142856

Hildebrandt, A., Wilhelm, O., & Robitzsch, A. (2009). Complementary and competing factor analytic approaches for the investigation of measurement invariance. *Review of Psychology*, *16*, 87–102.

Hill, P. L., & Roberts, B. W. (2016). Chapter 11 - Personality and Health: Reviewing Recent Research and Setting a Directive for the Future. In K. W. Schaie & S. L. Willis (Eds.), *Handbook of the Psychology of Aging* (8th ed.,pp. 205–218). San Diego, CA: Academic Press. DOI:10.1016/B978-0-12-411469-2.00011-X

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. DOI:10.1080/10705519909540118

Hülür, G., Wilhelm, O., & Robitzsch, A. (2011). Intelligence differentiation in early childhood. *Journal of Individual Differences*, *32*, 170–179. DOI:10.1027/1614-0001/a000049

Janssen, A. B., Schultze, M., & Grötsch, A. (2015). Following the ants: Development of short scales for proactive personality and supervisor support by Ant Colony Optimization. *European Journal of Psychological Assessment*, *33*, 409–421. DOI:10.1027/1015-5759/a000299

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78–89. DOI:10.1016/j.jrp.2014.05.003

Karaboga, D., & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of Global Optimization*, *39*, 459–471.DOI: 10.1007/s10898-007-9149-x

Karaboga, D., Gorkemli, B., Ozturk, C., & Karaboga, N. (2014). A comprehensive survey: artificial bee colony (ABC) algorithm and applications. *Artificial Intelligence Review*, *42*, 21–57. DOI: 10.1007/s10462-012-9328-0

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychological Medicine*, *42*, 1879–1890. DOI:10.1017/S0033291711002674

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, *12*, 321–344. DOI:10.1080/15305058.2011.643517

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*, 411–431. DOI: 10.1080/00273170802285743

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 151–173. DOI:10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*, 285–300. DOI:10.1037/a0033266

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72. DOI: 10.1207/s15328007sem1301_3

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694. DOI: 10.2466/pr0.1957.3.3.635

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, *72*, 143. DOI: 10.1037/h0021704

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40. DOI:10.1037//1082-989X.7.1.19

Marcoulides, G. A., & Drezner, Z. (2003). Model specification searches using ant colony optimization algorithms. *Structural Equation Modeling*, *10*, 154–164. DOI: 10.1207/S15328007SEM1001_8

McDonald, R. P. (2013). *Test theory: A unified treatment*. Abingdon, Oxon:Psychology Press.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143DOI:10.1016/0883-0355(89)90002-5

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299. DOI:10.1037/1082-989X.1.3.293

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. DOI:10.1007/BF02294825

Millsap, R. E. (2012). *Statistical Approaches to Measurement Invariance*. Routledge. DOI:10.4324/9780203821961

Molenaar, D., Dolan, C. V., Wicherts, J. M., & van der Maas, H. L. J. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence*, *38*, 611–624. DOI:10.1016/j.intell.2010.09.002

Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. In *Structural equation modeling: A second course(2nd ed.*, pp. 395–436). Charlotte, NC: IAP Information Age Publishing.

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 86–98. DOI: 10.1080/10705511.2012.634724

Moshagen, M., & Auerswald, M. (2018). On congruence and incongruence of measures of fit in structural equation modeling. *Psychological Methods, 23,* 318-336DOI:10.1037/met0000122

Murray, A. L., Booth, T., & Molenaar, D. (2016). Personality differentiation by cognitive ability: An application of the moderated factor model. *Personality and Individual Differences*, *100*, 73–78. DOI:10.1016/j.paid.2016.03.094

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Christensen, G. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425. DOI: 10.1126/science.aab2374

Nye, C. D., Allemand, M., Gosling, S. D., Potter, J., & Roberts, B. W. (2016). Personality trait differences between young and middle-aged adults: Measurement artifacts or actual trends? *Journal of Personality*, *84*, 473–492. DOI:10.1111/jopy.12173

Olaru, G., Schroeders, U., Wilhelm, O., & Ostendorf, F. (in press). "Grandpa, do you like roller coasters?": Identifying age-appropriate personality indicators. *European Journal of Personality*.

Olaru, G., Schroeders, U., Wilhelm, O., & Ostendorf, F. (2018). A confirmatory examination of age-associated personality differences: Deriving age-related measurement-invariant solutions using ant colony optimization. *Journal of Personality 86*, 1037-1049. DOI:10.1111/jopy.12373

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality*, *59*, 56–68. DOI:10.1016/j.jrp.2015.09.001

Orth, U., Robins, R. W., & Widaman, K. F. (2012). Life-span development of self-esteem and its effects on important life outcomes. *Journal of Personality and Social Psychology*, *102*, 1271–1288DOI:10.1037/a0025558

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*, 556–563. DOI:10.1016/S0092-6566(02)00505-6

Preacher, K. J., Wichman, A. L., Briggs, N. E., & MacCallum, R. C. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: SAGE Publications, Inc.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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*, 203–212. DOI10.1016/j.jrp.2006.02.001

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354–373. DOI:doi.org/10.1037/a0029315

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1–25. DOI:10.1037/0033-2909.132.1.1

Robitzsch, A. (2019). sirt: Supplementary item response theory models. R package version 3.1-80. https://CRAN.R-project.org/package=sirt

Yves Rosseel (2012). *lavaan: An R Package for Structural Equation Modeling*. Journal of Statistical Software, 48, 1-36. URL http://www.jstatsoft.org/v48/i02/

Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment*, *63*, 506–516. DOI:10.1207/s15327752jpa6303_8

Schroeders, U., Wilhelm, O., & Olaru, G. (2016a). Meta-Heuristics in short scale construction: Ant Colony Optimization and Genetic Algorithm. *PLOS ONE*, *11*, e0167110. DOI:10.1371/journal.pone.0167110

Schroeders, U., Wilhelm, O., & Olaru, G. (2016b). The influence of item sampling on sex differences in knowledge tests. *Intelligence*, *58*, 22–32. DOI:10.1016/j.intell.2016.06.003

Schultze, M. (2017). *Constructing Subtests Using Ant Colony Optimization* (Dissertation). Freie Universität Berlin. Retrieved from https://www.researchgate.net/publication/326319153_Constructing_Subtests_Using_Ant_Colony_Optimization.

Schultze, M. (2018). *stuart: Subtests Using Algorithmic Rummaging Techniques*. R Package Version 0.7.3). https://CRAN.R-project.org/package=stuart

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software*, *53*, 1-37.. DOI:10.18637/jss.v053.i04

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*.[EPub], New York: Routledge.

Small, B. J., Hertzog, C., Hultsch, D. F., & Dixon, R. A. (2003). Stability and change in adult personality over 6 years: Findings from the victoria longitudinal study. *The Journals of Gerontology: Series B*, *58*, P166–P176. DOI:10.1093/geronb/58.3.P166

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100*, 330-348. DOI: 10.1037/a0021717

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*, 84–90. DOI:10.1016/j.jrp.2008.10.002

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*, 117–143. DOI:10.1037/pspp0000096

Soto, C. J., & John, O. P. (2018). Optimizing the length, width, and balance of a personality scale: How do internal characteristics affect external validity? *Psychological Assessment*. DOI:10.1037/pas0000586

Terracciano, A., McCrae, R. R., Brant, L. J., & Costa, P. T. (2005). Hierarchical linear modeling analyses of NEO-PI-R Scales in the Baltimore longitudinal study of aging. *Psychology and Aging*, *20*, 493–506. DOI:10.1037/0882-7974.20.3.493

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental psychology*, *45*, 1097. DOI: 10.1037/a0015864

Tucker-Drob, E. M., & Salthouse, T. A. (2008). Adult age trends in the relations among cognitive abilities. *Psychology and Aging*, *23*, 453-460. DOI: 10.1037/0882-7974.23.2.453

Vassend, O., & Skrondal, A. (1997). Validation of the NEO Personality Inventory and the five-factor model. Can findings from exploratory and confirmatory factor analysis be reconciled? *European Journal of Personality*, *11*, 147–166. DOI:10.1002/(SICI)1099-0984(199706)11:2<147::AID-PER278>3.0.CO;2-E

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, *29*, 39–47. DOI:10.1111/j.1745-3992.2010.00182.x

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *76*, 913–934. DOI:10.1177/0013164413495237

Wrzus, C., Hänel, M., Wagner, J., & Neyer, F. J. (2013). Social network changes and life events across the life span: A meta-analysis. *Psychological Bulletin*, *139*, 53–80. DOI:10.1037/a0028601

Wrzus, C., & Roberts, B. W. (2017). Processes of personality development in adulthood: The TESSERA framework. *Personality and Social Psychology Review*, *21*, 253–277. DOI:10.1177/1088868316652279

Wrzus, C., Wagner, G. G., & Riediger, M. (2016). Personality-situation transactions from adolescence to old age. *Journal of Personality and Social Psychology*, *110*, 782–799. DOI:10.1037/pspp0000054

Wu, H., & Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*. New York: John Wiley & Sons.Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales

with 200 items. *Journal of Research in Personality*, *44*, 180–198. DOI:10.1016/j.jrp.2010.01.002

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122. DOI: 10.1177/1745691617693393

Yoon, M., & Lai, M. H. (2018). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 201–213. DOI: 10.1080/10705511.2017.1387859

III

A Confirmatory Examination of Age-Associated Personality Differences: Deriving Age-Related Measurement Invariant Solutions using Ant Colony Optimization

Gabriel Olaru[1], Ulrich Schroeders[1], Oliver Wilhelm[2], & Fritz Ostendorf[3]

1: University of Kassel

2: Ulm University

3: Bielefeld University

Status – accepted

Olaru, G., Schroeders, U., Wilhelm, O., & Ostendorf, F. (2018). A Confirmatory Examination of Age-associated Personality Differences: Deriving Age-related Measurement-invariant Solutions using Ant Colony Optimization. *Journal of Personality*, *86,* 1037-1049.

DOI: 10.1111/jopy.12373

Abstract

Objective. The goal of this study was to examine age-associated personality differences using a measurement-invariant representation of the higher-order structure of the Five Factor Model.

Method. We reanalyzed the German NEO-PI-R norm sample (N = 11,724) and applied Ant Colony Optimization in a multi-group confirmatory factor analysis setting in order to select three items per first-order factor that would optimize model fit and measurement invariance across 18 age groups ranging from 16 to 65 years of age.

Results. Ant Colony Optimization substantially improved absolute and relative model fit under measurement invariance constraints. However, the results showed that even when selecting items, measurement invariance across a large age span could not be guaranteed. Strong measurement invariance for *Extraversion* and *Agreeableness* could not be established. The age-associated mean-level differences of the first-order factors of *Neuroticism* and *Conscientiousness* supported the maturity hypothesis. The mean levels of the first-order factors of *Openness* varied substantially from each other across age.

Conclusion. Findings on age differences in personality can be particularly distorted in older age groups. Testing for and ensuring measurement invariance with item selection procedures can help solve this problem. The higher-order structure of personality should be accounted for when personality development is examined.

Keywords: personality development, maturity hypothesis, measurement invariance, item selection

The early debate on personality development focused on the question of whether personality changes occur at any point across the life span. Although personality was originally perceived to be carved in stone (Costa & McCrae, 1994), a lot of contradictory evidence has called into question the stability perspective (e.g., Helson, Jones, & Kwan, 2002; Mroczek & Spiro, 2003; Srivastava, John, Gosling, & Potter, 2003). In an ambitious attempt to summarize the findings of several meta-analyses on personality change, Roberts, Walton, and Viechtbauer (2006) analyzed 92 samples in a comprehensive meta-analysis and found substantial age-associated differences in personality at the mean level. Not only was personality found to change in early adolescence, but in contrast to prior expectations, the development also carried on through adulthood (Costa & McCrae, 1997; McCrae & Costa, 1999, 2002). However, the large majority of research on personality development has relied on sum scores, and the impact of measurement bias with respect to age is hence unclear. In addition, information is generally examined at the level of the broad personality factors instead of the more fine-grained first-order factors. Both assessment and methodological shortcomings call into question the validity of these findings.

In the present study, we examined differences in personality across age groups based on a latent representation of personality that maximizes measurement invariance across age. First, we discuss the concept of age differences in personality, and from a methodological perspective, we highlight how age differences might affect different aspects of the structure of personality. Second, we present an overview of the findings on personality change across age. The goal of the present study was to examine age differences in personality after maximizing model fit and measurement invariance. We then compare our findings on age-associated differences in the mean and covariance structure with the existing literature.

The Five Factor Model of Personality

The Five-Factor Model (FFM) is a widely accepted and well-evaluated model of personality that captures the personality traits of *Neuroticism*, *Extraversion*, *Openness*, *Agreeableness,* and *Conscientiousness*. The FFM is a parsimonious and relatively exhaustive representation of personality. As such, the traits represent very broad trait domains and should be seen as overarching second-order factors atop more specific first-order factors, often referred to as facets (e.g., Costa & McCrae, 1995; John & Srivastava, 1999; Soto & John, 2009). Assessing these first-order factors can capture personality with higher fidelity than scales that are based on the broad second-order factors (Ashton, Jackson, Paunonen, Helmes, & Rothstein, 1995). In addition, the first-order factors can provide a stronger prediction of various outcomes than the second-order factors (Ashton, 1998). However, whereas consensus on the five second-order factors of personality is relatively high, no commonly accepted first-order factor structure has been identified. Therefore, the first-order factors and hence the composition of the second-order factors can vary to a substantial degree across measurement instruments. In order to fully understand the meaning of the FFM, it is thus critically important to consider the underlying first-order factor structure.

Age Differences in Personality

From a psychometric point of view, age-related personality differences can manifest in different aspects of the model, for example, in the factor means or the factor loadings (see Figure 1). Most studies on personality development have focused on normative trends but have often overlooked structural age differences. Structural age differences can be examined by comparing the covariance between a set of parameters across age groups, for instance, the relationships between latent variables (e.g., factor correlations) or items (e.g., factor loadings).

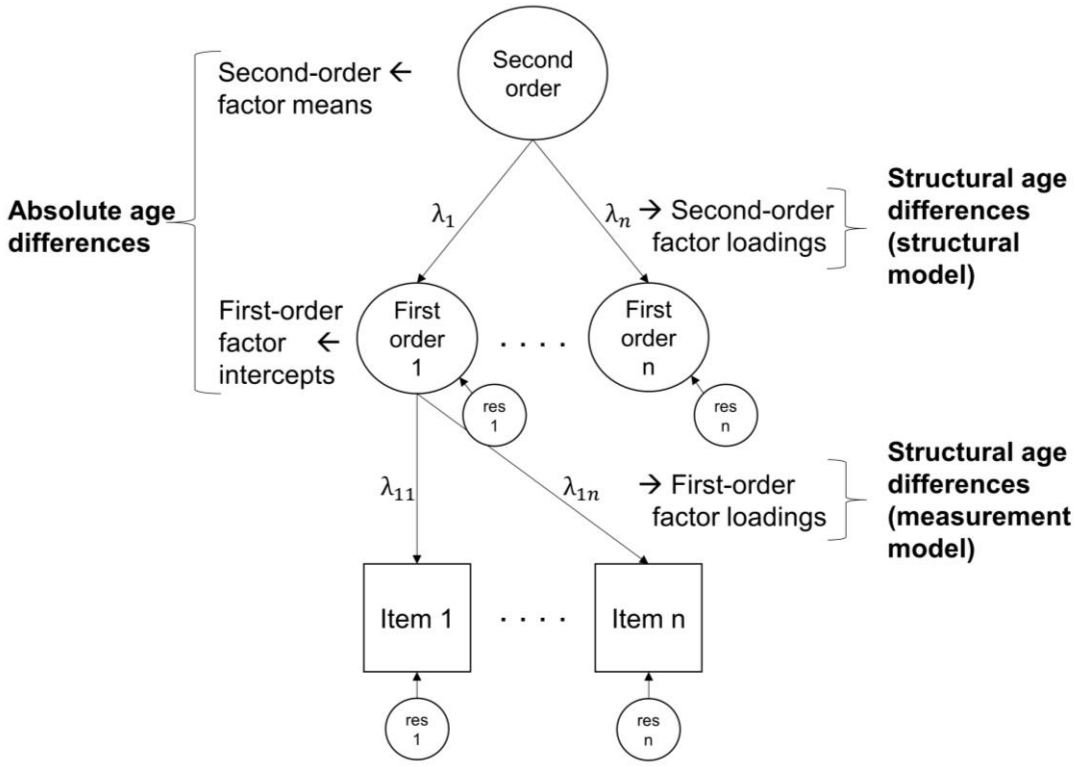


Figure 1. Age-associated differences in a higher-order factor model. Latent variables are depicted as circles. Manifest variables are represented by squares. Res = Residual variance.

Structural equality across age groups ensures that the model parameters are unbiased and the derived factors comparable across groups. Therefore, examinations of other types of age differences at the factor level are meaningful only when structural equality is established across age. Structural differences have been examined in a cross-sectional context predominately in the framework of measurement invariance testing by means of multigroup confirmatory factor analysis (MGCFA; e.g., Allemand, Zimprich, & Hertzog, 2007; Nye, Allemand, Gosling, Potter, & Roberts, 2015). Depending on the measurement invariance levels or, in other words, the parameters set to equality across groups, different comparisons can be made across groups (Meredith, 1993). Measurement invariance is usually tested by sequentially imposing stricter model constraints (e.g., equal factor loadings, item intercepts, or item residuals) across groups and examining the impact of these additional constraints on the model fit. Depending on the most stringent level of measurement invariance achieved, different aspects of the model can be compared across groups. For instance, weak invariance

(equal factor loadings across groups) is necessary for comparisons of factor correlations or for comparisons of correlations with covariates across groups. Strong measurement invariance (equal factor loadings and item intercepts across groups) is required to meaningfully compare factor means across groups. In a higher-order model, the constraints are first applied to the measurement model and then to the structural model (Chen, Sousa, & West, 2005). The levels of measurement invariance that we tested in this study and the corresponding types of age differences are presented in Table 1.

Table 1. *Steps of Invariance Testing*

Model	Invariance level	Type of age differences examined
1	Configural invariance	-
2	First-order factor loadings invariant	Structural differences (first-order factors)
3	First-order factor loadings and item intercepts invariant	Absolute differences (first-order factors)
4	First- and second-order factor loadings and item intercepts invariant	Structural differences (second-order factor)
5	First- and second-order factor loadings, item intercepts, and first-order factor intercepts invariant	Absolute differences (second-order factor)

Empirical Findings on Age Differences in Personality

With respect to absolute age differences, many cross-sectional and longitudinal studies have examined mean-level differences in personality across age. Cross-sectional studies typically provide larger sample sizes and can cover broader age spans, but can suffer from cohort effects. While longitudinal studies are less prone to such effects when examining developmental processes, but are also more expensive and suffer from attrition. A consideration of both these approaches is hence desirable for the examination of personality development across age. Overall, *Neuroticism* has been found to decline across the life course, whereas *Agreeableness* and *Conscientiousness* have been found to increase in cross-

sectional (McCrae et al., 1999; Soto, John, Gosling, & Potter, 2011; Srivastava, John, Gosling, & Potter, 2003) and longitudinal studies (Roberts et al., 2006; Terracciano, McCrae, Brant, & Costa, 2005). These normative differences across age have been associated with a maturation process that has been observed in the entire population (e.g., Helson & Moane, 1987; McCrae et al., 2000) and seems to be independent of sex (Helson, Jones, & Kwan, 2002; Roberts et al., 2006) and culture (McCrae et al., 1999). Findings on *Conscientiousness*, *Agreeableness,* and *Neuroticism* have been relatively consistent, but findings on *Extraversion* and *Openness* have tended to differ greatly between studies (e.g., Allemand, Hertzog, & Zimprich, 2007; McCrae et al., 1999; Roberts et al., 2006; Terracciano, McCrae, Brant, & Costa, 2005). Moreover, longitudinal studies that have examined age differences at the first-order factor level have found that aspects of *Extraversion* that are related to social dominance increase with age, whereas others such as Gregariousness or Excitement Seeking decline (Roberts et al., 2006; Terracciano et al., 2005). Similarly, some first-order factors of *Openness* have not been found to decrease with age (Terracciano et al., 2005) but might follow a curvilinear trajectory with a maximum in middle adulthood (Roberts, Walton, & Viechtbauer, 2006). Such first-order factor-level findings emphasize the importance of examining age-related personality differences on a more fine-grained level rather than the level of second-order factors alone. This is even more important because, on a more general stance, the findings on a trait level depend on the composition of the trait itself. In other words, the personality factors assessed with different measurement instruments might differ substantially in their composition on the first-order factor level. However, only when the underlying first-order factors and hence second-order factors are similar, meaningful comparisons across measures and studies are feasible.

The structural stability of personality across age has been examined in a number of cross-sectional (Allemand, Zimprich, & Hendriks, 2008; Allemand, Zimprich, & Hertzog,

2007; Nye et al., 2015) and longitudinal studies (Allemand et al., 2007; Robins et al., 2001; Small, Hertzog, Hultsch, & Dixon, 2003). Structural stability was usually examined by testing for measurement invariance across age groups. However, because the underlying first-order factor level was not assessed in these studies, the question of whether the internal structure of the personality traits differs across age groups was not addressed. Examining structural stability in a hierarchical model can help determine whether the measured constructs differ across age as a result of, for instance, changes in the factor loadings. Structural stability has received less attention than normative age differences in the personality development literature. This is unfortunate because an invariant measurement model over time or age is a necessary prerequisite for a meaningful interpretation of mean-level differences in personality across age.

The Issue of Model Fit

The current understanding in personality assessment is that the indicators constitute valid representations of the second-order factors that they are intended to measure. It is assumed that these indicators represent the unidimensional latent trait in question, which is a measurement by fiat. However, when these broad personality measures have been evaluated with confirmatory factor analyses, these models have been found to have inadequate model fit to data (Borkenau & Ostendorf, 1990; Church & Burke, 1994; McCrae, et al., 1996; Vassend & Skrondal, 1997). Accordingly, there tends to be a discrepancy between theoretical assumptions about personality and the empirical underpinning. The fit of broad personality models tends to suffer from violations to the simple structure and high correlated item uniqueness, which is often found in multidimensional self-report inventories. However, the validity of the model is a prerequisite to the interpretation of personality scores as representations of the underlying traits (Borsboom, 2006).

A common solution to the lack of fit in personality scales is to estimate the models on the basis of aggregates of the manifest indicators, so-called parcels (Little et al., 2002). However, in the present case, we decided not to use parceling because the method is applicable only when the scale is unidimensional (Bandalos & Finney, 2001; Little, Cunningham, Shahar, & Widaman, 2002; Marsh, Lüdtke, Nagengast, & Morin, 2013) and the indicators used for parceling are measurement invariant (Little, Rhemtulla, Gibson, & Schoemann, 2013; Meade & Kroustalis, 2006).

Item Selection with Ant Colony Optimization

In the present article, we present a recently proposed method (i.e., Ant Colony Optimization; ACO) to identify and remove problematic items that negatively impact model fit. This item selection procedure can be applied to improve both the absolute model fit and the relative model fit between different levels of measurement invariance (Schroeders, Wilhelm, & Olaru, 2016a). The usefulness of item selection for improving model fit has been demonstrated in a number of recent articles on short-scale construction (Janssen, Schultze, & Groetsch, 2015; Olaru, Witthoeft, & Wilhelm, 2015; Schroeders, Wilhelm, & Olaru, 2016a, 2016b). Finding a measurement model that is unidimensional and measurement invariant across age can be seen as a combinatorial problem in which one aims to identify the best item set out of a multitude of possible item sets. For comprehensive instruments such as the NEO-PI-R, which captures 30 first-order factors of the FFM, the outcome space for the smallest possible model that can be applied to capture each second-order factor with three manifest indicators per first-order factor contains roughly $3 * 10^{52}$ possible item combinations. Obviously, applying an exhaustive search to explore the entire outcome space for the best solution is not computationally feasible. We thus applied a meta-heuristic algorithm called *Ant Colony Optimization* (ACO; Leite, Huang, & Marcoulides, 2008; Marcoulides & Drezner, 2003), which is capable of tackling this issue by finding an optimal (or close-to-

optimal) solution in an iterative process. ACO algorithms were originally based on the foraging behavior of ants (Deneubourg, Aron, Goss, & Pasteels, 1990) and have been applied to derive optimized short scales ((Janssen, Schultze, & Grötsch, 2015; Leite et al., 2008; Olaru et al., 2015). Similar to the way in which ants use pheromones to attract other ants to the shortest route to a food source, ACO uses virtual pheromones to increase the attractiveness of items that yield better psychometric properties (e.g., model fit). Initially, ants will randomly explore the space between their nest and the food source. During their search, each ant leaves a pheromone trail. On shorter routes, more pheromones accumulate in a given time frame. Higher levels of pheromones attract more ants, and this in turn increases pheromone levels further until almost all ants will follow the shortest route. In our context, item sets correspond to the possible routes an ant can take. In the first iteration, a number of randomly selected item sets (i.e., ants) are compared with regard to a predefined criterion such as model fit (i.e., length of the route between the nest and the food source). This optimization criterion can also be a combination of several psychometric criteria (e.g., model fit and reliability). Similar to pheromones accumulating on shorter routes, virtual pheromones for items in the item subset with the best criterion value increase at the end of each iteration. A higher concentration of pheromones increases the probability of items being drawn from the item pool in the upcoming iterations, similar to ants preferring routes with higher pheromone levels. This procedure is repeated until the procedure cannot further improve the criterion over a number of iterations or a predefined criterion value is reached. Note that as ACO is a probabilistic approach, it will not necessarily find the optimal solution. Users are advised to compare results across several runs of the algorithm in order to ensure that an optimal solution is found (Dorigo & Stützle, 2010; Leite, Huang, & Marcolides, 2008).

Using ACO to construct short scales has a number of advantages over classical procedures: First, the selection algorithm is not tied to any statistical method and can be

adapted to solve every combinatorial problem with a quantitative outcome. Second, instead of removing items sequentially, ACO searches for item samples of a fixed size. Thus, it is not affected by sequence effects in the selection of items. Third, in large outcome spaces, the probabilistic approach of ACO is computationally much less demanding than exhaustive algorithms. And finally, ACO can be used to maximize several criteria simultaneously, for example, reliability and model fit (see also Janssen, Schultze, & Grötsch, 2015; Schroeders et al., 2016a, 2016b).

Research Aims

The main goal of this article was to examine age-related personality differences in a measurement-invariant higher-order model of the FFM factors and to detect sources of misfit when applying increasingly strict constraints. The prerequisite to examining normative and structural age differences is an age-invariant measurement model (i.e., invariant item intercepts and factor loadings on the first-order factor level). To achieve this overall goal of studying age-associated personality differences, we demonstrate (a) how to achieve measurement invariance using ant colony optimization while retaining decent model fit, (b) the importance of measuring personality at the first-order factor level, and (c) how the lack of measurement invariance biases the results commonly reported in the literature on personality development.

Method

Sample

The sample consisted of 11,724 participants (7,505 female), representing the nonclinical part of the German Revised NEO-Personality Inventory validation sample (Ostendorf & Angleitner, 2004). Compared with the general population, the given convenience sample was well-educated such that 67.0% of the sample had obtained or had worked toward a high school degree that qualified them to attend a university (compared with

19.3% in the population). A large portion (40.2%) were university students, and 13.3% of the sample had at least a college degree. Age ranged from 16 to 91 years with a mean of 29.9 years (*SD* = 12.1). In order to assess measurement invariance for the multi-group analyses, the sample was split into 18 age groups with about 500 participants in each. There were not enough participants older than 65, and thus, these 168 participants were dropped. Age ranges, sample sizes, and sex ratios for the 18 groups are presented in Table 2.

Table 2. *Sample Characteristics Across Age Groups*

Group	Age	*N*	% Female
1	16-18	524	74
2	19	670	80
3	20	962	78
4	21	955	64
5	22	875	60
6	23	757	54
7	24	696	55
8	25	572	53
9	26	477	56
10	27-28	771	57
11	29-30	622	60
12	31-32	476	62
13	33-35	529	62
14	36-39	540	65
15	40-43	484	74
16	44-48	577	68
17	49-54	530	66
18	55-65	529	63

Measures

Personality was measured with the German adaptation of the NEO-PI-R (Ostendorf & Angleitner, 2004), which measures the FFM second-order factors and the underlying first-order factors. Every second-order factor (e.g., *Neuroticism*) is assessed with six first-order factors (e.g., Anxiety, Anger Hostility, Depression, Self-Conscientiousness, Impulsivity, Vulnerability), measured with eight items each, resulting in a total of 240 items. Participants indicated their agreement with the statements on a 5-point scale, with the response options 4 (*strongly agree*), 3 (*agree*), 2 (*neutral*), 1 (*disagree*), and 0 (*strongly disagree*).

Statistical Analysis

Model specification.

Every second-order factor was examined separately. We specified a higher-order factor model with items loading on their corresponding first-order factor (e.g., Competence) and the six first-order factors loading on a second-order factor (e.g., *Conscientiousness*). The model was identified by constraining the unstandardized loading of the first indicator on each factor to 1. The number of items per first-order factor to be selected was set to three out of an available eight. Three was chosen as the minimum because this number ensures that the local parts of models will be just identified. In our data not all five response categories were used in all item in every age group. As a result, some thresholds could not be computed with a model estimator specifically designed for categorical responses. Hence, all models were estimated with a Maximum Likelihood estimator assuming a continuous response format, which is a common and appropriate approach to model categorical data with at least five response categories (Beauducel & Herzberg, 2006; Rhemtulla, Brosseau-Liard, & Savalei, 2012).

Measurement invariance.

On the basis of our research questions, we specified several levels of invariance for the hierarchical model by constraining certain groups of measurement parameters (cf. Chen, Sousa, & West, 2005). These constraints and the types of change that can be examined with each subsequent step of measurement invariance testing are listed in Table 1. The first age group was specified as the reference group by setting the factor means to zero.

Model evaluation.

Using common standards (Hu & Bentler, 1999; Marsh, Hau, & Grayson, 2005), we evaluated model fit with the *Comparative Fit Index* (CFI) and the *Root Mean Square Error of Approximation* (RMSEA). We use statistics based on the χ^2 with care because they are

sensitive to sample size (Bentler & Bonett, 1980). Based on the measurement invariance literature, we determined the level of invariance with (a) ΔCFI (Cheung & Rensvold, 2002) and (b) the 90% confidence intervals for the RMSEA (cf. Allemand, Hertzog, & Zimprich, 2007; MacCallum, Browne, & Sugawara, 1996) of two consecutive models. For a model to be accepted, it had to yield acceptable absolute fit (CFI > .90 and RMSEA < .08) and acceptable relative fit compared with the preceding step of invariance testing (ΔCFI < .01 and overlapping 90% confidence intervals for the RMSEA). Models were estimated in *Mplus* 7 (Muthén & Muthén, 2004-2017).

Item selection.

We applied ACO to select three indicators per first-order factor so that two criteria would be maximized: First, the overall fit of a unidimensional model needed to be good and, second, the relative model fit between consecutive models in measurement invariance testing across age needed to indicate no deterioration. Fit measures were computed for a model with strong measurement invariance (see Model 3 in Table 1) because this is the prerequisite for the further examination of age differences in personality. The optimization function included the CFI and RMSEA (overall fit) as well as the ΔCFI between the two decisive steps of invariance testing (i.e., Model 3 and Model 2 in Table 1). All three criteria were logit-transformed in order to scale the range of the values between 0 and 1 and maximize the differentiation around the critical cutoff values (see also Janssen, Schultze, & Groetsch, 2015; Schroeders et al., 2016a, 2016b). The logit-transformation ensures that the different optimization criteria are scaled on a common metric and thus weighted equally in the overall optimization function. By exploring different cutoff values, the CFI and RMSEA were transformed as follows, indicating acceptable to good model fit for CFI > .91 and RMSEA < .06 (see Equations 1 and 2).

$$\varphi_{CFI} = \frac{1}{1+e^{91-100CFI}} \quad (1)$$

$$\varphi_{RMSEA} = 1 - \frac{1}{1+e^{6-100RMSEA}} \quad (2)$$

ΔCFI was transformed with a cutoff at ΔCFI = .01, with lower values indicating an acceptable increase in model misfit between invariance levels. To further emphasize the differentiation around the critical cutoff, the value was additionally squared (see Equation 3).

$$\varphi_{\Delta CFI} = 1 - \left(\frac{1}{1+e^{1-100\Delta CFI}}\right)^2 \quad (3)$$

The overall optimization function was the sum of these three criteria. ACO was implemented in R (R core team, 2014).

Examination of age-related differences.

On the basis of the model identified by ACO, age differences were examined across age groups. If adequate measurement invariance levels were achieved (see Models 3 and Model 5 in Table 1), absolute age differences were analyzed by comparing factor means across age groups. Structural age differences were examined by testing invariance constraints on the first- or second-order factor loadings (see Models 2 and Model 4 in Table 1). If these invariance tests failed, we could assume that considerable structural age differences were present.

Results

Absolute Model Fit and Factor Saturation

The single FFM factor models based on the original scales (i.e., eight items per first-order factor) yielded an insufficient CFI ranging from .693 for *Agreeableness* to .768 for *Neuroticism* and an acceptable RMSEA ranging from .051 for *Openness* to .060 for *Extraversion*. The discrepancy between the RMSEA and CFI can be explained by the low complexity of the models and the large sample size, resulting in better RMSEA values. Another possible explanation is that the CFI is generally lower when loadings are smaller, as the null model will fit better in that case (Moshagen & Auerswald, 2017). This is supported

by the relative low factor saturation ω (McDonald, 1999). For the full scales, factor saturation ranges from ω = .48 to ω = .83 across all 30 first-order factors (M = .70; Median = .71; see Online Supplement Table 1 for full table).

ACO succeeded in identifying item subsets with acceptable absolute model fit for all second-order factors (see Table 3; see Online Supplement Table 2 for a list of the selected items). In order to examine the extent to which ACO was able to improve model fit, we compared our results to the distribution of model fit indices based on chance. To do so, we randomly selected 1,000 18-item models for *Neuroticism* and computed fit indices for the most critical invariance levels (Model 1, Model 2 and Model 3). The 99th percentile of CFI ($\mathrm{CFI_{M1}} = .928;\ \mathrm{CFI_{M2}} = .925;\ \mathrm{CFI_{M3}} = .906$) and 1st percentile of RMSEA ($\mathrm{RMSEA_{M1}} = .044;\ \mathrm{RMSEA_{M2}} = .043;\ \mathrm{RMSEA_{M3}} = .048$) of the randomly selected models were inferior to the model fit of the ACO item selection (see Table 3). However, factor saturation of the first-order factors decreased as a result of the reduced item numbers (ω range for ACO selection: .37 - .79; M = .59; Median = .59; see Online Supplement Table 1 for full table).

Measurement Invariance

Because of the low overall fit of the full 48-item models, all further examinations of measurement invariance based on these models would be invalid. Taken together, no aspect of continuity can be studied on the basis of the complete models with 48 items per second-order factors. Model fit estimates for the ACO models with increasing measurement invariance constraints are presented in Table 3. According to the relative fit between consecutive models, different levels of measurement invariance were reached for the factors. The ΔCFI criterion was most sensitive, showing only weak measurement invariance (Model 2) for *Extraversion* and *Agreeableness*. The RMSEA confidence intervals supported strong invariance across age groups for all five models (Model 3). Both criteria supported invariant second-order factor loadings across age groups for *Neuroticism*, *Openness*, and

Conscientiousness. In other words, structural continuity was supported for these second-order factors. *Conscientiousness* was the only second-order factor that yielded acceptable fit indices for both the absolute and relative tests of model fit for the most restrictive model. Again, we compared the relative fit of our solution to a selection by chance. The 1st percentile of ΔCFI across the randomly selected 18-item *Neuroticism* models with acceptable absolute model fit (Model 3) was higher than the critical cut-off for strong measurement invariance ($\Delta\mathrm{CFI}_{\mathrm{M2-M1}} = .001,\ \Delta\mathrm{CFI}_{\mathrm{M3-M2}} = .011$). In comparison, ACO was able to establish strong measurement invariance for this factor.

To investigate the source of model misfit between weak (Model 2) and strong measurement invariance (Model 3), we examined the modification indices of the models. Additionally, constraining the item intercepts to equality across age groups created the largest misfit for the oldest age groups (i.e., 49 to 54, 55 to 65 years). This pattern was *cum grano salis* consistent across all five traits. When removing the two oldest age groups, strong invariance could also be established for *Extraversion* and *Agreeableness* across the reduced age span.

Table 3. *Measurement Invariance Across 18 Age Groups*

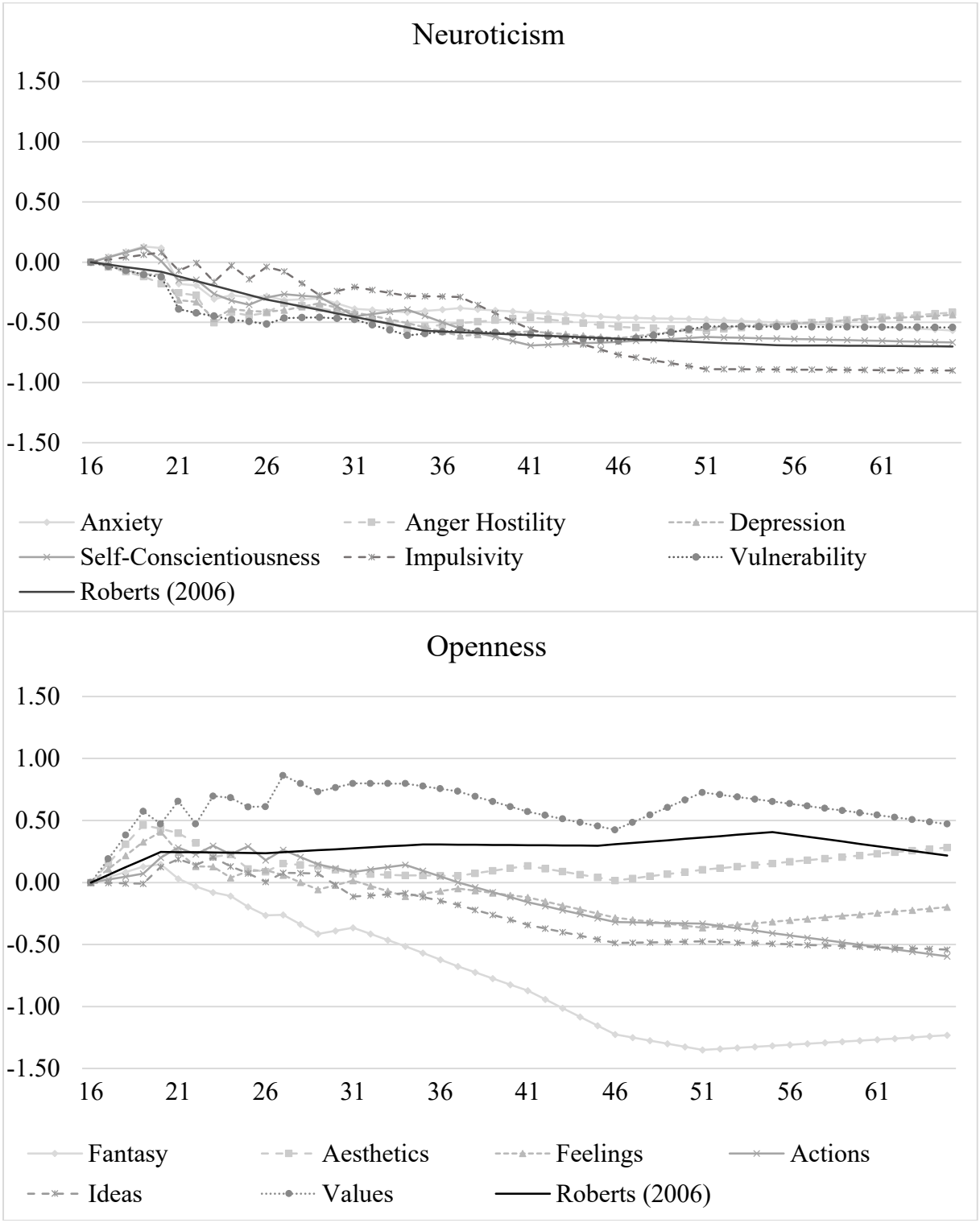
		Neuroticism			Extraversion			Openness			Agreeableness			Conscientiousness		
Mo.	*df*	χ^2	CFI	RMSEA	χ^2	CFI	RMSEA	χ^2	CFI	RMSEA	χ^2	CFI	RMSEA	χ^2	CFI	RMSEA
1	2322	4789	.949	.041 [.033,.036]	4895	.931	.042 [.040,.043]	4561	.938	.039 [.037,.040]	4593	.930	.039 [.037,.041]	-	-	-
2	2526	5075	.947	.040 [.038,.041]	5207	.928	.041 [.039,.042]	4882	.935	.038 [.037,.040]	4887	.927	.038 [.037,.040]	5765	.931	.045 [.043,.046]
3	2730	5693	.939	.041 [.040,.043]	5949	.914	.043 [.041,.044]	5318	.929	.038 [.037,.040]	5486	.915	.040 [.038,.041]	6389	.922	.046 [.044,.047]
4	2815	5842	.938	.041 [.039,.042]	6067	.913	.042 [.041,.044]	5457	.927	.038 [.037,.040]	5637	.912	.040 [.038,.041]	6506	.921	.045 [.044,.047]
5	2900	6498	.926	.044 [.043,.045]	8410	.853	.054 [.053,.056]	6894	.890	.046 [.045,.048]	6155	.899	.042 [.040,.043]	6823	.916	.046 [.044,.047]

Note. Mo. = Model; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation (RMSEA); 90% Confidence Interval (CI) of the RMSEA is printed below the RMSEA value in brackets; Achieved levels of invariance are underlined; Model 1: Configural invariance; Model 2 = First-order factor loadings invariant; Model 3 = First-order factor loadings and item intercepts invariant; Model 4 = First, and second-order factor loadings and item intercepts invariant; Model 5 = First, and second-order factor loadings, item intercepts, and first-order factor intercepts invariant; Model 1 for *Conscientiousness* did not converge.

Absolute Age Differences

The age-associated normative levels for all models with strong measurement invariance across the entire age span are presented in Figure 2. The mean level trajectories reported in the comprehensive meta-analysis by Roberts and colleagues (2006) are presented as a reference line. In the following, we will summarize the most relevant findings. With respect to *Neuroticism*, the mean level decreased with age. The age-associated mean-level differences found in the current study were similar to the findings reported by Roberts et al. (2006). The strongest deviation from the overall trend described by Roberts et al. (2006) was found for the first-order factor Impulsivity. When the second-order factor loadings were fixed to equality across age groups, Impulsivity had by far the weakest correlation with *Neuroticism* ($\lambda = .30$ compared with an average $\lambda = .86$ across the other first-order factors). This indicates that Impulsivity has a high uniqueness compared with the other first-order factors of *Neuroticism*. The age-associated mean-level differences for *Openness* were less consistent across first-order factors, advocating for the heterogeneity of the construct. The mean levels of most first-order factors of *Openness* increased early in the life span, began to decrease at the age of 20, and continued to decrease across the life span. However, there were large differences in the normative levels of the first-order factors across age. For example, Openness to Fantasy decreased by 1.30 *SD* across the age range, whereas the effects on the other first-order factors were small (absolute $d = 0.20$) to medium (absolute $d = 0.60$). Finally, invariance testing allowed us to examine age differences at the second-order factor level of *Conscientiousness*. The age-associated mean-level differences found in this study were similar to the trend described in the literature (Roberts et al., 2006). Note that age-associated increases in the first-order factors Achievement Striving and Deliberation were substantially lower than for the other first-order factors or the general *Conscientiousness*

second-order factor. Mean-level differences across age for *Extraversion* and *Agreeableness* across the reduced age span are presented in the Online Supplement Figure 1.



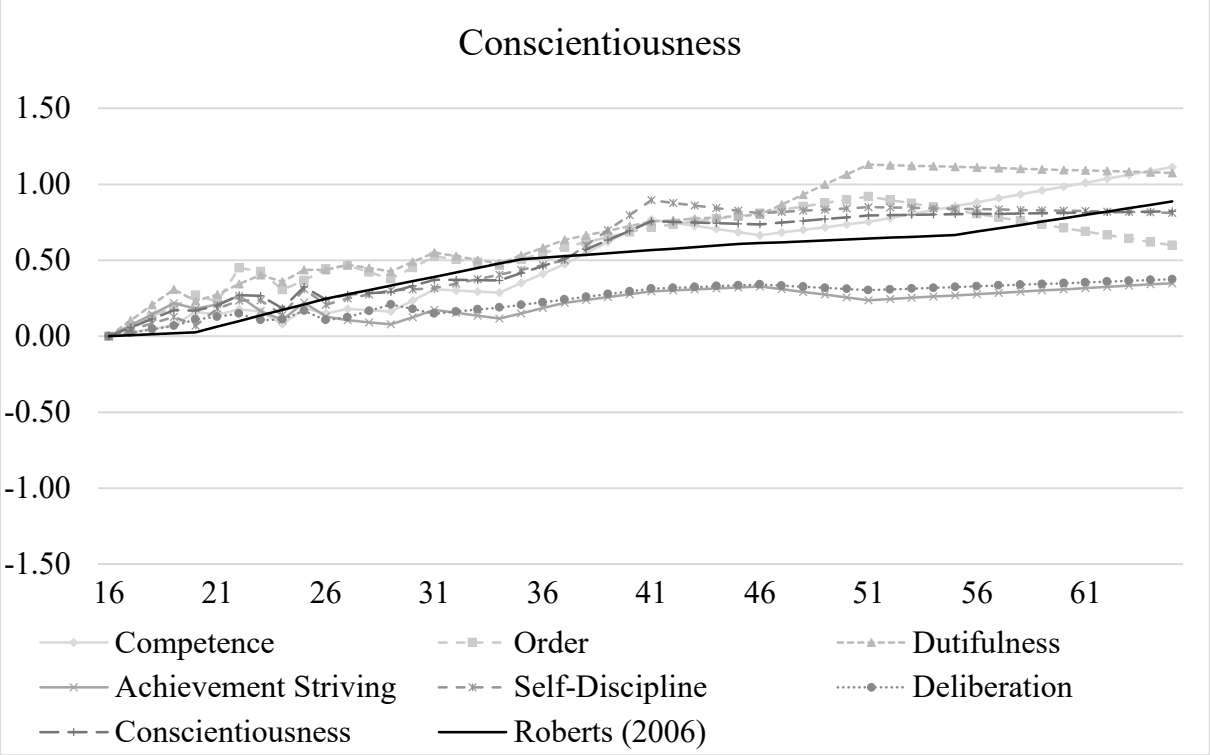


Figure 2. Mean levels of first- and second-order factors across age. Values on the y-axis are standardized factor scores, which can be interpreted as standardized z-values.

In order to examine the importance of establishing measurement invariance in personality development studies, we compared manifest scores to the derived factor scores. The increasing misfit in older age groups distorted the examination of absolute age differences substantially. Differences between the mean levels of the latent factors and manifest scores (transformed into the effect size Hedges' g) increased with age up to a range from d = -.60 to d = .95. In general, the manifest scores seemed to systematically underestimated absolute age differences, in particular for first-order factors with large mean-level differences across age (for a comparison of latent and manifest scores across age see Online Supplement Figure 2).

Discussion

In this article, we investigated age differences in personality across an age range of 50 years. We were seeking measurement models of personality that met strict standards of measurement invariance. One motivation to pursue this goal was to establish a psychometrically sound basis for age comparisons in personality dispositions. Where

applicable, we examined aspects of age-associated differences in personality and related the results to previous findings. Next, we will first discuss the results on age-associated personality differences. Subsequently, we will address methodological concerns regarding the use of item selection procedures in examining personality development on a more general stance. Finally, we will discuss implications of our findings for research on personality development and provide recommendations for future studies.

Age-associated Personality Differences

This study replicated previous findings on age-related differences at the normative level of personality. More precisely, the maturity hypothesis was supported by increasing levels of *Conscientiousness* over age and a decline in *Neuroticism*, which strongly resemble the findings reported by Roberts and colleagues (2006). Moreover, all first-order factors of *Neuroticism* and *Conscientiousness* were affected by the maturation process. The age-associated mean-level differences of the *Openness* first-order factors varied substantially from each other with respect to magnitude and direction. Findings on age-related changes of *Openness* are much less consistent across studies as for instance on *Neuroticism* and *Conscientiousness*. Depending on the measurement instrument and the first-order factors of *Openness* considered, the results may vary substantially. While the mean-levels of Openness to Values and Openness to Aesthetics resemble the curve-linear trajectory reported by Roberts and colleagues (2006), the other first-order factors decrease across the examined age range, as has been found in other studies too (McCrae et al., 1999, Terracciano, McCrae, Brant, & Costa, 2005). Depending on the mix of first-order factors that constitute the second-order factor, the findings in the literature vary. The issue of first-order factor-compilation is also mirrored in the age-associated mean-level differences of *Extraversion* and *Agreeableness* (see Online Supplement Figure 1). There is high consensus in the research literature concerning the development of *Agreeableness* over age. Findings concerning age-trajectories

are much less stringent for *Extraversion* (McCrae et al., 1999; Soto, John, Gosling, & Potter, 2011; Srivastava, John, Gosling, & Potter, 2003; Terracciano, McCrae, Brant, & Costa, 2005; Roberts et al., 2006). Correspondingly, in this study age-associated mean-level differences of *Agreeableness* across the reduced age span are consistent across first-order factors, whereas the age-associated normative differences of the *Extraversion* first-order factors differ substantially from each other in direction and magnitude.

In addition to providing a more precise measurement of personality over age, capturing personality traits in a higher-order factor model allows for a deeper examination of structural changes in personality. In a higher-order factor model, the centrality of the first-order factors for the overarching second-order factor can be examined by scrutinizing the second-order factor loadings. The higher the second-order loading, the more relevant or central the first-order factor is to that second-order factor. Examining whether this relationship changes across the life span can provide additional insights into developmental processes. For instance, Excitement Seeking might be a central first-order factor of *Extraversion* in young age, but decrease in relevance as a first-order factor of *Extraversion* in later years. Similarly, *Extraversion* might be defined better by Gregariousness and Assertiveness during young adulthood, and by Positivity and Warmth in old age. In order to examine such structural differences of the FFM across 50 years of age, we constrained the second-order factor loadings to equality across age groups. This level of invariance was supported for *Neuroticism*, *Conscientiousness*, and *Openness*, that is all second-order factors with invariant measurement level. Accordingly, for these FFM factors we can infer that the relevance of first-order factors in a common second-order factor does not change across age.

Methodological Approaches for Establishing Measurement Invariance

Removing non-invariant items to compare groups in MGCFA has been criticized before (Cheung & Rensvold, 1999), as the construct coverage might be narrowed. However,

the overall fit for the full 48-item models does not support the long form as an adequate representation of the latent traits. We decided that sampling from the original set of manifest variables might result in a sacrifice of construct coverage, but it is helpful when the model fit is deficient and indicators are not measurement invariant across age, which we deem pivotal. We retained the first-order factor structure of the personality model in order to reduce the impact of the item selection on construct coverage. ACO substantially increased both the absolute and relative model fit and resulted in an acceptable level of model fit for all traits at the configural invariance level. As shortening a scale will usually decrease reliability (Spearman, 1910), we suggest also including reliability as an optimization criterion in cases where the starting model fit is not as problematic as in the present study. The item selection procedure used in this study was not applied to present an ideal subset of items but to solve the specific problem of questionable unidimensionality of the traits as well as age-associated measurement invariance. In other contexts, a different set of selection criteria might be more appropriate (for an overview of possible selection criteria, see Janssen, Schultze, & Groetsch, 2015; Leite, Huang, & Marcoulides, 2008; Schroeders et al., 2016a, 2016b).

One goal of this article was to present a psychometrically sound alternative to the commonly applied parceling procedure when examining personality differences in MGCFA. It should be mentioned that a number of more sophisticated procedures with less strict testing assumptions have been suggested as other alternatives, such as Exploratory Structural Equation Modeling (ESEM; Asparouhov & Muthén, 2009). ESEM combines characteristics of CFA and EFA by allowing all cross-loadings while still providing model fit indices. ESEM has been suggested to be a viable alternative to CFA in the domain of personality, as personality models typically suffer from high misfit due to cross-loadings (Marsh et al., 2010). However, higher-order models cannot be estimated in ESEM (see Marsh et al., 2009), and these played an important role in our structural age-difference argument. Second, using

ESEM in a multi-group context is problematic, as cross-loadings have to be constrained to equality as well. In comparison, our item selection procedure was able to substantially eliminate cross-loadings and correlated uniqueness, as well as maximize measurement invariance. Less strict alternative approaches to classical invariance testing have also been suggested, most notably the so-called Alignment method (Asparouhov & Muthén, 2014). Instead of forcing measurement invariance on every parameter, the Alignment technique rotates the violations of invariance across a few items and tries to minimize it on others (similar to EFA rotation). The rotation does not affect the model fit, which remains the level of the configural invariance model. Alignment seems like a reasonable alternative to overly strict measurement invariance constraints, but it still requires a plausible model for configural invariance. The full models in this study did not fulfill this requirement, and plausible and adequate models could be established only after item selection.

Implications for Personality Development Research

Examining latent representations of personality rather than manifest aggregated scores is still the exception in the personality development literature. Personality researchers tend to neglect the issue of measurement invariance across the hierarchyof personality models. In the construction of well-known personality measures confirmatory model fit and age-associated measurement invariance were not central aspects of the development process, although it is of crucial importance when studying age-associated personality differences. If measurement invariance is not given, the derived scale scores may be distorted due to method artifacts. For instance, comparing an adolescent to an older adult person based on age-variant items such as “Do you like to go bungee jumping?” is evidently problematic, because this item of sensation seeking hinges upon the physical health of participants. Selecting measurement-invariant items as we did in this study is an appropriate way to reduce age-associated heterogeneity and allows for psychometrically justified age comparisons. Unfortunately, we were not able to

find a measurement-invariant solution across the entire age range for *Extraversion* and *Agreeableness*. While this finding is problematic from an assessment perspective, the lack of measurement invariance can also indicate developmental changes in personality that go beyond simple changes in factor or scale means. For example, one might find that the ideal item sets for measuring the personality traits is different at various stages in life, which implies that the way personality manifests at different stages across life changes. For instance, participating in social events may be a representative indicator of high Gregariousness for adolescents, whereas a good indicator for elderly people might rely on maintaining close social relations. Identifying the least invariant items can provide insights into which manifestations of personality are most affected by age.

Limitations

A limitation of this study is the use of cross-sectional data and MGCFA to examine age differences in personality. Cross-sectional and longitudinal results concerning age-related differences do not necessarily converge and a replication of the present result based on longitudinal data is desirable. The biggest issue of MGCFA is the categorization of a metric variable, in this case, categorizing a span of 50 years of age into 18 categorical age groups, thus reducing the precision of the analysis. Fortunately, the large overall sample size resulted in over half of the groups encompassing only 1 or 2 years of age. The age ranges for the groups were smallest in early adulthood, the period in life in which prior studies have suggested most of the change in personality should occur. We recommend a replication with a fine-grained distinction of age in older participants. Age is often taken as a proxy for cognitive decline, but we think that life events such as retiring, becoming grandparents, or facing the death of a spouse provide important stages for significant personality changes that are not primarily associated with age. Therefore, besides a more fine-grained resolution of

age amongst older participants, we suggest that researchers examine the impact of important contextual variables (e.g., critical life events) with longitudinal designs.

Conclusion

In this article we presented a procedure for maximizing invariance and examining age differences in a hierarchical representation of personality. We showed the importance of accounting for the hierarchical structure of the personality traits for both theoretical and practical reasons. Ignoring the composition of the personality factors at the first-order factor level can lead to unclear construct definitions and thus questionable comparability of findings across studies that employ different instruments. Examining only the broad personality factors can also result in a loss of information or, in the worst case, a distorted picture of the underlying developmental trajectories. We showed that even when the most measurement-invariant items are specifically selected with sophisticated procedures, adequate measurement invariance for popular measures of personality cannot always be achieved, which can be an indication of age-related differences in the manifestation of personality. A much deeper examination of the differences in the structure of personality as a function of age might be necessary to fully understand how personality develops across the life span.

References

Allemand, M., Zimprich, D., & Hendriks, A. A. (2008). Age differences in five personality domains across the life span. *Developmental psychology*, *44*, 758.

Allemand, M., Zimprich, D., & Hertzog, C. (2007). Cross-Sectional Age Differences and Longitudinal Age Changes of Personality in Middle Adulthood and Old Age. *Journal of personality*, *75*, 323-358.

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

Ashton, M. C., Jackson, D. N., Paunonen, S. V., Helmes, E., & Rothstein, M. G. (1995). The criterion validity of broad factor scales versus specific facet scales. *Journal of Research in Personality*, *29*, 432-442.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 397-438.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495-508.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal, 13*, 186–203.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588.

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences*, *11*, 515-524.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425.

Browne, M. W., & Du Toit, S. H. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research*, *27*, 269-300.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471-492.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1-27.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255.

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the big five and Tellegen's three-and four-dimensional models. *Journal of Personality and Social Psychology*, *66*, 93.

Costa, P. T., Jr., & McCrae, R. R. (1994). Set like plaster: Evidence for the stability of adult personality. In T. F. Heatherton & J. L. Weinberger (Eds.), *Can personality change?* (pp. 21–40). Washington, DC: American Psychological Association.

Costa Jr., P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment*, *64*, 21-50.

Costa Jr., P. T., & McCrae, R. R. (1997). Stability and change in personality assessment: the revised NEO Personality Inventory in the year 2000. *Journal of Personality Assessment*, *68*, 86-94.

Deneubourg, J. L., Aron, S., Goss, S., & Pasteels, J. M. (1990). The self-organizing exploratory pattern of the argentine ant. *Journal of Insect Behavior*, *3*, 159-168.

Dorigo, M., & Stützle, T. (2010). Ant colony optimization: overview and recent advances. In *Handbook of metaheuristics*. Springer US.

Helson, R., & Moane, G. (1987). Personality change in women from college to midlife. *Journal of Personality and Social Psychology*, *53*, 176.

Helson, R., Jones, C., & Kwan, V. S. (2002). Personality change over 40 years of adulthood: hierarchical linear modeling analyses of two longitudinal samples. *Journal of Personality and Social Psychology*, *83*, 752.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal*, *6*, 1-55.

Janssen, A. B., Schultze, M., & Grötsch, A. (2015). Following the Ants. *European Journal of Psychological Assessment*.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, *2*, 102-138.

Leite, W. L., Huang, I. C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*, 411-431.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*, 151-173.

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*, 285.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130.

Marcoulides, G. A., & Drezner, Z. (2003). Model specification searches using ant colony optimization algorithms. *Structural Equation Modeling*, *10*, 154-164.

Marsh, H.W., Hau, K., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary psychometrics* (pp. 275–340). Mahwah, NJ: Erlbaum.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological assessment*, *22*, 471.

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models. *Psychological methods*, *18*, 257.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 439-476.

McCrae, R. R., & Costa Jr., P. T. (1999). A five-factor theory of personality. *Handbook of Personality: Theory and Research*, *2*, 139-153.

McCrae, R. R., Costa, P. T., de Lima, M. P., Simões, A., Ostendorf, F., Angleitner, A., ... & Chae, J. H. (1999). Age differences in personality across the adult life span: parallels in five cultures. *Developmental psychology*, *35*, 466.

McCrae, R. R., Costa Jr., P. T., Ostendorf, F., Angleitner, A., Hřebíčková, M., Avia, M. D., ... & Saunders, P. R. (2000). Nature over nurture: temperament, personality, and life span development. *Journal of Personality and Social Psychology*, *78*, 173.

McCrae, R. R., Zonderman, A. B., Costa, P. T.-J., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, *70*, 552–566.

McDonald, R. P. (1999). Test theory: A unified approach.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, *9*, 369-403.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543.

Moshagen, M., & Auerswald, M. (2017). On Congruence and Incongruence of Measures of Fit in Structural Equation Modeling.

Mroczek, D. K., & Spiro, A. (2003). Modeling intraindividual change in personality traits: Findings from the Normative Aging Study. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *58*, P153-P165.

Muthén, L. K., & Muthén, B. O. (2004-2017). Mplus. *Statistical analysis with latent variables. Version*, *7*.

Nye, C. D., Allemand, M., Gosling, S. D., Potter, J., & Roberts, B. W. (2015). Personality Trait Differences Between Young and Middle-Aged Adults: Measurement Artifacts or Actual Trends? *Journal of Personality*, *84*, 473.

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: testing competing models for designing short-scale big-five assessments. *Journal of Research in Personality*, *59*, 56-68.

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

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1.

Robins, R. W., Fraley, R. C., Roberts, B. W., & Trzesniewski, K. H. (2001). A longitudinal study of personality change in young adulthood. *Journal of Personality*, *69*, 617-640.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*, 354–373.

Schroeders, U., Wilhelm, O., & Olaru, G. (2016a). The influence of item sampling on sex differences in knowledge tests. *Intelligence*, *58*, 22-32.

Schroeders, U., Wilhelm, O., & Olaru, G. (2016b). Meta-heuristics in short scale construction: Ant Colony Optimization and genetic algorithm. *PLoS ONE*, *11*(11).

Small, B. J., Hertzog, C., Hultsch, D. F., & Dixon, R. A. (2003). Stability and change in adult personality over 6 years: Findings from the Victoria Longitudinal Study. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *58*, 166-176.

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*, 84-90.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, *100*, 330.

Spearman, C, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*, 1041.

Team, R. C. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

Terracciano, A., McCrae, R. R., Brant, L. J., & Costa Jr., P. T. (2005). Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging*, *20*, 493.

Vassend, O., & Skrondal, A. (1997). Validation of the NEO Personality Inventory and the five-factor model. Can findings from exploratory and confirmatory factor analysis be reconciled? *European Journal of Personality*, *11*, 147-166.

IV

“Grandpa, do you like roller coasters?”:

Identifying Age-Appropriate Personality Indicators.

Gabriel Olaru[1], Ulrich Schroeders[1], Oliver Wilhelm[2], & Fritz Ostendorf[3]

1: University of Kassel

2: Ulm University

3: Bielefeld University

Status – accepted

Olaru, G., Schroeders, U., Wilhelm, O., & Ostendorf, F. (2019). 'Grandpa, do you like roller coasters?': Identifying Age-Appropriate Personality Indicators. *European Journal of Personality, 33*, 264-278. DOI: 10.1002/per.2185

Abstract

Personality development research heavily relies on the comparison of scale means across age. This approach implicitly assumes that the scales are strictly measurement-invariant across age. We questioned this assumption by examining whether appropriate personality indicators change over the lifespan. Moreover, we identified which types of items (e.g., dispositions, behaviors, interests) are particularly prone to age effects. We reanalyzed the German NEO-PI-R normative sample (N = 11,724) and applied a Genetic Algorithm to select short scales that yield acceptable model fit and reliability across locally-weighted samples ranging from 16 to 66 years of age. We then examined how the item selection changes across age points and item types, respectively. Emotion-type items seemed to be interchangeable and generally applicable to people of all ages. Specific interests, attitudes, and social effect items—most prevalent within the domains of Extraversion, Agreeableness and Openness—seemed to be more prone to measurement variations over age. A large proportion of items was systematically discarded by the item selection procedure, indicating that independent of age many items are problematic measures of the underlying traits. The implications for personality assessment and personality development research are discussed.

Keywords: personality measurement, personality development, item selection, Genetic Algorithm, Local Structural Equation Modeling

In personality development research, responses to a common set of questions are compared across people of different ages. Mean differences in responses to this fixed set of questions are attributed to differences in respondents' age. This approach is the backbone and the foundation of personality development research (Roberts, Walton, & Viechtbauer, 2006). However, is it appropriate to attribute all these differences solely to age? Response patterns across age may also be affected by age-related covariates that are unrelated to personality development (Hofer, Flaherty, & Hoffman, 2006). These covariates include differences in cognitive abilities (Gnambs & Schroeders, 2017), situational transitions across life stages (Bleidorn, Hopwood, & Lucas, 2018; Wrzus & Roberts, 2017; Wrzus, Wagner, & Riediger, 2016), physical and mental constraints of elderly people, or differences in the vocabulary used by different generations. This is particularly relevant to current personality inventories developed based on assumptions about the Big Five or Five Factor model, as they (Costa & McCrae, 1995) include a wide range of specific, complex, and situational item types. In general, the personality items currently used are much more complex (e.g., "I feel embarrassed when talking in front of people") and specific (e.g., "I like to go to the ballet") than items with simple adjective ratings (e.g., "I am diligent"). The German taxonomy of personality item-trait relations (Angleitner, John, & Löhr, 1986) and personality descriptive terms (Angleitner, Ostendorf, & John, 1990) provide an exhaustive representation of item types that are commonly applied in personality questionnaires (see also, Fiske & Cox, 1979; Wiggins, 1979). A wide range of item types can be used, including questions about habits, feelings, wishes, attitudes, or social effects (see Table 1 and Coding Manual under https://osf.io/muvtc/ for an overview). Different perspectives on the diversity of the item sampling exist: for example, the developers of the BFI-2 (Soto & John, 2017) and the Eysenck Personality Questionnaire (Eysenck & Eysenck, 1975) both made the conscious decision to reduce method effects of heterogeneous item samples by only incorporating trait

attributes, emotions/cognitions, and general behavioral items in their questionnaires. In contrast, the widely used NEO-PI-R (Costa & McCrae, 1992) and HEXACO inventories (Ashton & Lee, 2009; Lee & Ashton, 2004) include nearly all item types specified by Angleitner, John, and Löhr (1986), which might potentially result in a more exhaustive measure of the underlying traits. In this article, we are interested in the influence of different item types on the measurement of personality across age. More specifically, how much are interests and attitudes affected by trends and cohort effects (e.g., "I am bored when watching ballet or modern dance"?)? Can work-habit related items (e.g., "I meet my deadlines") be applied to children or retired people in the same fashion? How much are items affected by situational cues (Rauthmann, Sherman, & Funder, 2015), given the situational transitions across life stages (Bleidorn et al., 2018; Wrzus & Roberts, 2017; Wrzus et al., 2016)?

The awareness for possible biases in personality measurement across age has grown substantially over the past two decades. An increase of studies has tested for and established measurement invariance across age (Allemand, Zimprich, & Hendriks, 2008; Allemand, Zimprich, & Hertzog, 2007; Nye, Allemand, Gosling, Potter, & Roberts, 2016; Olaru, Schroeders, Wilhelm, & Ostendorf, 2018; Small, Hertzog, Hultsch, & Dixon, 2003). The goal of these studies is typically to establish measurement invariance across age for a fixed set of items and subsequently compare factor means. In this article, the opposite approach was taken. Instead of establishing measurement invariance, we sought to identify personality items that can only be applied to specific age ranges and are thus most affected by *measurement variance*. This approach is suited to identify the most problematic items when examining personality scores across age. Detecting the most age-specific items helps explain how age-invariant personality measures can be developed. In the following, we explain in more detail our rationale and the procedures used to identify these problematic items and corresponding traits. We then present a taxonomy of item types that we will use to classify

the personality items. Finally, we will scrutinize whether specific item types, such as behaviors or interests, are particularly problematic across age.

Model Fit and Measurement Invariance

Before one can examine mean-level differences of personality across age, two measurement assumptions must be met. First, the scale score must constitute a valid representation of the underlying latent trait (Borsboom, 2006a, 2008). And second, the measurement has to be equivalent across age (i.e., the relations between manifest and latent variables has to be equal across age; Borsboom, 2006b). These two requirements can be tested by specifying a confirmatory measurement model and constraining certain model parameters (e.g., factor loadings and item intercepts) to equality across age groups. While a number of studies in personality development follow this procedure (Allemand et al., 2008, 2007; Nye et al., 2016; Olaru, Schroeders, Wilhelm, et al., 2018; Small et al., 2003), results show that obtaining measurement invariance across broad age ranges can be challenging. Satisfying model fit and measurement invariance was only achieved by parceling items into aggregates (Allemand et al., 2008, 2007; Small et al., 2003), by data-driven modifications to the measurement models (Nye et al., 2016), or by item selection (Olaru, Schroeders, Wilhelm, et al., 2018). In addition, less restrictive procedures have been proposed as alternatives to strict testing procedures in personality research. For example, *Exploratory Structural Equation Modeling* (Asparouhov & Muthén, 2009; Morin, Marsh, & Nagengast, 2013) allows for cross-loadings in the model, while the *Alignment* procedure (Asparouhov & Muthén, 2014) redistributes non-invariance across items. Although these approaches do indeed improve model fit, they incorporate variability into the model at the expense of enforcing measurement invariance. It remains unclear how much variability can be found in personality measurement if item parameters are not specifically constrained across age. In other words, how much do the appropriate indicators of the underlying traits change across

age? As a note, in the present context the term “appropriate” is used to describe items that fit the measurement model well and have sufficiently high loadings. This operationalization provides an objective indicator of which items are adequate representations of the underlying traits. If, for instance, the factor loadings of an item drops in specific age ranges, it can no longer be assumed that this item measures the presupposed underlying trait.

In general, responses to a questionnaire can be understood as a function of items used, persons tested, and measurement occasion (see Generalizability Theory; Brennan, 1992). In the case of personality development research, a fixed-item set is typically applied at different age points, either on persons of varying age (cross-sectional) or on the same persons at different measurement occasions (longitudinal). The choice of items, the age of respondents, and the interaction between items and age of the respondents all affect the measurement of personality. For instance, some items may be more representative of the underlying trait and thus a more adequate measure of personality (= item effect). Furthermore, the assessment tool might only work (i.e., adhere to psychometric standards) within a limited age range (= age effect), because the initial development process was focused on specific samples such as students, or because some of the items assess behaviors that are only relevant in specific life stages. For instance, situational behavioral items, such as “I keep my workplace tidy”, may be representative of Orderliness for working people, but have little-to-no relevance outside the working age range. After retirement, non-work-related items, such as “I keep my household tidy”, or more general adjective items, such as “I am tidy”, could be more appropriate indicator of Orderliness. Age effects and the interaction of item wording and age can severely bias the study of age-related differences in personality. We thus examined a) whether the items used to measure personality traits are equally representative of the underlying traits, and b) whether some items are only representative in a specific age range. To do so, we used

item selection procedures to identify the best-fitting items at given age points and to compare the selected items across age.

Item Sampling: The Genetic Algorithm

A Genetic Algorithm (GA; Eisenbarth, Lilienfeld, & Yarkoni, 2015; Schroeders, Wilhelm, & Olaru, 2016a; Yarkoni, 2010) was used to extract item sets that best capture the underlying traits at different age ranges (i.e., good model fit and reliability) out of a larger item pool. GAs are often used in the computational sciences to solve complex combinatorial problems. Finding a model with a reduced item set that yields sufficient model fit and reliability can also be seen as a combinatorial problem. In the present case, the problem is to find suitable item subsets among over a trillion possible item combinations per personality factor. Instead of examining all possible models, the GA applies the evolutionary principles of selection, cross-over, and mutation to heuristically find a working—but not necessarily the best—solution. Just as how life species adapt to their environment by passing on desirable traits to later generations, the GA improves item sets over several iterations in regards to a user-defined optimization function (e.g., model fit and reliability). In the first step of the iterative search process, the GA draws a number of random item samples that are evaluated based on the optimization function. Note that any psychometric criterion can be used in the optimization function, including combinations of several criteria such as reliability, model fit, measurement invariance and correlations with external outcomes (Schroeders, Wilhelm, & Olaru, 2016b). The best item subsets are then selected and randomly recombined to form new item subsets, similar to the way parents' genes are recombined to create offspring. During the recombination process, some items can be randomly removed or added (= random mutations), which ensures that no item is categorically excluded during the search process. The newly-derived item sets are evaluated based on the optimization function. Again, the best item solutions are recombined and "mutated" to create the next sample of item sets. This

process is either repeated for a pre-defined number of iterations (= generations), or until a pre-defined convergence criterion is met. A simplified illustration of the item selection procedure by means of a Genetic Algorithm is presented in Figure 1. As a side note, for the present analysis GA was favored over other well-established item selection algorithms (e.g., Ant Colony Optimization; Janssen, Schultze, & Grötsch, 2015; Leite, Huang, & Marcoulides, 2008; Olaru, Witthöft, & Wilhelm, 2015; Schroeders, Wilhelm, & Olaru, 2016b), because it is faster in finding a large number of models that fulfill a pre-defined optimization criterion. However, given the same optimization criterion and the same goal (e.g., identify the single best solution), both procedures should yield similar results.

Generation	Operation	Solution	Items 1	2	3	4	5	6	Criterion (CFI)
1	Randomly select item samples and evaluate (=create initial population and evaluate fitness)	A1	■		■		■		A1: .91
		B1		■	■			■	B1: .85
		C1	■			■		■	C1: .80
1	Identify best solutions (=selection of fittest individuals)	A1	■		■		■		A1: .91
		B1		■	■			■	B1: .85
2	Recombine best solutions and evaluate (=parents creating offspring)	A2	■	■	■			■	A2: .85
		B2		■	■		■		B2: .93
		C2	■		■		■	■	C2: .92
2	Identify best solutions (=selection of fittest individuals)	B2		■	■		■		B2: .93
		C2	■		■		■	■	C2: .92
3	Recombine best solutions…	A3			■		■	■	

Figure 1. A simplified illustration of item selection with a genetic algorithm. The example represents the search for the optimal item combination (out of six items) to maximize CFI values. Mutations (= randomly adding or removing items during the recombination phase) are not presented for reasons of clarity.

The present study explores how strongly personality measures are affected by item effects (i.e., are some items generally better indicators of the underlying traits?) and by age effects (i.e., do the best indicators of personality change across age?) on a sample with a

broad age range. To identify the most prototypical indicators of personality, GA was applied to identify item sets that yield good model fit and reliability at different age points, ranging from 16 to 66 years. Items were classified items on the taxonomy suggested by Angleitner, John, and Löhr (1986) to examine whether item and age effects are related to the item content (e.g., attribute, behavior, interest). As the underlying measurement instrument and sample, the German NEO-PI-R was used because it provides a large item sample (240 items) with a wide range of item types. Moreover, a large standardization sample (N = 11,724) spans across a broad age range. In the following section, we will first present the sample and measurement instrument used. Then, we will elaborate on the procedures used to sample participants by age and items by model fit and reliability. Patterns in the item selection rates across age and the connection to item types will be subsequently presented and discussed.

While we did not have any expectations of item types on item effects, we argue that a higher selection rate of attribute, emotion/cognition, and behavioral-type items would support the underlying considerations (i.e., these items represent the best indicators of personality) in the development of the BFI (Soto & John, 2017) and Eysenck's Personality Questionnaire (Eysenck & Eysenck, 1975). In terms of age effects, we expect attribute-type items to show the most stable item selection rates across age, as adjective descriptors should be applicable to persons of all age. In contrast, interest, attitude, and biographical-type items are expected to be most affected by cohort effects and transitions across life, thus showing the largest selection differences across age. This expectation also applies to "social effect"-type items, which might be most affected by differences in the social networks across different life stages (Wrzus et al., 2016). Besides these rather broad notions, the analyses concerning item type were exploratory rather than confirmatory in nature (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

Method

Sample

The sample consisted of 11,724 participants (7,505 females) and represented the non-clinical part of the *German Revised NEO-Personality Inventory* validation sample (Ostendorf & Angleitner, 2004)[2]. Data for the validation sample was collected in over 50 studies from 1992 to 2004, with the weighted mean date of the data collection in 1999. The majority of subsamples were of small to medium size ($N = 33$ to 324) and were collected as part of internships or theses. One large subsample was collected as part of the research project BiLSAT (Spinath, Angleitner, Borkenau, Riemann, & Wolf, 2002). The resulting sample contained 12,552 participants. Participants were discarded due to implausible response patters ($N = 324$), two data-check questions (e.g., "Did you answer honestly?"; $N = 86$), over 10% missing responses or missing demographics ($N = 373$), and an age below 16 years ($N = 45$), which yielded a remaining sample size of 11,724. Compared to the general population, the given convenience sample was well-educated, with 67.0% having or working towards a high school degree qualifying for university (compared to 19.3% in the population). A large portion (40.2%) were university students at the time of the assessment; 13.3% of the sample had at least a college degree. Age ranged from 16 to 91 years, with a mean of 29.9 years ($SD = 12.1$). In this sample, twin pairs were substantially oversampled, as 757 twin pairs from the BiLSAT (Spinath et al., 2002) project were included. Technically, monozygotic twins substantially increase the dependence between what should be independent subjects. This redundancy between observation units is somewhat smaller between dizygotic twins and even smaller for other siblings. Given that mono- and dizygotic twins were oversampled over the whole age range, the effective N available at the age points in our analysis is somewhat

[2] The data set is subject to copyright regulations. Under certain circumstances the data set can be reused for scientific purposes; please contact Dr. Ostendorf at upsyf007@uni-bielefeld.de for further information.

smaller. Given the overwhelming power available in the present analysis, this artifact is extremely unlikely to affect the substantive results reported here.

Measures

Personality was measured using the German 240-item adaptation of the NEO-PI-R (Ostendorf & Angleitner, 2004), which measures the Five Factor Model (FFM) of personality and the underlying facets. Each factor contains six facets that are measured by eight items each, for a total of 48 items per factor. Positively- and negatively-coded items were almost equally distributed across factors. Participants indicated their agreement to a given statement on a five-point scale, and response options included *strongly agree* (4), *agree* (3), *neutral* (2), *disagree* (1), and *strongly disagree* (0). We chose the NEO-PI-R for this study for several reasons. First, the underlying facet structure provides one of the broadest representations of the Five Factor Model and is well-founded theoretically (Costa & McCrae, 1992, 1995). As such, it is popular and often perceived as a "gold standard" of personality measurement. The large number of items and diversity of item types also made the NEO-PI-R particularly interesting for the current study.

In this study, items were categorized according to the personality item classification by Angleitner, John, and Löhr (1986). The categories "symptoms" (physical reactions) and "bizarre items" were removed, as such item types were not present in the NEO-PI-R. In order to provide a more nuanced differentiation of item types, the categories "abilities, talents, or their absence" and "pure evaluations" from the German taxonomy of personality descriptive terms (Angleitner et al., 1990) were also included. The resulting item categories and brief descriptions are presented in Table 1. Nine advanced students of psychology or related fields classified the NEO-PI-R items (see https://osf.io/muvtc/ for the full coding manual and the coding sheets of all raters). In case of doubt, raters chose more than one category per item. Fleiss' Kappa indicated moderate rater agreement ($\kappa = .51$). Items were classified by the

category that was mostly selected and chosen by over half the raters. The average percentage rater agreement was 77% for the final classification. Rater agreement was lowest for the category “Abilities” at 68% and highest for “Social effects”-items at 85% (see Table 1 for overview). A relative large number of 26 items could not be classified due to low rater agreement. Three items were allocated to two categories. The resulting item classification can be seen in Figure 2 (see Table 1 for an overview).

Table 1. *Item Type Classification*

Item type	Description	Example	N Items					Percent Agree.
			N	E	O	A	C	
a) Character traits, temperament	Stable character traits, often described with adjectives.	*I am friendly.*	8	11	8	14	17	.75
b) Abilities, talents, or their absence	Dispositions that describe skills or abilities or a lack thereof.	*I solve difficult tasks easily.*	2	0	2	0	7	.68
c) Emotions, moods, cognitions	Thoughts or feelings that are typically described with frequency terms (e.g., "often").	*I am often sad.*	24	6	9	2	0	.80
d) Behavioral tendencies, activities	Openly observable behavior or behavioral tendencies.	*I often go to parties.*	2	8	5	3	8	.74
e) Pure evaluations	An evaluation of the self, other people or an indicator of self-worth.	*I am worthless.*	6	2	0	10	4	.74
f) Attitudes, worldviews	Explicit political or cultural views or opinions on groups.	*All politicians are thieves.*	0	0	7	10	2	.82
g) Interests, wishes	Descriptions of hobbies, interests or wishes to perform actions.	*I find football boring.*	0	9	15	0	0	.78
h) Social effects, reactions of others	Items in which the influence on others is explicitly described.	*At parties I am rarely the center of attention.*	2	4	0	6	1	.85
i) Biographical facts	Items that refer to the past.	*I had problems with the law when I was young.*	0	1	1	0	1	.74

Note. N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness; Percent Agree. = Inter-rater agreement with final classification in percent (by item-type).

Statistical Analysis

Person (Age) Sampling.

Samples representing different age spans were created by using the weighting approach applied by local structural equation modeling (LSEM; Hildebrandt, Lüdtke, Robitzsch, Sommer, & Wilhelm, 2016; Hildebrandt, Wilhelm, & Robitzsch, 2009). In this approach, models are not estimated based on separate age groups, but rather on sample weights (Wu & Zhang, 2006). More specifically, persons are weighted by their distance to specific age points, with maximum weights at the respective focal point and decreasing weights with increasing age distance. This approach has the benefit that the continuous age variable is not categorized, which can lead to a loss of information about individual differences in groups and carries a higher risk of missing non-linear relations (MacCallum, Zhang, Preacher, & Rucker, 2002). Weighted samples were derived for every age from 16 to 66 years in steps of ten years, based on a Gaussian function around each focal age points. Ten years was selected between age points because the analysis was very demanding computationally (about three to four days per factor-age combination on a single CPU) and no differences were expected between appropriate items on a smaller level. 66 years was the highest age point because only a small number of participants had a higher age. Sample weighting and subsequent model estimation were performed in *Mplus 7.4* with the MLR estimator. The effective sample size at each age point can be seen in Table 2.

Table 2. *Results of the Item Selection*

	Models found at age						Models that fit at least at one age point	Models that fit all age points
	16	26	36	46	56	66	absolute / relative	absolute / relative
Neuroticism	7,807	14,199	10,144	8,357	9,461	4,313	16,601 / 3.32%	665 (2,895) / 0.13% (0.58%)
Extraversion	2,011	2,370	1,561	2,706	1,185	213	5,213 / 1.04%	26 (86) / 0.01% (0.02%)
Openness	6,405	16,063	12,394	5,287	3,656	3,722	26,259 / 5.25%	70 (419) / 0.01% (0.08%)
Agreeableness	597	501	198	107	236	791	1,547 / 0.31%	14 (35) / 0.00% (0.01%)
Conscientiousness	6,282	3,554	916	453	1,174	707	6,973 / 1.39%	87 (260) / 0.02% (0.05%)
N_{eff}	2,383	4,795	1,574	1,039	639	224		

Note. N_{eff} = effective weighted sample size. The number in parenthesis is the number of models that fit all age points when excluding the age point 66, which had a substantially lower effective sample size than the other age points.

The power of finding close fit with the RMSEA increases with the sample size and degrees of freedom of the model (Kim, 2005; MacCallum, Browne, & Cai, 2006). The most restrictive model tested in the current analysis (three items per facet) had 129 degrees of freedom. In this case, at least N = 95 was required to have adequate levels of power ($1 - \beta \geq .80$) for finding close fit with RSMEA ≥ .06. At the age focal point 66, the effective sample size was 224, which was sufficient for a power of 1 for all possible models. In contrast, the power of CFI testing decreases with degrees of freedom and is also dependent on factor loadings and correlations. In the case of the most restrictive model, the required sample size for adequate power levels of testing with CFI ≥ .90 was around 150 (given our factor loadings and correlations). This number increased to N = 220 for models with four items per facet. Some models with a higher number of items might hence be rejected due to low power, but this problem would only affect large models at the focal age point 66.

Item Sampling.

Each trait was examined separately. The model consisted of a second-order factor atop of six first-order factors (i.e., facets). Item numbers ranged from 18 to 48 (full model) items for every factor. The minimum number of items was 18, because at least three items per first-order factor (i.e., facet) are required for model identification without additional constraints to the model. Item samples were drawn separately for each weighted age sample with the GA algorithm implemented from the GA package (Scrucca, 2013) in R (R Development Team, 2017). The GA was applied to identify solutions with adequate model fit (CFI ≥ .90 and RMSEA ≤ .06; Hu & Bentler, 1999) and loading structure ($\lambda \geq .33$; Tabachnick & Fidell, 2007). Whereas the RMSEA was adequate for the starting models, the CFI was particularly problematic. A possible reason might be the overall relatively low loadings, which results in a smaller difference between the measurement and baseline model, hence disproportionally affecting CFI (Moshagen & Auerswald, 2017). Therefore, the GA

optimization procedure was focused on CFI and the minimal factor loading of the models. Both values were log-transformed around their critical cutoff (CFI = .90; minimal $\lambda = .33$) to maximize differentiation around this point (Janssen et al., 2015; Schroeders et al., 2016b, 2016a). The final optimization function was the sum of the two log-transformed values[3].

Not only was the best model found in each GA run examined, but also all models found during the search process that met model fit requirements and yielded no factor loading below .33. To reduce random error due to the non-exhaustive search applied in this study, all models that met our requirements on at least one age point were also subsequently tested on all other age points. This step ensured that the independent search processes applied at each age point did not negatively affect the stability of the findings across age.

Results

None of the full 48 item models for any FFM factor fit the data sufficiently well at any given focal age point (CFI $\geq .90$; RMSEA $\leq .06$; $\lambda \geq .33$). To reach acceptable model fit and loadings, the GA had to remove around half of the items. This detail does not imply that half of the items should be psychometrically flagged: rather, it means that the full model is too complex and suffers from a high number of cross-loadings and residual correlations. Moreover, this problem is common in personality research and most likely cannot avoided (Marsh et al., 2010). The number of fitting models at each age point varied substantially across factors (see Table 2), with the largest number of models found for Neuroticism and Openness, and the lowest number of adequate models found for Agreeableness. The number of unique adequate models (i.e., item subsets with acceptable model fit and factor loadings for at least one age sample) identified for each FFM factor (independent of age) ranged approximately from 1,500 to 16,000 (from a total of approximately 500,000 models

[3] The R-Script used to run the Genetic Algorithm and corresponding Mplus files (for Neuroticism) can be downloaded at: https://osf.io/muvtc/

examined per factor). Of this subset of models that fitted the data at any *specific* age point, only about one percent fitted across *all* age points (see Table 2).

Age and Item Effects by Factor

The impact of age and item effects on the item selection of appropriate personality items was examined by portraying relative item selection frequency for all items across age in Figure 2. To quantify changes in the item selection frequency across age (and differences within age), the standard deviation of the relative item selection frequency was computed both across and within age points (see Table 3). A high standard deviation *across* age indicates that some items were only selected at specific age points, which means that different sets of items are needed to measure personality at different stages of life. A high standard deviation *within* age point indicates that some items were generally selected more often than others and are more appropriate indicators of the underlying factor in general—independent of age. Patterns were not tested for significance, as the results are biased by the level (i.e., absolute vs. relative selection frequency), the number of indicators (i.e., single indicators vs. aggregates across factors/item types), and the type of testing procedure (i.e., χ^2-test vs. variance decomposition). In the following, the most relevant findings are presented.

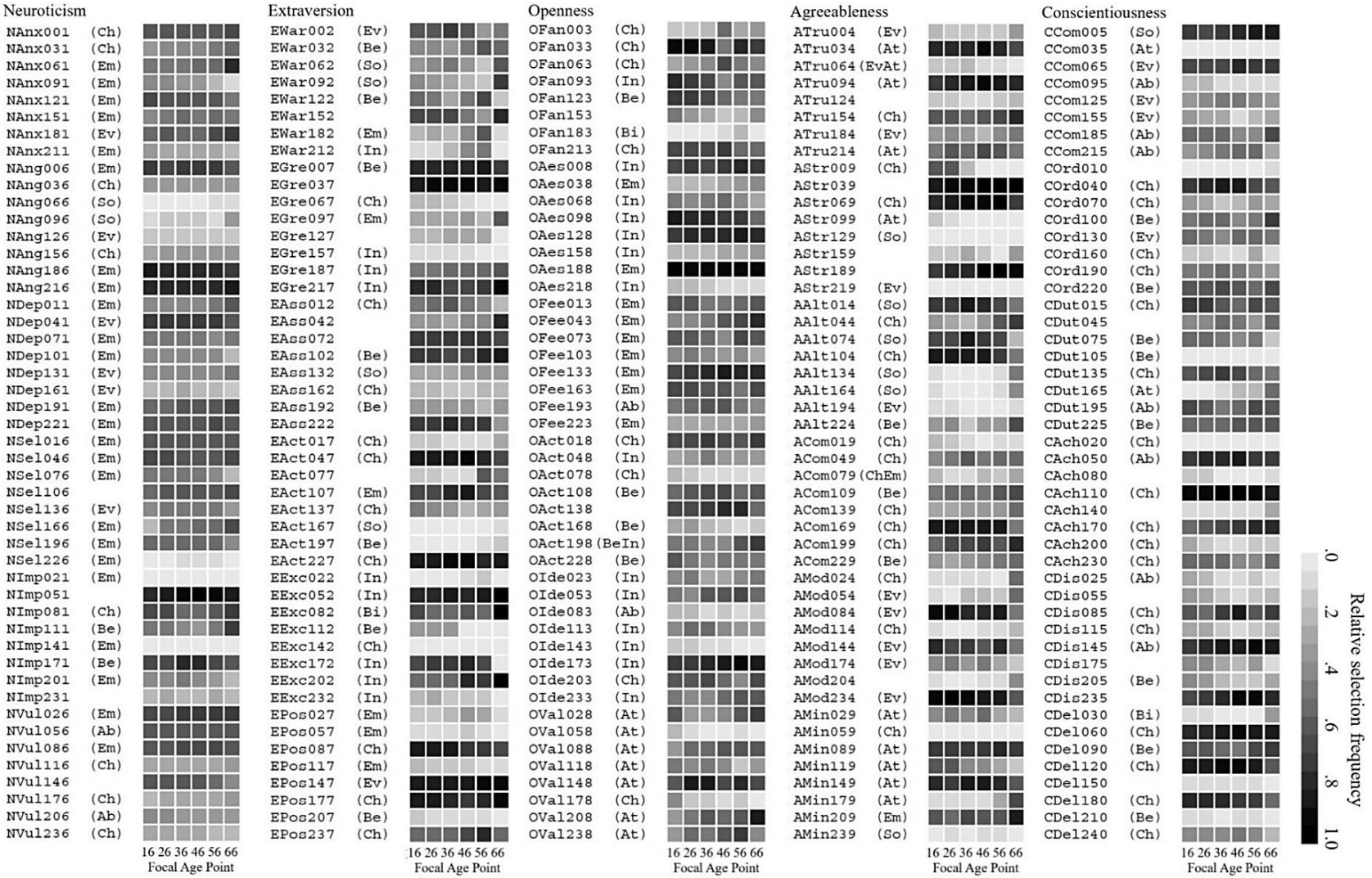


Figure 2. Item selection probability across focal age points. Item names can be interpreted as follows: The first letter represents the factor, the next three letters the corresponding facet, followed by the NEO-PI-R item number (see Table 3 for full facet names). The item type is presented in parenthesis after the item name. Ch = Character traits, temperament; Ab = Abilities, talents, or their absence; Em = Emotions, moods, cognitions; Be = Behavioral tendencies, activities; Ev = Pure evaluations; At = Attitudes, worldviews; In = Interests, wishes; So =Social effects, reactions of others ; Bi = Biographical facts. Some items could not be classified due to low rater agreement.

Table 3. *Item Selection Fluctuations and Item Type Composition of the NEO-PI-R Scales*

	Item selection			Item types of the full scales in percent									Deviation	
	Mean	*SD* (item)	*SD* (age)	Cha	Abi	Emo	Beh	Eva	Att	Int	Soe	Bio	χ^2	χ^2 uniform
Neuroticism	.44	.23	.05	.18	.05	.55	.05	.14	.00	.00	.05	.00	42.87*	96.73*
Anxiety	.46	.14	.07	.25	.00	.62	.00	.12	.00	.00	.00	.00	11.64	25.75*
Anger	.40	.34	.03	.25	.00	.38	.00	.12	.00	.00	.25	.00	9.40	12.25
Depression	.46	.17	.05	.00	.00	.62	.00	.38	.00	.00	.00	.00	19.64*	30.25*
Self-Consciousness	.45	.20	.06	.00	.00	.86	.00	.14	.00	.00	.00	.00	21.63*	40.57*
Impulsivity	.40	.33	.07	.17	.00	.50	.33	.00	.00	.00	.00	.00	8.13	15.00
Vulnerability	.44	.20	.04	.43	.29	.29	.00	.00	.00	.00	.00	.00	12.11	14.86
	Mean	SD (item)	SD (age)	Cha	Abi	Emo	Beh	Eva	Att	Int	Soe	Bio	χ^2	χ^2 uniform
Extraversion	.43	.30	.11	.27	.00	.15	.20	.05	.00	.22	.10	.02	14.82	29.90*
Warmth	.40	.14	.18	.00	.00	.14	.29	.14	.00	.14	.29	.00	10.76	7.14
Gregariousness	.46	.35	.07	.17	.00	.17	.17	.00	.00	.50	.00	.00	10.46	12.00
Assertiveness	.48	.24	.10	.40	.00	.00	.40	.00	.00	.00	.20	.00	8.01	11.20
Activity	.39	.38	.09	.57	.00	.14	.14	.00	.00	.00	.14	.00	5.88	17.43*
Excitement-Seeking	.39	.35	.13	.12	.00	.00	.12	.00	.00	.62	.00	.12	30.81*	23.50*
Positive Emotions	.44	.39	.06	.38	.00	.38	.12	.12	.00	.00	.00	.00	4.44	14.50
	Mean	SD (item)	SD (age)	Cha	Abi	Emo	Beh	Eva	Att	Int	Soe	Bio	χ^2	χ^2 uniform
Openness	.44	.23	.09	.17	.04	.19	.11	.00	.15	.32	.00	.02	30.07*	38.98*
Fantasy	.43	.24	.13	.57	.00	.00	.14	.00	.00	.14	.00	.14	14.37	17.43*
Aesthetics	.50	.31	.07	.00	.00	.25	.00	.00	.00	.75	.00	.00	35.33*	37.00*
Feelings	.49	.15	.08	.00	.12	.88	.00	.00	.00	.00	.00	.00	26.88*	48.25*
Actions	.42	.22	.08	.25	.00	.00	.50	.00	.00	.25	.00	.00	15.08	19.00*
Ideas	.40	.24	.08	.12	.12	.00	.00	.00	.00	.75	.00	.00	35.62*	34.75*
Values	.40	.23	.13	.12	.00	.00	.00	.00	.88	.00	.00	.00	62.42*	48.25*

	Mean	SD (item)	SD (age)	Cha	Abi	Emo	Beh	Eva	Att	Int	Soe	Bio	χ^2	χ^2 uniform
Agreeableness	.41	.32	.11	.31	.00	.04	.07	.22	.22	.00	.13	.00	34.09*	44.00*
Trust	.45	.32	.08	.12	.00	.00	.00	.38	.50	.00	.00	.00	26.41*	21.25*
Straightforwardness	.40	.44	.09	.40	.00	.00	.00	.20	.20	.00	.20	.00	5.59	7.60
Altruism	.38	.32	.16	.25	.00	.00	.12	.12	.00	.00	.50	.00	29.53*	16.75*
Compliance	.43	.26	.09	.67	.00	.11	.22	.00	.00	.00	.00	.00	10.26	32.00*
Modesty	.40	.35	.15	.29	.00	.00	.00	.71	.00	.00	.00	.00	30.37*	30.29*
Tender-Mindedness	.40	.31	.11	.12	.00	.12	.00	.00	.62	.00	.12	.00	30.91*	23.50*
	Mean	SD (item)	SD (age)	Cha	Abi	Emo	Beh	Eva	Att	Int	Soe	Bio	χ^2	χ^2 uniform
Conscientiousness	.41	.30	.08	.42	.18	.00	.20	.10	.05	.00	.02	.02	31.86*	55.40*
Competence	.39	.29	.07	.00	.38	.00	.00	.38	.12	.00	.12	.00	28.80*	14.50
Order	.40	.26	.07	.57	.00	.00	.29	.14	.00	.00	.00	.00	7.73	20.00*
Dutifulness	.43	.24	.10	.29	.14	.00	.43	.00	.14	.00	.00	.00	10.32	12.29
Achievement	.40	.39	.06	.83	.17	.00	.00	.00	.00	.00	.00	.00	12.88	33.00*
Self-Discipline	.41	.33	.10	.40	.40	.00	.20	.00	.00	.00	.00	.00	15.44	11.20
Deliberation	.45	.38	.08	.57	.00	.00	.29	.00	.00	.00	.00	.14	16.65*	20.00*

Note. Mean = Average item selection frequency for this factor or facet; SD (item) = Standard deviations (Variability) of the relative item selection frequency within age points; Higher numbers indicate that some items were selected more often than others and are generally more appropriate indicators of the underlying factor; SD (age) = Average Standard Deviation of the item selection across age points; Higher numbers indicate that the selection frequency of the items more substantially changes across age and that some items are not applicable across all age points; Cha = Character traits, temperament; Abi = Abilities and skills; Emo = Emotions and cognition; Beh = Behavioral tendencies; Judg = Judgment; Att = Attitudes; Int = Interests; Soe = Social effects; Bio = Biographical (see coding manual at https://osf.io/muvtc/ for a detailed description of the item types). Two measures of deviation from the norm are given: The χ^2-value shows the deviation from the expected distribution based on the existing item type distribution across all items. The uniform χ^2-value shows the deviation from a uniform distribution of item types. The critical test-statistic is $\chi^2(8) = 15.50$. Significant values are marked with an asterix (*).

Neuroticism yielded the lowest fluctuations both across and within age. This finding is in line with a previous study using the same inventory (i.e., NEO-PI-R) and sample (i.e., German NEO-PI-R normative sample), where scalar measurement invariance was also found for an abbreviated version on a facet level (Olaru, Schroeders, Wilhelm, et al., 2018). Adequate measurement models included nearly all items (low item effect), and this composition did not change substantially across age (low age effect; see the relative homogenous item selection frequencies in Figure 2 and the low standard deviations in Table 3). While a particularly large number of Openness models that fitted at some age point were found (see Table 2), the item variation across age was larger for this factor than for Neuroticism (see changing item selection frequencies across age in Figure 2 and relative high standard deviation across age in Table 3). This finding indicates that some items are only appropriate for the Openness factor in restricted age ranges and are replaced by other items when persons of different ages are assessed. Conscientiousness was particularly affected by high item effects (see high within age contrast between item-selection frequencies in Figure 2 and high within age standard deviation in Table 3). This finding means that some of the Conscientiousness items generally show higher item uniqueness, cross-loadings or residual correlations. Unfortunately, our sample consisted mostly of students and working people, which might explain the relative small age effects. Only 20 participants were older than 65 years in the current study. A broader age span might shed more light on the adequacy of the behavioral tendency items of the Conscientiousness scale that are often work-related (e.g., “I keep my workplace tidy”). Appropriate items changed most substantially across age for the factors Extraversion and Agreeableness, especially in the facets Warmth (E), Altruism (A), Modesty (A), and Excitement-Seeking (E). Similar issues were found with these two factors in a previous study examining the measurement invariance of the NEO-PI-R scales across age using the same data and similar item selection procedures (Olaru, Schroeders, Wilhelm, et al., 2018). Overall, item effects were much stronger than age effects across all factors. The results

suggest that personality items cannot be used interchangeably and that they are definitely not homogenous or drawn randomly from a hypothetical item universe (see also Loevinger, 1965). In our current analysis, some items seemed to be more decisive and essential for the underlying trait than others. This pattern is also evident when examining the three items per facet selected by us in the aforementioned study (Olaru, Schroeders, Wilhelm, et al., 2018): items included in the short-scale also showed a much higher selection probability in this study (selected: .55 on average; unselected: .35 on average; see https://osf.io/muvtc/ for the full table of item selection rates in this study and in Olaru et al., 2018).

Age and Item Effects by Item Type

Item types (Angleitner et al., 1986) were not equally distributed across traits, but are confounded with the trait measured (see Table 3 for an overview of the factor content distribution). The item type distributions were tested across factors for uniformity with χ^2-independence tests and found that nearly all scales except for Extraversion deviate from a uniform distribution (Table 3). For instance, Neuroticism in the NEO-PI-R is assessed mainly with the use of emotion-type items. Openness is measured using a large number of interest and attitude-items. Hence, we advise caution when attributing age or items effects solely to the underlying trait or item type, as the two item groups are confounded. The item type classification also yielded relatively heterogeneous groups with items measuring several different traits. Table 4 presents age and item effects by item types classified in this study (see https://osf.io/muvtc/ for classifications and selection rates at the item level).

Table 4. *Item Selection Frequency and Fluctuation by Item Type*

Item type	Mean	*SD* (items)	*SD* (age)
Character traits, temperament	.47	.31	.10
Abilities, talents, or their absence	.40	.24	.08
Emotions, moods, cognitions	.44	.24	.07
Behavioral tendencies, activities	.40	.25	.09
Pure evaluations	.44	.31	.08
Attitudes, worldviews	.42	.27	.12
Interests, wishes	.45	.28	.10
Social effects, reactions of others	.26	.29	.11
Biographical facts	.37	.37	.11
Not classified	.43	.28	.09

Note. SD (item) = Average Standard Deviation of the item selection withing age points; SD (age) = Average Standard Deviation of the item selection across age points.

In the following, we present the most salient findings. For instance, “Social effect”-type items (e.g., “I am the center of attention”; “Others think of me highly”) were systematically discarded by the item selection procedure and show among the highest age effects. Biographical items (e.g., “I used to play theater as a child”)—although rarely used in the NEO-PI-R—also stood out due to relative high item and age effects. The largest age effects could be found for Attitude items, which were most often used to measure Openness for Values (O), Tender-Mindedness (A) and Trust (A) in the context of moral, ethics, religion and open-mindedness. While interest items generally yielded average age effects, the subset of interest items measuring Excitement-Seeking (e.g., “I like the thrill of roller coasters”; “I avoid watching scary movies”) seem to decrease in relevance across age. On the positive side, emotion and cognition items (e.g., “I am often sad”) yielded the lowest item and age effects. The item types were most prevalent in the Neuroticism (N) factor, as well as the Openness for Feelings (O) and Positive Emotions (E) facets. In general, the item type composition of the factors after the item selection remained relatively unchanged across age (see Online Supplement at https://osf.io/muvtc/).

Discussion

In personality development research, people of different ages are typically asked to respond to a common and fixed set of items. Responses are then compared on an aggregated level across age with the implicit assumption that the measurement is age-invariant (i.e., strictly measurement invariant). Given the high specificity and subjectivity of personality items, and the lack of measurement invariance across age for broad personality inventories, we questioned this assumption. We examined how measurement of personality is affected by age when the item set is allowed to vary. More specifically, we tried to identify the most appropriate items in six age samples ranging from 16 to 66 years of age. We additionally classified and grouped items (e.g., character traits, emotions, behavioral tendencies) to examine whether specific item types are particularly affected by item and age effects. This approach provides a different perspective on measurement invariance and personality development across age. If selection probabilities in the implemented item selection procedure change substantially across age, this might indicate changes in the common item variance and subsequently factor composition across age. Such a variation would also affect the way we conceive the measurement of personality over the life span. Note that the age differences examined in this study are unrelated to normative differences at the item and factor level, but represent a type of structural change in personality (Caspi & Roberts, 2001). As such, the traits that show the strongest absolute differences across age (e.g., Neuroticism and Conscientiousness; see Roberts et al., 2006) do not necessarily show the strongest age effects in the variance-covariance structure. Strictly speaking our results are only valid for a specific inventory and sample (i.e., the German NEO-PI-R and corresponding normative sample); however, we have no theoretical reason to believe that the general findings will be different with other measures or samples.

Age and Item Effects by Factor

Not all factors were affected to the same degree by age and item effects in the item selection probability. Surprisingly, Neuroticism and Conscientiousness, which typically show among the strongest mean-level changes across age (Roberts et al., 2006), yielded the lowest age effects in this study. Findings on the structural change in the personality factors is lacking, as structural continuity (i.e., measurement invariance) needs to be established before normative values can be compared. As a result, many studies use parceling techniques or similar to achieve this goal (Allemand et al., 2008, 2007; Jackson et al., 2009). But as mentioned before, we were able to achieve scalar measurement invariance for a shortened version of the Neuroticism and Conscientiousness scales in a previous study (Olaru, Schroeders, Wilhelm, et al., 2018) using the same questionnaire (i.e., NEO-PI-R) and sample. We were unable to do so for the Extraversion and Agreeableness scales, which also showed the highest age effects in this study, potentially explaining the difficulty to establish measurement invariance across age. Given the surprisingly low proportion of measurement models that fit across all age points (e.g., around 0.01% of all examined models for Extraversion and Agreeableness), the assumption of measurement invariant scales is extremely implausible. The high prevalence of attitude, specific interest, “social effects”, and biographical items on these two factors might explain these effects. The high prevalence of these items in the NEO-PI-R Openness, Extraversion, and Agreeableness scales might have resulted in the relatively high age effects found for these factors.

Age and Item Effects by Item Type

“Social effect”-type items were systematically discarded by the GA and seem to be problematic measures across broad age spans. We think that these items should be avoided as measures of personality in general, as they change the point of view from the respondent to that of other people. Respondents are forced to interpret how they are perceived and/or treated by others, which has been shown to lead to diverging evaluations in previous research

(Connolly, Kavanagh, & Viswesvaran, 2007). The measurement based on this item type can also be confounded with the respondents' social relation and the level of Neuroticism. This item type also showed large age effects, which were in line with expectations: decreasing levels of self-consciousness across life (Olaru, Schroeders, Wilhelm, et al., 2018) and a transition of social relations from friends and colleagues to family among adults (Wrzus et al., 2016) can bias the measurement with these items across age. Cohort effects might have also resulted in the high age effects, as younger participants might generally experience more social criticism or even bullying due to a higher exposure through social media than older respondents.

Biographical items (e.g., "I had trouble with the law when I was young") should also be avoided when comparing participants of different age. The distance between the respondents' age and the age in question can disturb the examination of age-associated differences. The events described in these items seem to be particularly relevant to respondents that show agreement on these items, as they need to remember these events much later in life. Older respondents also have a higher chance of showing the behaviors in question. In particular, biographical items that referred to specific life stages such as childhood were omitted by the item selection procedure. In this case, past *open* behavior might not necessarily be a good predictor of future behavior, as personality volatility is strongest in childhood and early adolescence (Roberts & DelVecchio, 2000; Roberts et al., 2006). This might be particularly problematic since the retrieval of such semantic memories is strongly influenced by the current self-concept: Memories are much more likely to be retrieved when they support current self-concepts (Baumeister & Cairns, 1992; Swann & Schroeder, 1995). Similarly, memories may be altered to match the self-concepts (Epstein, 1973; McAdams, 1993). Generally, it has been shown that the retrieval of relevant episodic memories does not improve trait judgments (nor the other way around; Klein, Loftus, & Sherman, 1993; Klein, Loftus, Trafton, & Fuhrman, 1992). Interestingly, even amnesia does

not affect the ability to make accurate trait judgments despite a lack of episodic memories (Klein, Loftus, & Kihlstrom, 1996). As such, we question the purposefulness of including biographical items to objectify the self-evaluation of personality traits.

A relation between self-concepts and emotion-related items can also be found in self-report measures (Robinson & Clore, 2002). When comparing retrospective emotion judgments with on time reports, people scoring high in Neuroticism tend to report more or stronger negative emotions, whereas people high in Extraversion will typically overestimate positive emotions (Charles & Carstensen, 2010). Instead of trying to remember how often somebody felt angry, depressed, insulted or tired during the last years, people tend to rely on their self-concept to provide an estimate of how likely they experience these emotions on a more general level (Charles & Carstensen, 2010). Relying on relative stable self-beliefs (instead of much more volatile state-specific episodic memories) might explain the age-effects found for the emotion items. Differently put, emotion items seem to be interchangeable (i.e., low item-effects) because the base for self-assessment are common and stable self-concepts (e.g., vulnerability, sadness, or general emotionality). Older individuals typically experience (or recall) less negative and more positive emotions (Charles & Carstensen, 2010), but this finding on the mean level did not affect the age stability of the items.

Attitude items (e.g., “All people deserve respect”) showed among the strongest age effects in this study. The NEO-PI-R Attitude items were most prevalent on Openness for Values, Tender-Mindedness and Trust and captured mostly morals, ethics, empathy, and open-mindedness (vs. a fixed set of beliefs). Similar to the personality traits in general, attitudes tend to stabilize with age (Glenn, 1980), based on the same processes of situation (or social network) selection, behavior and subsequent evaluations (Caspi & Roberts, 2001; Wrzus & Roberts, 2017). However, the fluctuations found at the group level were generally unsystematic and thus difficult to interpret. We refrained from interpreting single items,

particularly since it is likely that cohort and selection effects (i.e., an overrepresentation of students in the younger age groups) might have impacted the item selection effects.

While interest items (e.g., “I find ballet boring”) were not uniformly affected by age effects, we would like to point out that the measurement of very specific interests may be affected by age differences. For instance, the items used to measure Excitement-Seeking (E) mostly capture the desire to engage in specific activities (e.g., “I like the thrill of roller coasters”). These items decreased in relevance with increasing age (see Figure 2), as respondents may become less inclined to or physically capable of engaging in such activities. Capturing Excitement Seeking via the desire to feel extreme activity-induced emotions might result in a more comparable measurement across age. Interest items might also be impacted by cohort effects, as the availiability of the activities in question can change across time (e.g., increased availiability of and exposure to extreme sports nowadays).

The classification of items was done after test construction and yielded heterogeneous groups of items, in particular across different traits. This assessment is supported by the high item uniqueness, as expressed by high item variability effects. In addition, the German NEO-PI-R items are also particularly long (10.1 words per item on average) compared to other German inventories (e.g., German BFI-2 has an average of 6.2 words per item) and are thus more complex linguistically. Even relatively simple items types, such as those measuring temperament and character traits, are surprisingly long in the NEO-PI-R (8.4 words per item). Previous research pointed out that, even in linguistically simple measures (such as the Rosenberg Self-Esteem Scale), the dimensional structure might change dependent of respondents’ reading ability (Gnambs & Schroeders, 2017). As such, we advise caution when generalizing the item-type findings of this study to other personality inventories. In addition, differences in the selection rates were much smaller at the aggregated level (i.e., by item type or factor) than some differences found at the individual item level. With the exception of the problematic “social effect” and biographical items, all item types can provide valid measures

of the underlying personality traits. Issues in personality measurement seem to be more related to some problematic items, rather than item types. These items can be eliminated using item selection procedures, such as GA or Ant Colony Optimization (Janssen et al., 2015; Leite et al., 2008; Olaru, Schroeders, Hartung, & Wilhelm, 2018; Olaru, Schroeders, Hartung, et al., 2018; Olaru et al., 2015; Schroeders et al., 2016a). In the following section, the advantages and disadvantages of this approach will be discussed.

The Advantages of Item Sampling for Personality Research

None of the full scales were found to fit adequately—neither across the full age range, nor at a specific age point. When item selection reduced scales to half their original size, a fraction of all possible measurement models showed adequate fit. Surprisingly, the degree to which specific items were discarded or selected across the six age samples was relatively stable. That is, large item effects on selection probability indicate that some NEO-PI-R items are generally less appropriate for the underlying traits than others. By using item selection procedures as the one presented in this study (for a tutorial on item selection in personality research, see also Olaru, Schroeders, Hartung, et al., 2018), either these problematic items can be removed or the model complexity can be reduced while still maintaining construct coverage (Yarkoni, 2010). Other researchers have recently examined differences in model fit and reliability between personality questionnaires that adhere or violate widely accepted item construction norms (Pargent, Hilbert, Eichhorn, & Bühner, 2018). As they did not find systematic differences between the “improved” and “deteriorated” version of the NEO-FFI personality questionnaires, they questioned the ability of current psychometric indices to distinguish between items of high and low quality. The item selection process presented in this study should be able to distinguish between more and less appropriate items, as the item selection has consistently discarded or preferred certain items over others across all age samples.

Note that problems with finding adequate personality measurement models under the strict evaluation of model fit indices in CFA are well known and the implications are controversially discussed in the literature (Borkenau & Ostendorf, 1990; Borsboom, 2006a; Church & Burke, 1994; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996; Olaru et al., 2015; Vassend & Skrondal, 1997). Often, the high number of cross-loadings is attributed to the subjectivity of the self-report method (Hopwood & Donnellan, 2010), which has led to more permissive model testing procedures such as ESEM (Asparouhov & Muthén, 2009; Morin et al., 2013) or *Alignment* (Asparouhov & Muthén, 2014). If the issue was related to the measurement method in general—hence affecting all items equally—item selection frequencies would be equal across all items (i.e., no item effect), apart from random fluctuation. However, given the large differences in item selection, which were stable across age, the issue of missing measurement invariance over age seems to be attributable to specific items rather than to the measurement method in general. Therefore, we again want to point out that item selection procedures (Olaru, Schroeders, Wilhelm, et al., 2018; Olaru et al., 2015; Schroeders et al., 2016a) are better-suited than laxer testing procedures when tackling the issue of poor model fit.

Item selection for the sake of improving the psychometric properties of a measure has been criticized for narrowing the breadth and depth of the construct or inflating Type 1 and Type 2 error (Credé, Harms, Niehorster, & Gaye-Valentine, 2012; Kruyen, Emons, & Sijtsma, 2012, 2013). However, meaningful comparisons of personality scores across age can only be made after two prerequisites are met. First, the scale score must be a valid unidimensional representation of the underlying latent trait (Borsboom, 2006a, 2008). Second, the relation between the manifest and latent variables (i.e., items and personality factors) must be invariant across age (Borsboom, 2006b). Unmodified full scales in personality measurement do not seem to meet either of these prerequisites (Borkenau & Ostendorf, 1990; McCrae et al., 1996; Olaru, Schroeders, Wilhelm, et al., 2018; Olaru et al., 2015; Small et al., 2003; Vassend

& Skrondal, 1997). Merging manifest indicators into a smaller number of aggregates for model testing—also known as parceling—might indeed improve model fit (Little, Cunningham, Shahar, & Widaman, 2002) and measurement invariance (Little, Rhemtulla, Gibson, & Schoemann, 2013), but only by masking item and age effects of single manifest variables and thus only by tackling the problem of non-invariant scales on a superficial level. Instead, we recommend eliminating the bias in comparisons across age by first eliminating non-measurement invariant items before scores are compared (see Olaru, Schroeders, Wilhelm, et al., 2018). To tackle the bandwidth-fidelity issue (Cronbach, 1960), one could also try to maintain the construct coverage of a scale by retaining the facet structure of personality models or by maximizing the correlation between the short and the long version during the optimization process.

From a conceptual perspective, personality traits are understood as the shared variance between the individual items. Item variance that cannot be explained by the extracted factors, the residual, is often simply interpreted as measurement error. However, recent advancements in personality assessment have demonstrated substantial cross-rater agreement, rank-order stability, heritability (Mõttus, Kandler, Bleidorn, Riemann, & McCrae, 2017; Mõttus et al., 2018) and criterion-related validity of the item residuals (Seeboth & Mõttus, 2018). The individual items have been labeled as *personality nuances* (McCrae, 2015) and represent the lowest level of the personality trait hierarchy (i.e., below the domain and facet level). This perspective on the importance of item uniqueness seems to be at odds with the assumption that a specific set items is affected by a shared set of processes (see the debate on correspondence vs. emergence, Baumert et al., 2017). But it also reminds us how much the empirical results we discuss in personality assessment are tied to a specific item sets and are far away from the hypothetical item universe of exchangeable items. What are some analogies between the theory of personality nuances and item selection in general, and measurement variance studied in the present contribution in particular? From an item selection perspective,

removing items and thus eliminating nuance variance could be problematic. For instance, Seeboth and Mõttus (2018) used elastic net regression to examine the criterion-related validity of item residuals, which corresponds to an item selection of the most predictive indicators. Similarly, the most measurement invariant items across age can be selected to provide the least biased estimate to compare mean levels across age (Olaru, Schroeders, Wilhelm, et al., 2018), or items can be freely drawn at each age to examine which nuances are most relevant for a specific age. That is, instead of relying personality assessment on a fixed set of measurement invariant items, it might be fruitful to use item sampling procedures to select the most relevant indicators with high item uniqueness.

Taken together, it is not recommended to view the full scale of personality inventories and similar instruments as a gold standard suitable for all research questions. Instead, it is best thought of as just one possible representation of an (unfortunately) often underspecified and underdetermined item universe. Current personality inventories are constructed based on a blend of rational and inductive construction techniques and subsequent item selection with the goal of improving factor loadings or reliability—the latter often reduced to Cronbach's Alpha (Costa & McCrae, 1995; Johnson, 2014; Soto & John, 2017). Model fit and measurement invariance across age have not been addressed specifically in the development of these inventories. Item selection with the goal of improving invariance can be understood as an additional development or item-selection step of identifying subsets of items that also fulfill these criteria (ideally, these criteria should be optimized simultaneously).

Future Directions

The current findings are limited to the item types and specific items present in the NEO-PI-R. An examination of item and age effects with additional item classifications is desirable, such as classifications based on linguistic complexity and adherence to item construction norms (see also Pargent et al., 2018). Examining variability effects in one personality inventory provides preliminary insight, as the coverage of item types and item

attributes is also restricted to the specific instantiation. The NEO-PI-R uses a wide variety of item types, but its items are more complex and specific overall than indicators from similar inventories. The BFI (Soto & John, 2009) and BFI-2 (Soto & John, 2017), for instance, measure personality via character trait-items of relative low complexity with no situational cues. In addition, about half of the BFI-2 (Soto & John, 2017) items contain more than one character trait per item and thus often represent "aggregates" of prototypical personality descriptive terms (e.g., "I see myself as someone who is outgoing, sociable"). It would be interesting to study how these relative short items fare in the context of measurement variability across age and whether "aggregate"-items can help reduce age and item effects. Regardless, even such character trait items might not be appropriate measures of personality across the entire life, as research has shown that different adjectives than the ones derived from lexical studies are needed to describe the personality of children (Kohnstamm, Mervielde, & Havill, 1998). An examination based on a large item pool containing a wide variety of personality measurements would be desirable to truly understand the interplay between item type, item wording, and measurement variations across age. The *Synthetic Aperture Personality Assessment* project is currently collecting data on 696 personality items from 92 public domain personality scales (Condon & Revelle, 2015), which provides an excellent database to tackle this question. In sum, future studies on age and item effects in personality measurement could make use of a much larger item pool.

In the present study, we found that item types used to measure personality factors were confounded with the factor in question. Factors and item types seem to be naturally connected by the underlying theory of personality to a certain degree (Angleitner et al., 1986). Neuroticism, for example, is predominantly framed in terms of emotions and cognitions, while Openness is primarily expressed through asking for interests and attitudes. Arguably, personality inventories do not have to adhere to this assumed inter-connection of item types and personality factors to be valid and reliable representations of the underlying traits. From

the classical perspective of latent traits representing the overlap between personality indicators, similar item types within factors are desirable for a homogenous measurement and thus high factor loadings or internal consistency. But given the new findings on the validity of personality nuances (Mõttus et al., 2017, 2018; Seeboth & Mõttus, 2018), a more diverse measurement of the personality traits may be desirable (McCrae, 2015). Future test development efforts in the realm of personality inventories could focus on developing inventories with an exhaustive factor x item-type coverage. For example, Conscientiousness-emotion-items (e.g., "I only feel well when everything is in order") and Extraversion-attitude-items (e.g., "The more the merrier") would fill gaps in the factor-item type matrix. A blueprint for such an item construction procedure can, for instance, be found in the *Personality Research Form* (Ashton, Jackson, Helmes, & Paunonen, 1998), which has been developed with the goal of measuring personality factors across a wide variety of item types (e.g., motives and goals, reactions to social situations, behaviors). Alas, the factors of this test do not exhaustively represent broadly accepted taxonomies like the Big Five model.

Given the results of this study and previous studies on measurement invariance of personality, two competing perspectives on age effects in personality measurement are discussed (see also, Church et al., 2011; McCrae, 2015). To begin with a psychometric perspective, it is desirable to have little-to-no age effects on what items constitute prototypical measures of the underlying trait. Factors, facets, and their constitution are the same across a broad age range. Only then can mean-levels of the personality factors be compared across age (or adjust for differences in age heterogeneous samples). Normative differences in the personality factors are then understood to reflect some form of universal personality development. From a second perspective, age-related changes in factors, facets, and their manifestations might be more profound. Such changes include failing traditional invariance tests (e.g., Church et al., 2011; Huang, Church, & Katigbak, 1997), but these changes might be influenced by other factors as well. This paper began with a more liberal perspective:

instead of just stressing the psychometric perspective by attempting to identifying invariant item sets, severe violations of invariance constraints were allowed. This liberalization was allowed as such a perspective can be considered enriching for the study of personality development. Although a wealth of information is lost when using only age-invariant items and age-differentiated scales with corresponding cues (e.g., school, work, interests) and age-appropriate situational demands (Rauthmann et al., 2015), this approach can provide a much more precise and exhaustive measure of personality at different life stages.

Both perspectives to personality measurement are viable and one perspective cannot be preferred over the other. Whether an age-invariant or an age-differentiated measure should be created depends on the underlying research goal. Both measurement intentions can be pursued with the methods presented in this article (for an in-depth tutorial, see Olaru, Schroeders, Hartung, et al., 2018). However, given that the age-differentiated perspective has hardly been pursued in research on personality development, it will be exciting to seek age-varying factors, facets, and nuances (McCrae, 2015) of established and potentially novel personality constructs. Can new, less impulsivity-saturated facets be found for Extraversion from middle-adulthood on? Is Conscientiousness less achievement-prone once people retire? How and with which manifestations does Neuroticism develop in early infancy? Do occupations with strongly entrepreneurial or social components affect the constitution of the Honesty-Humility factor? If the answers to such questions is "yes", then personality can no longer be easily compared across age. Moreover, this question of differentiation is not restricted to age alone but can be asked about other context variables such as sex, socio-economic status, ethnicity, culture as well.

Conclusion

This article examined how factor composition changes as a function of respondents' age. More specifically, changes in appropriate personality indicators as a result of age-associated changes were examined. This research effort relevant because many personality

development studies assume that manifest scale scores are measurement invariant across age. This assumption was challenged for several reasons: a) a lack of measurement invariance across age in most broad personality inventories, b) strong mean effects for some of the first-order factors, and c) an apparent lack of appropriateness of some items for some life stages. Some item types seem to be affected more by a lack of age invariance (e.g., attitudes, specific interests, social effects and biographical items) than others. For instance, older participants might not be physically able or willing to engage in some of the specific interests measured in Excitement Seeking (E), and interests and attitudes of young participants may deviate from those of older respondents. In contrast, emotion-type items seemed to be broadly applicable to persons of the age range tested in this study. However, age effects were relatively small compared to the substantial item selection differences found in this study, indicating that some items are generally less appropriate indicators of the underlying personality factors—independent of age. Item selection procedures can help identify sets of appropriate items and substantially improve the measurement of personality while also ensuring that the items are measurement invariant across age (Olaru, Schroeders, Hartung, et al., 2018; Olaru, Schroeders, Wilhelm, et al., 2018; Olaru et al., 2015; Schroeders et al., 2016a). Personality researchers are thus encouraged to test for and to establish measurement invariance across age with item selection procedures before examining age related mean level differences.

References

Allemand, M., Zimprich, D., & Hendriks, A. A. J. (2008). Age differences in five personality domains across the life span. *Developmental Psychology*, *44*, 758–770. https://doi.org/10.1037/0012-1649.44.3.758

Allemand, M., Zimprich, D., & Hertzog, C. (2007). Cross-sectional age differences and longitudinal age changes of personality in middle adulthood and old age. *Journal of Personality*, *75*, 323–358. https://doi.org/10.1111/j.1467-6494.2006.00441.x

Angleitner, A., Ostendorf, F., & John, O. P. (1990). Towards a taxonomy of personality descriptors in German: A psycho-lexical study. *European Journal of Personality*, *4*, 89–118. https://doi.org/10.1002/per.2410040204

Ashton, M. C., Jackson, D. N., Helmes, E., & Paunonen, S. V. (1998). Joint factor analysis of the personality research form and the Jackson Personality Inventory: comparisons with the big five. *Journal of Research in Personality*, *32*, 243–250. https://doi.org/10.1006/jrpe.1998.2214

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. *Journal of personality assessment*, *91*, 340-345. https://doi.org/10.1080/00223890902935878

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 397–438. https://doi.org/10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 495–508. https://doi.org/10.1080/10705511.2014.919210

Baumeister, R. F., & Cairns, K. J. (1992). Repression and self-presentation: When audiences interfere with self-deceptive strategies. *Journal of Personality and Social Psychology, 62*, 851-862. http://dx.doi.org/10.1037/0022-3514.62.5.851

Bleidorn, W., Hopwood, C. J., & Lucas, R. E. (2018). Life events and personality trait change: life events and trait change. *Journal of Personality*, *86*, 83–96. https://doi.org/10.1111/jopy.12286

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences*, *11*, 515–524. https://doi.org/10.1016/0191-8869(90)90065-Y

Borsboom, D. (2006a). The attack of the psychometricians. *Psychometrika*, *71*, 425. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D. (2006b). When does measurement invariance matter? *Medical Care*, *44*, S176-S181. http://dx.doi.org/10.1097/01.mlr.0000245143.08679.cc

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research & Perspective*, *6*, 25–53. https://doi.org/10.1080/15366360802035497

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*, 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Caspi, A., & Roberts, B. W. (2001). Personality development across the life course: The argument for change and continuity. *Psychological Inquiry*, *12*, 49–66.

Charles, S. T., & Carstensen, L. L. (2010). Social and emotional aging. *Annual Review of Psychology*, *61*, 383–409. https://doi.org/10.1146/annurev.psych.093008.100448

Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology*, *101*, 1068–1089. https://doi.org/10.1037/a0025290

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology*, *66*, 93–114. https://doi.org/10.1037//0022-3514.66.1.93

Condon, D., & Revelle, W. (2015). Selected personality data from the SAPA-Project: On the structure of phrased self-report items. *Journal of Open Psychology Data*, *3*. https://doi.org/10.5334/jopd.al

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). the convergent validity between self and observer ratings of personality: a meta-analytic review. *International Journal of Selection and Assessment*, *15*, 110–117. https://doi.org/10.1111/j.1468-2389.2007.00371.x

Costa, P. T., & McCrae, R. R. (1992). *Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo-ffi)*. Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: hierarchical personality assessment using the revised neo personality inventory. *Journal of Personality Assessment*, *64*, 21–50. https://doi.org/10.1207/s15327752jpa6401_2

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology*, *102*, 874–888. https://doi.org/10.1037/a0027403

Cronbach, L. J. (1966). *Essentials of Psychological Testing* (3rd ed.). New York, NY: Harper & Row.

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory–Revised (PPI-R). *Psychological Assessment*, *27*, 194–202. https://doi.org/10.1037/pas0000032

Epstein, S. (1973). The self-concept revisited: Or a theory of a theory. *American Psychologist*, *28*, 404.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire (junior and adult)*. Kent, UK: Hodder and Stoughton.

Fiske, S. T., & Cox, M. G. (1979). Person concepts: The effect of target familiarity and descriptive purpose on the process of describing others 1. *Journal of Personality*, *47*, 136–161.

Glenn, N. D. (1980). Values, attitudes, and beliefs. *Constancy and Change in Human Development*, 596–640.

Gnambs, T., & Schroeders, U. (2017). Cognitive abilities explain wording effects in the rosenberg self-esteem scale. *Assessment*, 107319111774650. https://doi.org/10.1177/1073191117746503

Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., & Wilhelm, O. (2016). Exploring factor model parameters across continuous variables with local structural equation models. *Multivariate Behavioral Research*, *51*, 257–258. https://doi.org/10.1080/00273171.2016.1142856

Hildebrandt, A., Wilhelm, O., & Robitzsch, A. (2009). Complementary and competing factor analytic approaches for the investigation of measurement invariance. *Review of Psychology*, *16*, 87–102.

Hofer, S. M., Flaherty, B. P., & Hoffman, L. (2006). Cross-sectional analysis of time-dependent data: mean-induced association in age-heterogeneous samples and an alternative method based on sequential narrow age-cohort samples. *Multivariate Behavioral Research*, *41*, 165–187. DOI: 10.1207/s15327906mbr4102_4

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346. DOI: 10.1177/1088868310361240

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. DOI: 10.1080/10705519909540118

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: differential item functioning in the neo personality inventory. *Journal of Cross-Cultural Psychology*, *28*, 192–218. DOI: 10.1177/0022022197282004

Jackson, J. J., Walton, K. E., Harms, P. D., Bogg, T., Wood, D., Lodi-Smith, J., … Roberts, B. W. (2009). Not all conscientiousness scales change alike: a multimethod, multisample study of age differences in the facets of conscientiousness. *Journal of Personality and Social Psychology*, *96*, 446–459. DOI: 10.1037/a0014156

Janssen, A. B., Schultze, M., & Grötsch, A. (2015). Following the ants. *European Journal of Psychological Assessment*, *33*, 409–421. DOI: 10.1027/1015-5759/a000299

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78–89. DOI: 10.1016/j.jrp.2014.05.003

Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*, 368–390. DOI: 10.1207/s15328007sem1203_2

Klein, S. B., Loftus, J., & Kihlstrom, J. F. (1996). Self-knowledge of an amnesic patient: Toward a neuropsychology of personality and social psychology. *Journal of Experimental Psychology: General*, *125*, 250. http://dx.doi.org/10.1037/0096-3445.125.3.250

Klein, S. B., Loftus, J., & Sherman, J. W. (1993). The role of summary and specific behavioral memories in trait judgments about the self. *Personality and Social Psychology Bulletin*, *19*, 305–311. http://dx.doi.org/10.1177/0146167293193007

Klein, S. B., Loftus, J., Trafton, J. G., & Fuhrman, R. W. (1992). Use of exemplars and abstractions in trait judgments: A model of trait knowledge about the self and others. *Journal of Personality and Social Psychology*, *63*, 739. http://dx.doi.org/10.1037/0022-3514.63.5.739

Kohnstamm, G. A. , Halverson, C. F., Jr. , Mervielde, I. , & Havill, V. L. (Eds.). (1998). *Parental descriptions of child personality: Developmental antecedents of the Big Five?* Hillsdale, NJ: Erlbaum.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: when is short too short? *International Journal of Testing*, *12*, 321–344. DOI: 10.1080/15305058.2011.643517

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013). On the shortcomings of shortened tests: a literature review. *International Journal of Testing*, *13*, 223–248. DOI: 10.1080/15305058.2012.703734

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*, 329–358. http://dx.doi.org/10.1207/s15327906mbr3902_8

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*, 411–431. DOI: 10.1080/00273170802285743

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 151–173. DOI: 10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*, 285–300. DOI: 10.1037/a0033266

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19–35. DOI: 10.1037/1082-989X.11.1.19

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40. DOI: 10.1037//1082-989X.7.1.19

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471–491. DOI: 10.1037/a0019227

McAdams, D. P. (1993). *The stories we live by: Personal myths and the making of the self*. New York, NY: Guilford Press.

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, *19*, 97–112. DOI: 10.1177/1088868314541857

McCrae, R. R., Zonderman, A. B., Costa, P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, *70*, 552–566. DOI: 10.1037/0022-3514.70.3.552

Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. In *Structural equation modeling: A second course, 2nd ed.* (pp. 395–436). Charlotte, NC, US: IAP Information Age Publishing.

Moshagen, M., & Auerswald, M. (2017). On congruence and incongruence of measures of fit in structural equation modeling. *Psychological Methods*, *Publish Ahead of Print*. DOI: 10.1037/met0000122

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*, 474. DOI: 10.1037/pspp0000100

Mõttus, R., Sinick, J., Terracciano, A., Hřebíčková, M., Kandler, C., Ando, J., . . . Jang, K. L. (2018). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology.* Advance online publication. http://dx.doi.org/10.1037/pspp0000202

Nye, C. D., Allemand, M., Gosling, S. D., Potter, J., & Roberts, B. W. (2016). Personality trait differences between young and middle-aged adults: measurement artifacts or actual trends? *Journal of Personality*, *84*, 473–492. DOI: 10.1111/jopy.12173

Olaru, G., Schroeders, U., Hartung, J., & Wilhelm, O. (2018). A tutorial on novel item and person sampling procedures for personality research. *Manuscript submitted for publication in the European Journal of Personality*.

Olaru, G., Schroeders, U., Wilhelm, O., & Ostendorf, F. (2018). A confirmatory examination of age-associated personality differences: Deriving age-related measurement-invariant solutions using ant colony optimization. *Journal of Personality, 86,* 1037-1049. DOI: 10.1111/jopy.12373

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality*, *59*, 56–68. DOI: 10.1016/j.jrp.2015.09.001

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

Pargent, F., Hilbert, S., Eichhorn, K., & Bühner, M. (2018). Can't make it better nor worse. *European Journal of Psychological Assessment*, 1–9. DOI: 10.1027/1015-5759/a000471

Rauthmann, J. F., Sherman, R. A., & Funder, D. C. (2015). Principles of situation research: Towards a better understanding of psychological situations. *European Journal of Personality*, *29*, 363–381. DOI: 10.1002/per.1994

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*, 3–25. DOI: 10.1037//0033-2909.126.1.3

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1–25. DOI: 10.1037/0033-2909.132.1.1

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, *128*, 934. DOI: 10.1037//0033-2909.128.6.934

Schroeders, U., Wilhelm, O., & Olaru, G. (2016a). Meta-heuristics in short scale construction: Ant Colony Optimization and Genetic Algorithm. *PLOS ONE*, *11*, e0167110. DOI: 10.1371/journal.pone.0167110

Schroeders, U., Wilhelm, O., & Olaru, G. (2016b). The influence of item sampling on sex differences in knowledge tests. *Intelligence*, *58*, 22–32. DOI: 10.1016/j.intell.2016.06.003

Scrucca, L. (2013). GA: A package for Genetic Algorithms in R. *Journal of Statistical Software*, *53*. DOI: 10.18637/jss.v053.i04

Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality, 32,* 186-201. DOI: 10.1002/per.2147

Small, B. J., Hertzog, C., Hultsch, D. F., & Dixon, R. A. (2003). Stability and change in adult personality over 6 years: findings from the Victoria longitudinal study. *The Journals of Gerontology: Series B*, *58*, P166–P176. DOI: 10.1093/geronb/58.3.P166

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*, 84–90. DOI: 10.1016/j.jrp.2008.10.002

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*, 117–143. DOI: 10.1037/pspp0000096

Spinath, F., Angleitner, A., Borkenau, P., Riemann, R., & Wolf, H. (2002). German Observational Study of Adult Twins (GOSAT): A Multimodal Investigation of Personality, Temperament and Cognitive Ability. *Twin Research, 5*, 372-375. doi:10.1375/twin.5.5.372

Swann, W. B., & Schroeder, D. G. (1995). The search for beauty and truth: A framework for understanding reactions to evaluations. *Personality and Social Psychology Bulletin*, *21*, 1307–1318. http://dx.doi.org/10.1177/01461672952112008

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Allyn & Bacon/Pearson Education.

Vassend, O., & Skrondal, A. (1997). Validation of the NEO Personality Inventory and the five-factor model. Can findings from exploratory and confirmatory factor analysis be reconciled? *European Journal of Personality*, *11*, 147–166. DOI: 10.1002/(SICI)1099-0984(199706)11:2<147::AID-PER278>3.0.CO;2-E

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. DOI: 10.1177/1745691612463078.

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, *37*, 395. http://dx.doi.org/10.1037/0022-3514.37.3.395

Wrzus, C., & Roberts, B. W. (2017). Processes of personality development in adulthood: The TESSERA framework. *Personality and Social Psychology Review*, *21*, 253–277. DOI: 10.1177/1088868316652279

Wrzus, C., Wagner, G. G., & Riediger, M. (2016). Personality-situation transactions from adolescence to old age. *Journal of Personality and Social Psychology*, *110*, 782–799. DOI: 10.1037/pspp0000054

Wu, H., & Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*. Hoboken, NJ: John Wiley & Sons.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*, 180–198. DOI: 10.1016/j.jrp.2010.01.002

V

Epilogue

In the following, I will first give a brief overview of the findings in the second and third manuscript, which examined research questions related to measurement invariance (or the DIF-paradox) from different perspectives: A) identifying measurement invariant indicators for a comparison of factor means across age groups, at the cost of potentially losing information relevant to specific age ranges vs. B) identifying indicators relevant at restricted age ranges for a precise measurement within, but not across age groups. I will subsequently link the results to existing personality development research and provide an outlook on further research in the field.

Manuscript 2: A Confirmatory Examination of Age-Associated Personality Differences: Deriving Age-Related Measurement Invariant Solutions using Ant Colony Optimization

In the first study, we used Ant Colony Optimization (ACO; Janssen et al., 2015; Leite et al., 2008; Olaru et al., 2015; Schroeders et al., 2016) to derive measurement invariant models of personality for subsequent examination of structural and normative differences in personality across age. We specified a higher-order model of personality that includes facet factors as well as the broad trait domains of personality based on the NEO-PI-R (Costa & McCrae, 1992; Ostendorf & Angleitner, 2004) and corresponding Five Factor Model of personality (Costa & McCrae, 1995). We then described which aspects of the model are prone to structural and normative changes across age. ACO was applied to maximize model fit and measurement invariance through item selection in a multi-group confirmatory setting with 18 age groups ranging from 16 to 65 years. Based on the measurement invariant models, we examined structural differences at the second-order level and normative differences where sufficient measurement invariance levels were achieved. We then compared the results to findings in personality development literature. The results showed problems of achieving measurement invariance for broad personality factors across broad age ranges, in particular for Extraversion and Agreeableness. Even though findings at the trait domain level mirrored previous findings on personality development (i.e., an increase in Emotional Stability,

Agreeableness and Conscientiousness across age; Roberts et al., 2006), mean-level patterns for the facets of most factors differed substantially across age. As such, we did not find support for measurement invariance at the second-order factor level for the factors – the only exception being Conscientiousness. This stresses the importance of examining personality development at the facet level, as normative trends at the trait domain level are not representative of the underlying facets and can thus be somewhat misleading. However, factor loadings remained stable across the entire age range, suggesting that the factor composition remains equivalent across age. The most noteworthy contributions of this article to the field of personality development are a) the use of a hierarchical model of personality that incorporates both broad trait domain and specific facet factors, b) the examination of structural differences in higher order models of personality and c) the demonstration of the purposefulness of item selection procedures for the optimization of model fit and measurement invariance.

Manuscript 3: "Grandpa, do you like roller coasters?": Identifying Age-Appropriate Personality Indicators.

In the second study, we used the item and person sampling procedures Genetic Algorithm (GA; Eisenbarth et al., 2015; Schroeders et al., 2016; Yarkoni, 2010) and Local Structural Equation Modeling (LSEM; Hartung, Doebler, Schroeders, & Wilhelm, 2018; Hildebrandt et al., 2016, 2009) to study the interaction between the NEO-PI-R personality items and respondents' age. More specifically, we tried to identify items that are valid measures of the underlying traits at specific age ranges, and subsequently compared selection frequencies across age. This allowed us to answer the question whether a common set of indicators is able to capture personality across a broad age range from 16 to 66 years, and which items are most age-sensitive. Results showed that age effects on the NEO-PI-R measurement of personality across age are comparatively small to item selection differences within age, indicating that a large set of the NEO-PI-R items shows a high item uniqueness and may be unrelated to the underlying factor – independent of the respondents' age. We

additionally categorized items based on the content assessed (e.g., temperament, emotions, interests) to examine whether specific item types are more prone to fluctuations in measurement across age. With the exception of the rarely used items measuring social reactions or biographical facts, all item types seemed to be prototypical personality indicators across age. This supports the notion that personality can be measured with a wide range of indicators (e.g., characteristics, emotions, interests, attitudes) without sacrificing validity. Contributions to the field of personality development constitute a) the use of a different perspective on personality development, namely item *measurement variance* across age, b) the examination of age effects on personality item types and c) the combination of both novel item and person procedures, in this case a GA and LSEM.

Item selection is not a new concept in personality assessment, as the large number of short-scales (Donnellan et al., 2006; Gerlitz & Schupp, 2005; Gosling et al., 2003; Rammstedt & John, 2007; Saucier, 1994) show. It is also a crucial step in the development of many broad personality inventories, where items are typically selected from the initially created item pool based on several criteria, typically main loadings in Principal Component Analysis (PCA), correlations with other personality inventories, and/or expert rated construct coverage (this also applies to the creation of many personality short scales). I discussed some of the flaws of these selection criteria in the manuscripts, such as the assumption that PCA is a latent modeling technique and the disregard for model fit (or the equalization of PCA simple structure with model fit). I also discussed some of the problems regarding person sampling in personality development research, such as the reliance on student samples when developing personality tests and the categorization of the continuous moderator age when examining personality differences across age. In the following, I provide an outlook on future research on personality with the use of the presented item and person sampling procedures.

Outlook

The HEXACO Model of Personality

The Big Five factors of personality have been derived using a lexical analysis of personality-descriptive adjectives across a wide variety of languages, including (among others) English, Dutch, German, Italian and Polish (Raad, Perugini, Hrebícková, & Szarota, 1998). In the lexical analysis, an exhaustive list of personality-descriptive list of adjectives is derived from a dictionary (e.g., *Webster's Unabridged Dictionary of the English Language* for American-English; Allport & Odbert, 1936), which typically consists of around 4% (Raad et al., 1998) of all dictionary entries. Dictionaries are typically used because these documents provide a well-maintained and exhaustive list of the examined language used by several generations. The initially derived list of all personality-descriptive adjectives is typically reduced based on ratings of familiarity, frequency of use, representativeness and uniqueness to provide a more manageable number of entries for subsequent analysis. Depending on the study, descriptions of abilities and evaluations – e.g., great, horrible – were also excluded. The remaining adjective markers are then applied as self-report measures and the underlying structure is analyzed using PCA. PCAs with five components generally identified the five factors known as the Big Five factors of personality (Goldberg, 1990). However, evidence in favor of the five factors is not undisputed. In particular, the fifth factor – often labeled Intellect, Imagination, Creativity, Unconventionality, or Openness – differs strongly across studies (see, Raad et al., 1998 for an overview; note that this also depends on whether abilities were excluded in the initial item sampling step). More recent examinations of the lexical studies also show strong support for a robust sixth factor, encapsulating descriptions of sincerity, fairness and lack of entitlement or greed (Honesty-Humility factor; Ashton et al., 2004). The extraction of six instead of five components also resulted in a change in the factors Neuroticism and Agreeableness, as descriptions of anger and irritability moved from the first to the latter, resulting in a less negative Emotionality and an Agreeableness vs. Anger factor.

Item sampling approaches can be used in this context to examine the shared and unique variance between the HEXACO and Five Factor Model. Ashton and Lee (2007) argue that the HEXACO model captures inter-individual differences in personality more exhaustively than the Five Factor Model or the Big Five and thus also provides higher prediction validity of many relevant outcomes. However, differences in the latent factors need to be separated from differences in the items used to measure these. By selecting items that provide the most distinct/overlapping measures between these two models of personality, the influence of the item sets used to measure these models can be separated from meaningful differences in the personality structure. For instance, if it is possible to create perfectly correlated scales of Honesty-Humility and Agreeableness/Conscientiousness in the Five Factor Model, one can assume that the unique variance captured by the Honesty-Humility scale is based on the indicators used. Questions of differentiation/de-differentiation between the two concurring models can also be examined across continuous moderators, such as the factors themselves (e.g., Honesty-Humility). This can help address the question for which persons the sixth factor is most distinct from the other personality factors.

Circumplex Models of Personality

The number of factors is not the only aspect in question. Another property of the Big Five factors that needs to be reevaluated is the orthogonality of the factors. In the aforementioned lexical studies, the five (or six) factors of personality were derived using a varimax (i.e., orthogonal) rotation of the principal components. The derived personality traits are typically understood as independent factors (Goldberg, 1993). However, correlated models of personality generally provide better model fit if simple structure of the model is enforced (McCrae, Zonderman, Costa, Bond, & Paunonen, 1996). These correlations are reduced when cross-loadings between factors are included in the model (Marsh et al., 2010; which is also an argument that has been made in favor of Exploratory Structural Equation Models of personality), but still substantial. One of the best-known models of personality that

discards the notion of orthogonality is the circumplex model of personality (Hofstee, De Raad, & Goldberg, 1992; Wiggins & Broughton, 1985). In the case of the Abridged Big Five Circumplex Model (AB5C), facets represent a blend of up to two overarching factors (Hofstee et al., 1992). This resulted in a total of 90 unipolar facets, or 45 bipolar facets respectively. For instance, Gregariousness is a central facet of Extraversion, but Friendliness or Poise represent blends of Extraversion and Agreeableness or Emotional Stability respectively. The facets themselves represent unidimensional constructs, and cross-loadings are only conceptualized at the facet level. However, personality models with cross-loadings at the item level have also been proposed as an alternative to overly strict unidimensional models of personality, most prominently by Marsh and colleagues (2010; 2014) in their article applying Exploratory Structural Equation Modeling (ESEM) to personality data. As mentioned earlier, this procedure reduces the correlations between the Big Five factors as shared variance between items across different factors is not forced into the factor correlations, but instead represented as cross-loadings. The resulting factor correlations thus correspond more to the original assumption of orthogonality between the Big Five. I criticized ESEM in the manuscripts for violating unidimensionality assumptions of the personality factors and deriving theoretically hard to justify cross-loadings. However, studies, such as the one conducted by Hofstee and colleagues (1992), provide evidence that most personality-descriptive adjectives may also be seen blends of the personality factors instead of unidimensional measures. As such, this assumption of simple structure might need to be challenged. The question arises whether cross-loadings arise at the item level or facet level (e.g., AB5C; Hofstee et al., 1992), or both. This question could be addressed by constraining factor main and cross-loadings to their expected relation with the underlying traits. We used item sampling procedures to select indicators that provide the most unidimensional measures of the Five Factor Model. Alternatively, the item selection procedures could be used in the context of circumplex models to identify the items that maximize model fit and the overlap

between empirical and theoretically assumed factor loadings (e.g., to ensure that the main factor loadings of the factors or facets are twice as large as the cross-loadings; see Hofstee et al., 1992)

Domain Sampling

I mentioned briefly in the introduction that the facets measured differ strongly across personality inventories, often depending on theoretical (e.g., HEXACO, circumplex, Big Five, or Five Factor Model) and practical (e.g., length of the scale) considerations. Research on the facet structure of personality suffers strongly from jingle-jangle fallacies, where it is unclear which constructs are distinct or redundant. Consider for instance the international personality item project (ipip.ori.org), which contains 463 different personality scales, apparently measuring 274 different constructs. Which of these constructs are redundant? How many of these domains can be subsumed below the higher-order Big Five or HEXACO factors? And which constructs exist "outside" of the dominant personality theories? The last question is often a matter of heated debate, with scale developers and personality researchers arguing whether a "new" construct is or is not a facet of the Big Five of other broad personality theories (see, Credé, Tynan, & Harms, 2017; Lee & Ashton, 2005, 2014; Pfattheicher, Geiger, Hartung, Weiss, & Schindler, 2017; Roberts, Lejuez, Krueger, Richards, & Hill, 2014). Similar to the items used to measure personality, the facet sets used can also be considered a random sample from a much larger facet universe (e.g., the 274 IPIP constructs). As such, questions on the influence of domain (i.e., facet) sampling on personality measurement provide a relevant research topic. To address this research question, the item selection algorithms ACO and the GA could be applied on broad personality datasets to sample domains – similar to how we sampled items. The Synthetic Aperture Personality Assessment (SAPA) currently collects data on 7,000 items measuring temperament, abilities, and interests from 92 openly-available personality inventories (Condon & Revelle, 2015), and provides an ideal basis for such examinations of the domain/facet structure of personality. Using the

optimization procedures presented in this dissertation, domains can for instance be sampled to provide the highest overlap/differences between concurring personality taxonomies (e.g., HEXACO and the Five Factor Model). In a similar fashion, a common and exhaustive facet structure of personality across all SAPA personality inventories can be derived. To achieve this, the optimization goal could be to maintain full construct coverage of all scales applied (i.e., all of the unique variance), while reducing redundancies within the selected domains (see, Yarkoni, 2010). By doing so, the jingle-jangle fallacy and the much-debated facet structure of personality could be addressed with just a single study.

Personality Nuances

In the previous sections, I presented alternatives regarding the factor and facet structure of the Big Five or Five Factor model. I now want to present a new conceptualization regarding the item level of personality (more specifically: item residuals). Generally, the common variance between items is extracted in factor analytic approaches to derive the latent personality factors. The remaining item variance, which is unique to the item (i.e., not shared between items), is then assumed to be measurement error or some other type of unrelated variance. Consider the following example: The 120 Item version of the IPIP NEO (Johnson, 2014) measures the Openness facet *Artistic Interests* using the following items: “Believe in the importance of art.”, “See beauty in things that others might not notice.”, “Do not like poetry. (reverse)” And “Do not enjoy going to art museums. (reverse)”. In the currently dominant conception of personality, it is assumed that people that believe in the importance of art will generally also see beauty in things and like poetry, as all these interests are caused by the underlying *Artistic Interests* trait. Intra-individual variations in the responses to these items (e.g., high interest in poetry, but low interest in art), which cannot be explained by the common *Artistic Interests* trait, are considered measurement or random error. However, some researchers have argued that these item residuals have a meaningful impact in our daily behavior (McCrae, 2015; Mõttus et al., 2018). Two people with the same *Artistic Interest*

score can behave very differently, one going to the museum on the weekend, the other listening to a poetry session. As such, the first person will be more sophisticated in art, whereas the other person will have a deeper understanding of poetry. The residual variance (i.e., not accounted for by the facet/factor) in these items can thus be seen as a meaningful latent trait as well, representing a personality trait even more specific than the facets. This lowest level of personality has been termed personality nuances (McCrae, 2015), and has shown heritability and incremental validity beyond the personality trait domains or facets alone (Mõttus, Kandler, Bleidorn, Riemann, & McCrae, 2017; Mõttus et al., 2018; Seeboth & Mõttus, 2018). The consequence of this would be to either use even longer personality inventories to also model the nuance trait level, or to treat single item residuals as meaningful latent factors (as long as the first is not possible due to measurement length constraints). Because only the latter approach is currently feasible, McCrae (2015) proposed a change in our conceptualization of personality models, namely a shift from reflective to formative models. In the following section, I provide more details on the differences between reflective and formative models and discuss the implications for personality research.

Formative Models of Personality Measurement

Currently, personality traits are modeled as reflective models, with the underlying personality traits causing the measured indicators (e.g., behaviors, interests). The derived latent factors only capture the shared variance between these indictors, whereas unique variance is modeled as residual variance or random error. Formative models on the other hand assume that the latent factor is caused by the indicators, and as such considers all variance as meaningful. In the following sections, I will discuss the implications of these approaches to personality models and how the two perspectives can be reconciled (see, Markus & Borsboom, 2013) and studied using item sampling procedures.

In reflective models, the indicators are considered to be replacable without changing the underlying latent construct, whereas changing the indicators in a formative model will

also change the resulting latent construct. In addition, reflective models require items to be highly correlated, whereas formative models have no such requirement. When personality items are sampled to provide as much construct coverage as possible, the overlap between items will inevitably be reduced (e.g., five and ten item personality inventory; Gosling et al., 2003). As such, formative models that encompass all of the item variance may be better suited to capture the breadth of the personality traits when using short personality measures (Bollen & Diamantopoulos, 2017; Myszkowski, Storme, & Tavani, 2018). The model used will also have an severe impact on the item sampling procedure applied: Whereas reflective models encourage sampling indicators with a shared common cause (causal theory of measurement; Markus & Borsboom, 2013) and thus a high overlap, formative models require a representative item sample from the broad item domain of personality items – not necessarily showing any shared variance (behavioral domain theory; Markus & Borsboom, 2013). This raises the interesting question whether personality indicators should be sampled based on distinctness, thus providing the broadest construct coverage, or based shared variance or centrality (as is typically done when selecting indicators based on main loadings). The first approach benefits from a formative model, which maintains all captured item variance, whereas the latter is dependent on reflective models, which rely on the shared item variance to extract representations of the latent personality factors.

Even though Seeboth and Mõttus (2018) never refer to their approach as a formative measurement model (which are regarded with skepticism in psychological measurement for not providing “true measurement models”; see Bollen & Diamantopoulos, 2017 for an overview), their use of elastic net regressions on the item residuals can be seen as a formative approach designed to maximize the predictive validity of personality measurement. As mentioned earlier, Myszkowski, Storme and Tavani (2018) also demonstrated the usefulness of formative personality models for maximizing the correlation of short personality measures with broader personality inventories – thus maximizing the construct coverage of the short

inventories. In my opinion, one of the strongest evidences in favor of using formative models of personality was presented by Yarkoni (2010), who created a 200 item short scale that was capable of capturing the full variance of "200 personality scales". Again, even though not specifically using the term *formative measurement model*, Yarkoni optimized the correlation between the scale sum scores using a Genetic Algorithm, thus also technically optimizing formative measurement models of personality by selecting items that contribute as much unique variance to the scale score as possible. In a more recent article, Yarkoni and Westfal (2017) argue that the robust prediction of future behavior (i.e., formative) can provide a deeper understanding of human behavior than the often-used approach of fitting small reflective models to specific person samples (with sometimes questionable generalizability across samples). Even though formative and reflective measurement models are often seen as concurring or incompatible approaches to psychological measurement (Bollen & Diamantopoulos, 2017), Markus and Borsboom (2013) tried to reconcile both approaches by creating a shared theoretical framework. Their model of psychological assessment includes both a reflective and formative part: reflective latent traits are the underlying cause of responses to manifest indicators, but the domain score, which is derived from these indicators, represents a formative construct. This common framework can provide a great model in which indicators can be sampled to optimize both the representation of the latent trait on the reflective part of the model, while also optimizing the predictive aspect of the formative part of the model, in line with the suggestions by Yarkoni and Westfal (2017). Person sampling approaches, such as LSEM, also provide a great method of investigating the robustness of the reflective and predictive (i.e., formative) aspect of personality models across a wide range of continuous moderator variables. Using LSEM, the relation between the reflective and formative construct could also be examined in this framework. For example, it would be interesting to investigate under which conditions the predictive validity is highest, depending for instance on the mean-level of the attribute. In my opinion, combining reflective and

formative approaches to personality measurement to optimize both explanation and prediction may be one of the most exciting opportunities for personality research.

However, I also want to point out that in my opinion the Big Five or Five Factor model are so popular, because these five abstract entities are capable of predicting a wide range of meaningful outcomes, such as job success, longevity and life satisfaction (Barrick & Mount, 1991; Barrick, Mount, & Judge, 2001; Ozer & Benet-Martinez, 2006; Poropat, 2009). The correlation between the personality factors and outcomes could unquestionably be increased by using predictive machine learning algorithms. However, much of the appeal of the Big Five rests on the simplicity and parsimony of their use (i.e., a robust model with only five factors), which would get lost in the process. With a focus on prediction, weights for every predictor-criterion combination would have to be derived separately to maximize correlations with each criterion. In addition, these outcome-focused formative weights will inevitably capitalize on an overlap between predictor and criterion (e.g., impulsivity item "overeating" and BMI; Vainik, Mõttus, Allik, Esko, & Realo, 2015). This is problematic because a sufficient explanatory distance is required when trying to predict future behavior (in contrast to just finding high correlations in a cross-sectional context): If a person is already overeating, it is likely that his/her BMI is already raised or that it is too late to prevent the outcome (i.e., raised BMI). This then translates to an increased cost in treatment compared to an early prevention – this is particularly important when dealing with mental or physical illnesses. In contrast, if a person's high impulsivity levels have already been detected in childhood or young adulthood (of course assuming a high stability in this trait across life), we would know that this person is potentially prone to overeating and could counteract this issue before it would occur. The latent Big Five factors are stable enough across life to provide a distant prediction of relevant outcomes several years ahead of their occurrence – the best example being longevity (Ozer & Benet-Martinez, 2006). Whether the derived formative models can maintain their predictive validity across such long time spans – or whether they

just capitalize on a closeness between predictor and outcome – needs to be addressed with extensive longitudinal research and methods adequate to investigate causality (e.g., *directed acyclic graphs*; Rohrer, 2018).

Network Analysis of Personality

Another new conceptualization of personality that deserves attention discards the notion of latent traits altogether and describes personality as an interplay between behaviors independent of underlying traits. This network perspective on personality rests on the theory of mutualism (Cramer, Waldorp, van der Maas, & Borsboom, 2010; Van Der Maas et al., 2017; Van Der Maas et al., 2006) and conceptualizes personality as behaviors that cause and reinforce each other instead of common latent traits causing related behaviors. To quote Cramer and colleagues (2012): "you can't like parties if you don't like people". From the network perspective, the relations we observe between personality items constitute reinforcement and inhibition processes between these behaviors instead of underlying latent entities. If somebody likes people, he or she will be more inclined to meet strangers and thus will also tend to visit more social events. Positive experiences at these social events will then further increase this person's positive perception of other people, thus further enforcing his tendency to search new contacts. Researchers supporting the mutualism theory (originally applied in intelligence research; Van Der Maas et al., 2006) even go so far to argue that the dominance of latent factor theory is based on pragmatism instead of theoretical meaningfulness: Factor theory allows researchers to extract a relative reliable representation of personality from a large set of mediocre indicators (e.g., self-reports; Van Der Maas et al., 2017) instead of having to develop highly valid indicators. The factors also provide a very parsimonious framework that allows several behaviors to be described with a much smaller set of latent factors. However, I want to point out that network analysis (e.g., Schmittmann et al., 2013) has gained much of its popularity in field of clinical psychology (Borsboom & Cramer, 2013; Cramer et al., 2010; van Borkulo et al., 2015), where a clear-cut number of

finite symptoms is associated with each affliction. The interplay between this finite set of symptoms can then be studied using network analysis (ideally on longitudinal data to examine causality). However, no such truly finite set of indicators can be derived in the case of personality traits. Instead, each personality inventory only represents a small sample of a virtually infinite personality item universe. Network analysis would require an exhaustive set of personality related behaviors, interests, emotions, etc. to provide a meaningful representation of personality processes, which would then be impossible to interpret due to the graphical nature of this method. Reflective models of personality do not have this requirement, as the indicators are typically seen as interchangeable.

The psychometric usefulness of network analysis in cross-sectional settings has also been challenged, in particular related to the lack of incremental information beyond factor analysis (e.g., centrality correlates perfectly with factor loadings; see, Hallquist, Wright, & Molenaar, 2019). The psychometric parameters derived in network analysis are also overly sensitive to spurious correlations (e.g., betweenness; Hallquist, et al., 2019). Depending on which indicators are used in the personality measurement, these parameters will vary dramatically (also making a combination of item sampling and network analysis unfeasible). Using a combination of network analysis and LSEM to study structural differences in personality across age (or any other moderator) may be visually pleasing, but lacks meaningful psychometric characteristics required for an investigation of measurement invariance. The generalizability of these findings will also be questionable, due to the item sample differences across the various personality inventory. Another downside of currently available network analysis tools is the inability to distinguish between sources of variance, such as for instance bifactor models of latent traits, which can separate item variance into general and more specific traits (Hallquist et al., 2019). Considering the high stability (Roberts & DelVecchio, 2000) and hereditary (50% or more; Bleidorn et al., 2010; Kandler, 2012; Kandler, Riemann, Spinath, & Angleitner, 2010) of personality traits, and a lack of a

finite set of indicators, I do not think that network perspectives on personality are capable of replacing the current latent variable theories of personality.

Conclusion

In the last sections of my dissertation I presented a wide range of contexts in which the presented item and person sampling procedures can be applied. Research using item selection methods is still in its infancy. Further applications can provide new perspectives on many research questions in the field of personality assessment (or psychological assessment in general). As presented in this dissertation, item sampling procedures can be used to derive short-scales that meet the theoretical assumptions of the underlying personality theories (e.g., unidimensionality in the Big Five or Five Factor Model; cross-loading patterns in circumplex models). Item and/or domain sampling can be used to investigate the range of shared or unique variance between concurring personality theories, such as the Big Five and HEXACO model of personality. The item sampling procedures can also be applied to identify indicators that maximize the unidimensionality and reliability of the reflective part of personality models, while also optimizing the prediction of the formative part of the assessment (i.e., the scale score). The person sampling procedure LSEM can then be used to examine the robustness of all aforementioned models across continuous moderator variables. Or a combination of the presented item and person sampling procedures can be used to derive robust models, or models with maximized measurement precision within restricted ranges on the moderator variable. In summary, the item and person sampling procedures presented in this dissertation are very flexible tools that can be applied in any context. This dissertation focused on research questions related to measurement invariance in personality development research, but also provides a foundation for many of the aforementioned research topics.

References

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., De Vries, R. E., Di Blas, L., … De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology*, *86*, 356. DOI: 10.1037/0022-3514.86.2.356

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1–26. DOI: 10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, *9*, 9–30. DOI: 10.1111/1468-2389.00160

Bleidorn, W., Kandler, C., Hülsheger, U. R., Riemann, R., Angleitner, A., & Spinath, F. M. (2010). Nature and nurture of the interplay between personality traits and major life goals. *Journal of Personality and Social Psychology*, *99*, 366. DOI: 10.1037/a0019982

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, *22*, 581. DOI: 10.1037/met0000056

Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121. DOI: 10.1146/annurev-clinpsy-050212-185608

Costa, P. T., & McCrae, R. R. (1992). *Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo-ffi)*. Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the revised NEO personality inventory. *Journal of Personality Assessment*, *64*, 21–50. DOI: 10.1207/s15327752jpa6401_2

Cramer, A. O., Van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., … Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*, 414–431. DOI: 10.1002/per.1866

Cramer, A. O., Waldorp, L. J., van der Maas, H. L., & Borsboom, D. (2010). Complex realities require complex theories: Refining and extending the network approach to mental disorders. *Behavioral and Brain Sciences*, *33*, 178–193. DOI: 10.1017/S0140525X09991567

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*, 492. DOI: 10.1037/pspp0000102

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP Scales: Tiny-yet-effective measures of the Big Five Factors of personality. *Psychological Assessment*, *18*, 192–203. DOI: 10.1037/1040-3590.18.2.192

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory–Revised (PPI-R). *Psychological Assessment*, *27*, 194–202. DOI: 10.1037/pas0000032

Gerlitz, J.-Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persoenlichkeitsmerkmale im SOEP. *DIW Research Notes*, *4*, 2005.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*, 14. DOI: 10.1037//0022-3514.59.6.1216

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*, 26. DOI: 10.1037/0003-066X.48.1.26

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504–528. DOI: 10.1016/S0092-6566(03)00046-1

Hallquist, M., Wright, A. G., & Molenaar, P. C. (2019). Problems with centrality measures in psychopathology symptom networks: Why network psychometrics cannot escape psychometric theory. 10.31234/osf.io/pg4mf

Hartung, J., Doebler, P., Schroeders, U., & Wilhelm, O. (2018). Dedifferentiation and differentiation of intelligence in adults across age and years of education. *Intelligence*, *69*, 37–49. DOI: 10.1016/j.intell.2018.04.003

Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., & Wilhelm, O. (2016). Exploring Factor Model Parameters across Continuous Variables with Local Structural Equation Models. *Multivariate Behavioral Research*, *51*, 257–258. DOI: 10.1080/00273171.2016.1142856

Hildebrandt, A., Wilhelm, O., & Robitzsch, A. (2009). Complementary and competing factor analytic approaches for the investigation of measurement invariance. *Review of Psychology*, *16*, 87–102.

Hofstee, W. K., De Raad, B., & Goldberg, L. R. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*, 146. DOI: 10.1037/0022-3514.63.1.146

Janssen, A. B., Schultze, M., & Grötsch, A. (2015). Following the ants: Development of short scales for proactive personality and supervisor support by ant colony optimization. *European Journal of Psychological Assessment*, 1–13. DOI: 10.1027/1015-5759/a000299

Kandler, C. (2012). Nature and nurture in personality development: The case of neuroticism and extraversion. *Current Directions in Psychological Science*, *21*, 290–296. DOI: 10.1177/0963721412452557

Kandler, C., Riemann, R., Spinath, F. M., & Angleitner, A. (2010). Sources of variance in personality facets: A multiple-rater twin study of self-peer, peer-peer, and self-self (dis)agreement. *Journal of Personality*, *78*, 1565–1594. DOI: 10.1111/j.1467-6494.2010.00661.x

Lee, K., & Ashton, M. C. (2005). Psychopathy, Machiavellianism, and narcissism in the Five-Factor Model and the HEXACO model of personality structure. *Personality and Individual Differences*, *38*, 1571–1582. DOI: 10.1016/j.paid.2004.09.016

Lee, K., & Ashton, M. C. (2014). The dark triad, the big five, and the HEXACO model. *Personality and Individual Differences*, *67*, 2–5. DOI: 10.1016/j.paid.2014.01.048

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*, 411–431. DOI: 10.1080/00273170802285743

Markus, K. A., & Borsboom, D. (2013). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology*, *31*, 54–64. DOI: 10.1016/j.newideapsych.2011.02.008

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory

structural equation modeling. *Psychological Assessment*, *22*, 471–491. DOI: 10.1037/a0019227

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110. DOI: 10.1146/annurev-clinpsy-032813-153700

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, *19*, 97–112. DOI: 10.1177/1088868314541857

McCrae, R. R., Zonderman, A. B., Costa, P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, *70*, 552–566. DOI: 10.1037/0022-3514.70.3.552

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*, 474. DOI: 10.1037/pspp0000100

Mõttus, R., Sinick, J., Terracciano, A., Hřebíčková, M., Kandler, C., Ando, J., … Jang, K. L. (2018). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*. Advance online publication. http://dx.doi.org/10.1037/pspp0000202

Myszkowski, N., Storme, M., & Tavani, J.-L. (2018). Are reflective models appropriate for very short scales? Proofs of concept of formative models using the Ten-Item Personality

Inventory. *Journal of Personality*. Advance online publication. DOI: 10.1111/jopy.12395

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality*, *59*, 56–68. DOI: 10.1016/j.jrp.2015.09.001

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae: NEO-PI-R ; Manual*. Retrieved from https://pub.uni-bielefeld.de/publication/1878577

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.*, *57*, 401–421. DOI: 10.1146/annurev.psych.57.102904.190127

Pfattheicher, S., Geiger, M., Hartung, J., Weiss, S., & Schindler, S. (2017). Old wine in new bottles? The case of self-compassion and neuroticism. *European Journal of Personality*, *31*, 160–169. DOI: 10.1002/per.2097

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*, 322. DOI: 10.1037/a0014996

Raad, B. D., Perugini, M., Hrebícková, M., & Szarota, P. (1998). Lingua franca of personality: Taxonomies and structures based on the psycholexical approach. *Journal of Cross-Cultural Psychology*, *29*, 212–232.

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*, 203–212. DOI: 10.1016/j.jrp.2006.02.001

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*, 3–25. DOI: 10.1037//0033-2909.126.1.3

Roberts, B. W., Lejuez, C., Krueger, R. F., Richards, J. M., & Hill, P. L. (2014). What is conscientiousness and how can it be assessed? *Developmental Psychology*, *50*, 1315. DOI: 10.1037/a0031109

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1–25. DOI: 10.1037/0033-2909.132.1.1

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*, 27–42. DOI: 10.1177/2515245917745629

Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's Unipolar Big-Five Markers. *Journal of Personality Assessment*, *63*, 506–516. DOI: 10.1207/s15327752jpa6303_8

Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, *31*, 43–53. DOI: 10.1016/j.newideapsych.2011.02.007

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant Colony Optimization and Genetic Algorithm. *PLOS ONE*, *11*, e0167110. DOI: 10.1371/journal.pone.0167110

Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality*. DOI: 10.1002/per.2147

van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B. W., Waldorp, L. J., & Schoevers, R. A. (2015). Association of symptom network structure with the course of depression. *JAMA Psychiatry*, *72*, 1219–1226. DOI: 10.1001/jamapsychiatry.2015.2079.

Van der Maas, H. L., Kan, K. J., Marsman, M., & Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *Journal of Intelligence*, *5*, 16. DOI: 10.3390/jintelligence5020016

Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*, 842–861. DOI: 10.1037/0033-295X.113.4.842

Wiggins, J. S., & Broughton, R. (1985). The interpersonal circle: A structural model for the integration of personality research. *Perspectives in Personality*, *1*, 1–47.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*, 180–198. DOI: 10.1016/j.jrp.2010.01.002

Anlage 1. Erklärung über den Eigenanteil an den veröffentlichten oder zur Veröffentlichung vorgesehenen wissenschaftlichen Schriften innerhalb meiner Dissertationsschrift

Universität Kassel, Fachbereich Humanwissenschaften Erklärung zur kumulativen Dissertationen im Promotionsfach Psychologie

Ergänzung zu § 5a Abs. 4 Satz 1 der Allgemeinen Bestimmungen für Promotionen an der Universität Kassel vom 13. Juni 2011

Antragssteller:

Gabriel Olaru Institut für Psychologie, Universität Kassel

On Measuring Some of the People Some of the Time with Some of the Items: The Search for Stability and Variation in Item Sets

Nummerierte Aufstellung der eingereichten Schriften

1. Olaru, G., Schroeders, U., Hartung, J., & Wilhelm, O. (2019). A Tutorial on Novel Item and Person Sampling Procedures for Personality Research. *European Journal of Personality, 33,* 400-419. DOI: 10.1002/per.2195

2. Olaru G, Schroeders U, Wilhelm O, Ostendorf F (2018). A Confirmatory Examination of Age-associated Personality Differences: Deriving Age-related Measurement-invariant Solutions using Ant Colony Optimization. *Journal of Personality. 86*, 1037–1049. DOI: 10.1111/jopy.12373

3. Olaru, G., Schroeders, U., Wilhelm, O., & Ostendorf, F. (2019). 'Grandpa, do you like roller coasters?': Identifying Age-Appropriate Personality Indicators. *European Journal of Personality, 33*, 264-278. DOI: 10.1002/per.2185

Darlegung des eigenen Anteils an diesen Schriften

Zu Nr. 1. Ich bin Erstautor des Textes. Datenauswertung und wurden vollständig von mir durchgeführt. Ergebnisdiskussion, Erstellung des Manuskripts Literaturrecherche und Programmierung wurden überwiegend von mir durchgeführt und in Teilen von Johanna Hartung, Ulrich Schroeders und Oliver Wilhelm. Die Studienkonzeption wurde in gleichen Teilen von mir, Johanna Hartung, Ulrich Schroeders und Oliver Wilhelm durchgeführt.

Zu Nr. 2. Ich bin Erstautor des Textes. Datenauswertung, Methodenentwicklung und Programmierung wurden vollständig von mir durchgeführt. Ergebnisdiskussion und Erstellung des Manuskripts wurden überwiegend von mir und in Teilen von Ulrich Schroeders, Oliver Wilhelm und Fritz Ostendorf durchgeführt. Die Konzeption der Studie wurde in gleichen Teilen von mir, Ulrich Schroeders und Oliver Wilhelm durchgeführt. Die Datenerhebung wurde vollständig von Fritz Ostendorf durchgeführt.

Zu Nr. 3. Ich bin Erstautor des Textes. Datenauswertung, Methodenentwicklung und Programmierung wurden vollständig von mir durchgeführt. Erstellung des Manuskripts wurden überwiegend von mir und in Teilen von Ulrich Schroeders, Oliver Wilhelm und Fritz Ostendorf durchgeführt. Die Konzeption der Studie und Ergebnisdiskussion wurde in gleichen Teilen von mir, Ulrich Schroeders und Oliver Wilhelm durchgeführt. Die Datenerhebung wurde vollständig von Fritz Ostendorf durchgeführt.

Anlage 2. Dokumentation der Daten

Bei den hier präsentierten Studien handelt es sich um Reanalysen bestehender Datensätze. Für diese Studien wurden keine zusätzlichen Daten erhoben. In der ersten Studie wurden die öffentlich zugänglichen IPIP-NEO-300 Daten von Johnson (2014) verwendet. Diese wurden online über die Webseite http://www.personal.psu.edu/~j5j/IPIP/ erhoben und sind unter https://osf.io/tbmh5/ frei verfügbar. In der zweiten und dritten Studie wurde der deutsche NEO-PI-R Normdatensatz von Ostendorf und Angleitner (2004) verwendet. Dabei handelt es sich um eine Sammlung mehrerer deutschsprachiger NEO-PI-R Datensätze. Der Datensatz wird zurzeit von Fritz Ostendorf verwaltet.

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78–89. DOI:10.1016/j.jrp.2014.05.003

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae: NEO-PI-R ; Manual*. Retrieved from https://pub.uni-bielefeld.de/publication/1878577