

1191²⁰²³

SOEP papers
on Multidisciplinary Panel Data Research

Using Within-Person Change in Three Large Panel Studies to Estimate Personality Age Trajectories

Ingo S. Seifert, Julia M. Rohrer, Stefan C. Schmukle

SOEPPapers on Multidisciplinary Panel Data Research at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPPapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPPapers are available at <http://www.diw.de/soeppapers>

Editors:

Carina **Cornesse** (Survey Methodology)

Jan **Goebel** (Spatial Economics)

Cornelia **Kristen** (Migration)

Philipp **Lersch** (Sociology, Demography)

Carsten **Schröder** (Public Economics)

Jürgen **Schupp** (Sociology)

Sabine **Zinn** (Statistics)

Conchita **D'Ambrosio** (Public Economics, DIW Research Fellow)

Denis **Gerstorf** (Psychology, DIW Research Fellow)

Martin **Kroh** (Political Science, Survey Methodology)

Stefan **Liebig** (Sociology)

David **Richter** (Psychology)

Jörg-Peter **Schräpler** (Survey Methodology, DIW Research Fellow)

Thomas **Siedler** (Empirical Economics, DIW Research Fellow)

C. Katharina **Spieß** (Education and Family Economics)

Gert G. **Wagner** (Social Sciences)

Katharina **Wrohlich** (Gender Economics)

ISSN: 1864-6689 (online)

German Socio-Economic Panel (SOEP)

DIW Berlin

Mohrenstrasse 58

10117 Berlin, Germany

Contact: soeppapers@diw.de



Using Within-Person Change in Three Large Panel Studies to Estimate Personality Age Trajectories

Ingo S. Seifert, Julia M. Rohrer, and Stefan C. Schmukle

Wilhelm Wundt Institute for Psychology, Leipzig University

Date of submission: November 24, 2022

Date first revision submitted: May 19, 2023

Date second revision submitted: July 13, 2023


Accepted: July 24, 2023

Seifert, I. S., Rohrer, J. M., & Schmukle, S. C. (in press). Using within-person change in three large panel studies to estimate personality age trajectories. *Journal of Personality and Social Psychology*.


Supplemental Materials: <https://osf.io/vctnz/>

© 2023, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. The final article will be available, upon publication, via its DOI: 10.1037/pspp0000482

Author Note

Ingo S. Seifert  <https://orcid.org/0000-0002-9798-0003>

Julia M. Rohrer  <https://orcid.org/0000-0001-8564-4523>

Stefan C. Schmukle  <https://orcid.org/0000-0002-6279-9618>

Additional material is provided on the Open Science Framework (OSF) and can be retrieved from <https://osf.io/8rjex/>. On the OSF, we share all the data analytic methods and code

necessary to reproduce our results directly from the original data sets without the need for any additional steps to prepare the data. The data sets used in this study are made available through the respective data-holding institutions. We are not allowed to make the data publicly available, but on the OSF, we provide information about how researchers can request these data for research purposes. On the OSF, for each study, we also share the relevant questionnaires as study materials.

This paper uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey conducted by the Australian Government Department of Social Services (DSS); data from the Socio-Economic Panel (SOEP) made available by the German Institute for Economic Research (DIW); and data from the Longitudinal Internet Studies for the Social Sciences (LISS) Panel administered by CentERdata (Tilburg University, the Netherlands). The LISS panel data were collected by CentERdata through its MESS project funded by the Netherlands Organization for Scientific Research. The findings and views reported in this paper are those of the authors and should not be attributed to the Australian Government, DSS, any of DSS' contractors or partners, DIW, or CentERdata. We are grateful to the Center for Information Services and High Performance Computing (ZIH) at TU Dresden for providing its facilities for high throughput calculations. Computations for this work were done (in part) using resources from the Leipzig University Computing Centre. This research was supported by funding for doctoral candidates from Leipzig University to Ingo S. Seifert.

Correspondence concerning this article should be addressed to Ingo S. Seifert, Wilhelm Wundt Institute for Psychology, Leipzig University, Neumarkt 9–19, 04109 Leipzig, Germany. Email: ingo.seifert@uni-leipzig.de

Abstract

How does personality change when people get older? Numerous studies have investigated this question, overall supporting the idea of so-called personality maturation. However, heterogeneous findings have left open questions, such as whether maturation continues in old age and how large the effects are. We suggest that the heterogeneity is partly rooted in methodological issues. First, studies may have failed to recover age effects, as they did not stringently separate within-person changes from confounding between-person differences. Second, items supposedly belonging to the same trait may show different individual trajectories, thus rendering results sensitive to the specific set of items used. We analyzed panel data from Australia ($N = 15,268$; Study 1), Germany ($N = 22,833$; Study 2), and the Netherlands ($N = 10,163$; Study 3) to investigate age trends in the Big Five on the levels of both scores and items. We applied a fixed effects approach that incorporates only within-person changes over time. Developmental trends in the Big Five scores were generally moderate to large and broadly confirmed personality maturation at younger ages. At older ages, maturation consistently continued for Neuroticism, whereas we found mixed evidence for such changes in Conscientiousness and Agreeableness. Furthermore, in each study, individual items showed age trends that diverged from the rest of the corresponding trait; and these differential patterns could be partly replicated across the three studies. Our results highlight the importance of items in the study of personality development and provide an explanation for previously unaccounted for variability in age trends.

Keywords: personality development, mean-level change, Big Five, panel studies, fixed effects modeling

Using Within-Person Change in Three Large Panel Studies to Estimate Personality Age Trajectories

How does personality change across the life span? Numerous studies have tried to answer the question of how traits develop with age (e.g., Anusic et al., 2012; Ashton & Lee, 2016; Donnellan & Lucas, 2008; Graham et al., 2020; Lucas & Donnellan, 2009, 2011; Marsh et al., 2013; Milojev & Sibley, 2017; Soto et al., 2011; Soto & John, 2012; Specht et al., 2011; Srivastava et al., 2003; Terracciano et al., 2005; Wortman et al., 2012). As summarized in comprehensive meta-analyses by Roberts et al. (2006) and more recently by Bleidorn et al. (2022), there is strong evidence of a pattern in young adulthood referred to as the maturity principle of personality development (e.g., Bleidorn et al., 2021; Caspi et al., 2005; Roberts et al., 2008; Roberts & Nickel, 2017)—that is, with respect to the Big Five personality traits, Neuroticism decreases, whereas Conscientiousness and Agreeableness increase.

Such maturational processes in personality are commonly explained by a social investment perspective (Bleidorn et al., 2013; Roberts et al., 2005; for a different account, see, e.g., McCrae et al., 2000; see also Specht et al., 2014). In adulthood, individuals are expected to commit to normative-conventional roles (e.g., as a romantic partner, parent, or employee; Lodi-Smith & Roberts, 2007) that, in turn, elicit personality development. However, the principle of social investment has been challenged, as the expected changes in personality after such role transitions have not emerged consistently across empirical studies (e.g., Asselmann & Specht, 2021; den Boer et al., 2019; Denissen et al., 2019; Deventer et al., 2019; Krämer et al., 2022; Specht et al., 2011; van Scheppingen et al., 2016; for an overview, see Bleidorn et al., 2018).

Further, whether or not personality maturation ends at a certain age remains unclear. In general, the maturation process is assumed to occur between young and middle adulthood (e.g.,

Bleidorn et al., 2013), mirroring the predominantly investigated age groups in this field of research (Bleidorn et al., 2021; Roberts et al., 2006). Changes in personality past middle adulthood are viewed as “very gradual and modest” (Costa et al., 2019, p. 430). Nonetheless, a continuing pattern of maturation has sometimes been found up to old age (e.g., Lucas & Donnellan, 2011; Wortman et al., 2012)—but results have been rather heterogeneous, with some studies even reporting the opposite age trend after middle adulthood (de-maturation; i.e., increases in Neuroticism as well as decreases in Conscientiousness and Agreeableness; e.g., Graham et al., 2020; Marsh et al., 2013).

Beyond these inconsistencies, studies have also disagreed about the *magnitude* of maturational change, with some even failing to support the predicted maturational development of personality in single dimensions. For example, Lucas and Donnellan (2011) and Specht et al. (2011) found a rather flat age trajectory for Neuroticism instead of the expected decrease. Similarly, Lucas and Donnellan (2011) reported that Agreeableness remained stable throughout young and middle adulthood with increases happening only at the oldest ages. Challenging the maturity principle even further, some studies have found maturity-*opposed* age trends in large age-heterogeneous samples: an increase in Neuroticism (Donnellan & Lucas, 2008), a decrease in Conscientiousness (Graham et al., 2020), and a decrease in Agreeableness (Milojev & Sibley, 2017; Specht et al., 2011).

Uncertainty regarding the development of personality with age exists not only for Neuroticism, Conscientiousness, and Agreeableness but also for Extraversion and Openness. Within the framework of the maturity principle, these two traits have often been somewhat neglected. In general, they are supposed to show less pronounced developmental patterns (Bleidorn et al., 2013). But, again, heterogeneous trends have been reported for both traits across

studies. For Extraversion, some studies found no or only very small age-related changes across adulthood (e.g., Soto et al., 2011), whereas other studies reported a pronounced downward trend that was even comparable in magnitude to the changes in the maturity-related traits (e.g., Marsh et al., 2013). Different aspects of Extraversion were found to show different developmental patterns that could potentially explain some of the heterogeneity in the reported age trends. The Social Dominance facet of Extraversion showed marked age-related increases, whereas trajectories for the Social Vitality facet were rather flat (Roberts et al., 2006). For Openness, some studies reported an increase across adulthood (e.g., Soto et al., 2011), whereas other studies found a decrease (e.g., Wortman et al., 2012). The exact definition of Openness is controversial, as the content of this trait tends to vary by theory and questionnaire (John, 2021; Schwaba et al., 2018; Schwaba, 2019; see also Saucier, 1992), pointing to the possibility that varying conceptualizations of Openness might explain diverging developmental trends across studies (Costa et al., 2019).

In summary, there is general agreement that personality changes across the life span. Meta-analyses have quantified these changes as medium to large, with effects of about 0.5 to 1 standard deviation across traits (e.g., Bleidorn et al., 2022; Roberts et al., 2006). But beyond this general agreement that personality matures in some way, the specific age trajectories reported for personality across studies vary considerably. This view was corroborated by Bleidorn et al.'s (2022) meta-analysis, which found for mean-level personality trends substantial between-study heterogeneity that has remained largely unexplained. One explanation for heterogeneity in mean-level trends is of course true substantive variability—maybe personality development *does* look different in different places at different times. However, before such

substantive variability can be concluded, the possibility that heterogeneous findings result from heterogeneous methodologies should be ruled out.

Methodological Considerations of Age Differences in Mean Levels of Personality

The literature on personality change across the life span is large. Therefore, and perhaps unsurprisingly, different methodological approaches and various statistical models have been applied. Accordingly, differences in the methodologies may explain at least some of the uncertainty in the age trajectories (for a similar argument, see Costa et al., 2019). Here, we focus on two key aspects: the successful identification of age effects as well as measurement invariance in personality. Failing to address these issues appropriately can introduce biases into individual studies, and importantly, these issues cannot be resolved simply by collecting more extensive data or by pooling estimates in meta-analyses.

Identifying Age Effects on Personality

A central challenge in the study of personality development is determining how to effectively separate actual age effects from cohort effects, which may confound any associations with age observed in the data. In purely cross-sectional designs (and regardless of sample size; e.g., Soto et al., 2011), age effects are confounded with cohort effects: A person who is 10 years older was born 10 years earlier. This confounding is a substantial problem because there is more and more evidence that year of birth affects personality (e.g., Brandt et al., 2022; Jokela et al., 2017; Smits et al., 2011; see also Hülür, 2017). In an intriguing study covering nearly 80% of the male Finnish population born between 1962 and 1976, Jokela et al. (2017) found cohort effects on personality traits comparable in magnitude to the so-called *Flynn effect*, the well-known increase in cognitive abilities across cohorts (Flynn, 1984, 1987). Across traits, the cohort differences in personality reported by Jokela et al. (2017) amounted to small-to-medium effects

(with changes ranging from 0.2 to 0.6 standard deviations relative to the earliest year of birth).

Age trajectories that are based on cross-sectional data are therefore an indistinguishable blend of such cohort effects and the age effects of interest, a fact that has been widely acknowledged by the authors of the affected studies (e.g., Donnellan & Lucas, 2008; Soto et al., 2011).

Thus, a priori, we cannot assume that age trajectories that were estimated on a single wave of cross-sectional data will recover the age effects of interest. In principle, age effects are conceptualized on the within-person level, that is, on an individual level, they capture the contrast between a person at a given age and the same person at a different age; on the population level, they capture the average of such individual-level age effects. Although in principle, it is possible to recover such average age effects from a repeated cross-sectional design (see Fitzenberger et al., 2022; Ion et al., 2022), longitudinal data are the preferred approach. With the help of longitudinal data, researchers can compare the same person at different ages. This comparison is still not guaranteed to recover the age effects of interest, as age is confounded with period: The same person 10 years older is observed 10 years later, and historical circumstances can change a lot in that time. In principle, the age-period-cohort conundrum cannot be solved by simply adding more (longitudinal) data or more complex statistical models (see Glenn, 2003; see also A. Bell & Jones, 2014). But an age trajectory that is pieced together from purely within-person comparisons is no longer confounded by cohort effects, and thus, fewer assumptions are necessary to interpret it as a reflection of actual age effects.

However, just because a study uses longitudinal data does not guarantee that its results exclusively rely on within-person comparisons. The data simultaneously contain between-person and within-person information, and depending on the statistical model employed, both can affect the estimated age trajectories. In studies on personality development, longitudinal data are

typically analyzed by means of latent growth modeling (LGM; e.g., Specht et al., 2011) or multilevel modeling (MLM; e.g., Graham et al., 2020; Terracciano et al., 2005). Both approaches often have limitations in practice.

In the case of LGM, age trends in the intercepts (representing between-person differences) and slopes (representing within-person changes) are required to follow a function that is either prespecified (e.g., linear, quadratic, or cubic) or estimated on the underlying data (so-called latent basis modeling). The resulting age trajectories, though, might be substantially biased if individuals' age trends deviate from the functional form (Wu & Lang, 2016). Thus, when there are interindividual differences in the intraindividual change patterns of personality, LGM will not necessarily successfully recover the average age trajectory.¹

In MLM, usually parametric functions reflecting effects of age on personality are included in the model with random intercepts and random slopes for individuals (e.g., Graham et al., 2020; Terracciano et al., 2005). However, random intercepts are not sufficient to fully control for between-person differences (Townsend et al., 2013), and thus, estimates in the typical implementation of MLM represent an indistinguishable blend of within-person and between-person information (Curran & Bauer, 2011). Therefore, age trajectories may still be confounded with cohort effects. This issue is especially relevant if analyses explicitly include cases that provide only cross-sectional information (e.g., Graham et al., 2020) and the extent to which developmental trajectories represent actual age effects remains unclear.

By contrast, fixed effects modeling (FEM) allows researchers to extract within-person information that is completely separated from between-person variation in the data (Boyce et al.,

¹ In addition, with respect to the presentation of the results, the slopes in LGM represent only *piecewise* changes in personality as a function of age (see, e.g., Specht et al., 2011; see also Lucas & Donnellan, 2011; Wortman et al., 2012). To obtain a cohesive mean-level trajectory across the life span, the slopes would need to be integrated, which is not typically done in this research tradition.

2013; for a general introduction to FEM, see Allison, 2009; Angrist & Pischke, 2009; McNeish & Kelley, 2019). As only within-person *changes* are captured, all time-constant factors (including between-person confounds) are partialized out (Brüderl & Ludwig, 2015; McNeish & Kelley, 2019). Thus, FEM is a good match for research on developmental trajectories of personality, where within-person changes with age are of interest, and so-called between-person effects, such as the confounding influence of cohort, are considered nuisances. Although FEM is widely used in the econometric literature, it is rarely applied in psychological research in its original form (McNeish & Kelley, 2019).

However, LGM, MLM, and FEM do not necessarily represent distinct classes of statistical approaches (Zyphur et al., 2020). LGM and MLM can be translated into each other so that they lead to identical results (see, e.g., Curran, 2003). Similarly, it is possible to translate FEM into the LGM framework (Allison, 2009; Allison et al., 2017; Andersen, 2022). And after person-mean centering, estimates from MLM are identical to FEM (A. Bell & Jones, 2015; Hamaker & Muthén, 2020; McNeish & Kelley, 2019). Despite these statistical overlaps, a certain approach may be more suitable than the others for addressing a particular research question. For example, when data cover the life span, LGM is cumbersome to implement because each year of age results in an additional observed variable (see Curran, 2003). Similarly, to obtain within-person estimates for many predictor variables in the MLM framework (e.g., when age-dummies are used for life span data, as in the present studies), repeated person-mean centering is necessary, whereas FEM provides such estimates without any additional preparation (for an illustration, see <https://osf.io/uycx6/> and Table S21 in the Supplemental Materials). This extra step may be one of the reasons why MLM with person-mean centering has not yet been used (as

far as we know) to study the effects of age on personality, despite multiple articles advocating for its use (Hamaker & Muthén, 2020; Wang & Maxwell, 2015).

Even if within-person trends are correctly separated from between-person trends, additional methodological challenges come along with the use of longitudinal data to examine age differences in personality (see also Kratz & Brüderl, 2021). Issues such as selective dropout and panel conditioning effects are additional sources of bias. For example, researchers should avoid excluding individuals with incomplete longitudinal data, as this might lead to an overly healthy sample (Seifert et al., 2022), which in turn means that average age trajectories of personality will be biased because changes in health are correlated with changes in personality (which might particularly be an issue for older age groups; Kornadt et al., 2018; Luo et al., 2022; Mueller et al., 2018). Further biases are possible. For example, studies have suggested that there may be an initial elevation bias (Anvari et al., 2022; Arslan et al., 2021; Shrout et al., 2018; but cf. Cerino et al., 2022): Participants seemingly overreport constructs at their first assessment in longitudinal studies. Such a bias would in turn distort age trajectories because participants will be youngest at this (biased) initial assessment.

Measurement Invariance

A second key methodological aspect of the mean-level development of personality is whether individual items that supposedly measure the same personality dimension follow the same trend. This question is usually addressed from the perspective of measurement invariance (Vandenberg & Lance, 2000; see also Nye et al., 2016): Is the latent personality construct comparable across age, so that observed age differences can be attributed to the underlying construct rather than to the particular indicators? Because we were interested in mean-level changes, there is a need to demonstrate metric invariance (i.e., assuming equal item factor

loadings across age) and additionally scalar invariance (i.e., assuming equal item intercepts across age) in personality across age in order to draw valid conclusions (Widaman et al., 2010).

Even though measurement invariance is such a crucial prerequisite, it is often not stringently tested in studies on personality development. According to a recent meta-analysis on personality development, tests of measurement invariance were part of only approximately 30% of the studies (Bleidorn et al., 2022). When measurement invariance is examined, metric invariance is usually supported; but for scalar invariance, results are less unequivocal, as much more marked losses in model fit are typically observed (e.g., Lucas & Donnellan, 2011; Milojev & Sibley, 2017; Wortman et al., 2012; see also Brandt et al., 2020). Researchers sometimes argue that measurement invariance is given when the model assuming scalar invariance fits the data sufficiently well (e.g., Specht et al., 2011). However, such a conclusion is not necessarily valid because a good overall model fit may still be observed even when the model fit decreases drastically after equality constraints are included for item intercepts, calling scalar invariance into question.

What does a lack of measurement invariance across age mean on a substantive level? A lack of *metric* invariance (i.e., varying item factor loadings) indicates that the same item indicators represent the personality trait to varying degrees across the life course. If some commensurable construct exists across the life course, the implication would then be that some indicators would become more or less suitable, thus pointing to measurement issues (e.g., partying might no longer be a good indicator of Extraversion in old age). Taking the items at face value and insisting that they do represent the construct would instead imply that the meaning of the construct changes with age (e.g., partying may measure something different than

Extraversion in old age). In any case, a lack of metric invariance precludes straightforward substantive comparisons.

A lack of *scalar* invariance (i.e., varying item intercepts) indicates that age changes in the latent trait alone are not sufficient for explaining age changes in the respective items. Thus, individual items do not follow the age trajectory of the latent factor in proportion to their respective factor loading—instead, their age trajectory may be steeper or flatter or may exhibit a different shape altogether. Again, such differences can be interpreted in different ways.

Assuming that all age effects on items must be mediated through some latent trait, differences in the item trends may point to subfactors that show different age trajectories (e.g., the splitting of Extraversion into the Social Dominance and Social Vitality facets; Roberts et al., 2006).

Focusing instead on the items, different trends may simply suggest that beyond the age changes in the latent trait, something else induces age changes in the items (e.g., partygoing may decline with age, regardless of Extraversion, because there are fewer opportunities or because of health issues). In other words, not all age changes in the items are mediated by the latent construct (Borsboom, 2023; Paulewicz & Blaut, 2022). In any case, a lack of scalar invariance indicates that considering only developmental trends on the trait level misses part of the picture. Thus, age trends should be examined on the subordinate item level where trajectories are likely to differ (Marsh et al., 2013; Mõttus et al., 2015; Wicherts & Dolan, 2010).

Previous studies have shown that when different items are used to measure a particular Big Five personality trait, these different items can show heterogenous mean-level patterns with age (e.g., Lucas & Donnellan, 2009; Mõttus & Rozgonjuk, 2021), although personality development is typically examined only for mean scores (e.g., Graham et al., 2020). Similarly, when items that cover similar aspects were bundled into facets, distinct developmental trends

were found for the facets of a trait (e.g., Jackson et al., 2009; Roberts et al., 2006; Schwaba et al., 2022; Soto et al., 2011; Terracciano et al., 2005). Therefore, developmental trends in the mean levels of the Big Five may depend on which personality items were selected. The fact that the same personality trait is often measured with different items in different studies may at least partially explain why developmental patterns vary considerably across studies.

The Present Studies

Numerous previous investigations have resulted in partially compatible, partially contradictory conclusions regarding the mean-level development of personality. The extent to which heterogeneity in results can be attributed to genuine substantive differences in personality development as opposed to differences in methodology, models, and measures remains unclear. To clarify the situation, we analyzed and contrasted results across three studies, consistently applying what we consider the most sensible model specifications and explicitly taking into account questions of measurement invariance.

We relied on three large panel studies: the Household, Income and Labour Dynamics in Australia (HILDA; Study 1) Survey; the German Socio-Economic Panel (SOEP; Study 2); and the Dutch Longitudinal Internet Studies for the Social Sciences (LISS; Study 3) Panel. These data offered several advantages: First, they provided longitudinal personality data that spanned up to 14 years. Second, the large sample sizes allowed us to detect even small changes with adequate precision. Third, the participants were diverse with respect to age and other characteristics (e.g., education, income, and marital status), such that we were able to estimate age trends that covered the entire adult life span and were representative of a broader population. To overcome the analytical limitations of previous studies, we used fixed effects modeling (FEM) with age dummies to obtain flexible nonparametric developmental trajectories

that were based solely on within-person changes (thus not confounded with between-person differences in personality) and could recover the average trajectory even in the presence of heterogeneous development. We furthermore explicitly tested for measurement invariance, and because we detected relevant violations, we investigated item-level trajectories. Finally, to get a better understanding of the extent to which separating within-person change from between-person differences matters for substantive conclusions, we additionally ran standard multilevel modeling (MLM) analyses and contrasted the results with our main findings. These results are reported after Study 3.

In each study, personality was assessed with the Big Five traits (i.e., Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness). Each panel study covered 12 to 14 years of personality development. In total, we analyzed longitudinal data from more than 48,000 individuals. Analyzing three samples from different countries allowed us to assess the robustness of our results.

General Statistical Approach

The HILDA (Study 1), SOEP (Study 2), and LISS (Study 3) studies share essential features: all three are panel studies with Big Five personality measures. We kept the analysis strategies across all studies as parallel as possible to maximize comparability. Accordingly, we first present the general methodological approach and later explain some study-specific points. For each of the three studies, informed consent was obtained from participants by the respective responsible institution. Ethical approval was not required for our research, as we analyzed existing and fully anonymized data.

Fixed Effects Modeling

To estimate mean-level age trajectories of personality that were informed solely by within-person information, we applied fixed effects modeling (FEM; Allison, 2009; Brüderl & Ludwig, 2015; Hamaker & Muthén, 2020; McNeish & Kelley, 2019). We ran separate analyses for each dependent variable, that is, for each of the Big Five mean scores but also for every single item in isolation. Reversed items were recoded so that higher values indicated higher construct expression. Following recommendations (Brüderl & Ludwig, 2015; Cameron & Miller, 2015), we used panel-robust standard errors for FEM (as proposed by Arellano, 1987) to account for serial correlation and heteroscedasticity.

We used dummy-coded age (separate dummies for each year of age, with the youngest age as the reference group) as the focal predictor of interest. As FEM is based on within-person variation only, all time-constant variables (e.g., cohort and sex, but also unobserved characteristics) are implicitly controlled for. However, time-varying variables may still bias the results; therefore, we included additional controls (in the General Discussion section, we consider the conditions under which FEM can recover unbiased age trajectories). First, we added a dummy variable that indicated the participants' first wave of personality data (1 = first wave, 0 = all other waves). We did so to control for a potential initial elevation bias (Anvari et al., 2022; Arslan et al., 2021; Shrout et al., 2018), which is a tendency to overreport a certain construct at the first assessment in longitudinal surveys and which may bias age trajectories.

Second, we controlled for changes in the mode of data collection within participants (self-report questionnaire vs. interview) that occurred for roughly 20% of the participants in Study 2 (SOEP). Interviews have been reported to be more susceptible to response biases than self-report questionnaires (e.g., social desirability; Richman et al., 1999; see also Ausmees et al.,

2022), leading to personality levels that vary by the mode of data collection (e.g., more socially desirable personality ratings in interviews; for an examination with SOEP data, see Donnellan & Lucas, 2008; Hilgert et al., 2016; Lang et al., 2011). Thus, because changes in mode might bias the age trajectory of personality, in Study 2, we additionally included dummy-coded information about the mode for each wave (1 = self-report questionnaire, 0 = interview). Mode changes were not an issue in Study 1 (HILDA) or Study 3 (LISS), as personality was always assessed with self-report questionnaires. These control variables were included in the FEM without explicitly modeling person-specific deviations from the average effect, a practice that is sufficient for controlling for potential confounds of the average age effects (but for how to relax this assumption in FEM, see Rüttenauer & Ludwig, 2023).

To facilitate the interpretation of the results across studies, the dependent variables were transformed into *T*-scores before being entered into the models (see, e.g., Donnellan & Lucas, 2008). For this procedure, we standardized the data to obtain a mean of 50 and a standard deviation of 10 at the participants' second observation (because the first observations were potentially distorted by an initial elevation bias; e.g., Shrout et al., 2018). Accordingly, the estimated within-person changes in personality are interpreted relative to the personality differences across participants. We applied Cohen's (1988) conventional criteria to interpret the effect sizes and determined that a difference of 2 *T*-score units equaled a small effect, a difference of 5 units a medium effect, and a difference of 8 units a large effect. We primarily use these labels to quantify the maximal differences in mean scores observed across the life span; when comparing our results with other studies, it should be kept in mind that the magnitude of change needs to be considered relative to the time span that was covered (e.g., a change of

2 *T*-score units across 1 year is much more dramatic than a change of 2 *T*-score units across 10 years; see Funder & Ozer, 2019).

As FEM uses only longitudinal information to make estimates, only participants who answered all the items for the same personality trait in at least two waves of measurement were included in the analyses. However, beyond this minimal requirement, participants with missing data at some measurement occasions were still included. Due to the sparse number of participants at the oldest ages, we limited the estimation to years of age for which we had at least 30 observations in order to obtain sufficiently precise estimates (for similar restrictions, see, e.g., Donnellan & Lucas, 2008; Specht et al., 2011). Inclusion criteria were applied separately for the Big Five dimensions, potentially resulting in different sample sizes for different traits. In Study 2 (SOEP), observations with missing information on the mode of data collection were excluded, which applied to roughly 0.5% of observations from participants with at least two measurements (with a total of 74,128 to 74,714 observations, depending on the trait). We used an ordinary least squares (OLS) approach to obtain the parameter estimates, which is standard for FEM (A. Bell & Jones, 2015; Brüderl & Ludwig, 2015; McNeish & Kelley, 2019).

Tests of Measurement Invariance

In order to be able to interpret age trends in personality mean scores as age trends in the underlying personality traits, measured constructs have to be comparable across age. Accordingly, there is a need to establish scalar measurement invariance, which is a prerequisite for meaningful mean-level comparisons of latent constructs (e.g., Widaman et al., 2010; see also Meredith, 1993). A lack of scalar measurement invariance suggests that items are showing heterogeneous age trends (Möttus et al., 2015).

In our case, invariance should be tested across both years of measurement and years of birth, as age is defined by both (age = date of measurement – date of birth). We used a structural equation modeling (SEM) framework to test for metric and scalar measurement invariance simultaneously across years of measurement and years of birth. Specifically, we used local structural equation modeling (LSEM; Hildebrandt et al., 2009, 2016; Olaru, Schroeders, Hartung, & Wilhelm, 2019), as this technique allowed us to *continuously* estimate parameters across birth years. Therefore, we did not need to artificially categorize the birth year variable (as would be necessary for multiple-group SEM; see, e.g., Lucas & Donnellan, 2011; Wortman et al., 2012), thus avoiding the accompanying loss of information (MacCallum et al., 2002). For each value of the continuous variable (in LSEM terminology, for each focal point), the sample is weighted: Participants with a birth year that matches the focal point of interest receive the largest weight (i.e., 1); the more a participant's birth year deviates from the focal point, the smaller the weight. We used a Gaussian kernel function, which is commonly used so that the weights around a focal point follow a normal distribution. The standard deviation of this distribution (referred to as bandwidth) regulates how influential the data points surrounding a focal point are; we used a bandwidth factor of $h = 2$, which is recommended (Hildebrandt et al., 2016; Olaru, Schroeders, Hartung, & Wilhelm, 2019) and most commonly used (e.g., Gnams & Schroeders, 2020; Olaru & Allemand, 2022; Seifert et al., 2022; Wagner et al., 2019).

For the measurement invariance analyses, we included the same observations as in FEM analyses (for an exception, see Study 3). To avoid unstable parameter estimates and estimation issues (e.g., nonpositive definite covariance matrices or Heywood cases), we restricted the range of focal points (i.e., birth years) so that a minimum of $n = 30$ participants were continuously observed across the whole span of focal points. Participants with birth years outside this range of

focal points still contributed to the estimation (with weights corresponding to their distances from the focal points of interest). We used a joint estimation approach, which allowed us to implement invariance constraints across birth years, and a full information maximum likelihood (FIML) estimator to include participants with missing data.² The analyses were performed separately for Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness.

To test for measurement invariance, we estimated and compared models with increasingly strict equality constraints. In a baseline model, each wave of measurement was represented as a separate latent factor. The number of manifest indicators per factor ranged from three to six items, depending on trait and study. All latent factors were allowed to covary with each other. Residual variances of identical items were allowed to covary across time (see, e.g., Little, 2013). This model was estimated for each focal point (i.e., each birth year) in LSEM. First, to test for *metric measurement invariance*, we restricted the baseline model by placing invariance constraints on the factor loadings simultaneously across years of measurement and years of birth. Then, to test for *scalar measurement invariance*, we additionally constrained the item intercepts to be equal across both years of measurement and years of birth.

The comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR) were obtained as common model fit indices (e.g., Kline, 2016) across the range of focal points. It is common practice to evaluate fit indices with Hu and Bentler's (1999) quasi-canonical cutoffs ($CFI \geq .95$; $RMSEA \leq .06$; $SRMR \leq .08$), albeit this practice is not without criticism (e.g., McNeish & Wolf, 2023). To test for

² In particular, FIML estimation allowed us to include participants who took part in some but not all waves of measurement in the measurement invariance analyses. These participants were also included in the FEM analyses without the need to apply any specific estimator because the underlying data table is structured in the long format (i.e., each measurement wave is a separate row, and only rows with missing variables are excluded).

measurement invariance, we evaluated differences in the fit indices between consecutive models. We primarily used the Cheung and Rensvold (2002) criterion of $\Delta CFI = .01$ to probe for metric and scalar measurement invariance, which is arguably the most widely used criterion, although recommendations disagree about the cutoff values and even about which fit indices should be used (e.g., Chen, 2007; Meade et al., 2008; Rutkowski & Svetina, 2014; see also Putnick & Bornstein, 2016). If measurement invariance is not given, single items may exhibit unique age trends and, thus, they might not be meaningfully represented by a composite (Marsh et al., 2013; Meredith & Horn, 2001). To examine the substantive impact of this potential developmental heterogeneity, differences across the age trends of individual items should be inspected when measurement invariance cannot be established (see, e.g., Möttus et al., 2015).

Additional Analyses

Comparing Fixed Effects Modeling With Standard Multilevel Modeling

With the present FEM approach, we relied exclusively on within-person changes to estimate age trends in personality. However, research has yet to determine the extent to which these trajectories are actually distinct from the more common analyses of longitudinal data with multilevel modeling (MLM) in studies on personality development (e.g., Graham et al., 2020; Terracciano et al., 2005) that incorporate both within-person *and* between-person information in the estimates (Curran & Bauer, 2011). To investigate the matter, we reanalyzed our data for the Big Five scores by replacing FEM with MLM.

As is typically done to analyze longitudinal data, we implemented MLM with random intercepts for individuals without person-mean centering (e.g., Graham et al., 2020; Terracciano et al., 2005). The methodology was kept as close as possible to our main analyses with FEM: First, we used the same individuals and observations. Second, we included age as a categorical

variable in the regression models. Third, we added the same control variables to the models (i.e., dummy-coded first wave [Studies 1 to 3] and response mode [Study 2]), without random slopes for individuals as in the FEM analyses. The results of these analyses are presented after Study 3.

Quantifying the Similarity of Item Trajectories Across Studies

To quantify the similarity of the item trends (within and across the three studies), we calculated the Fréchet distance (Fréchet, 1906; see also Alt & Godau, 1995; Genolini et al., 2016) as a descriptive measure for each pair of items that were used to measure the same personality trait. We explain and report these analyses in more detail in the Supplemental Materials (see Figures S20 to S24).

Within-Person Factor Structure

Our FEM approach focuses on *intraindividual* information; however, the Big Five were derived in a way that they primarily describe *interindividual* differences (e.g., Beck & Jackson, 2020; John, 2021). Thus, the question arises whether they are a suitable taxonomy for investigating within-person changes. To address this concern, we tested whether the same factor structure of personality held on the between-person *and* within-person levels using multilevel confirmatory factor analysis (MLCFA; Geldhof et al., 2014; Muthén, 1994; for more details, see the Supplemental Materials).

When we analyzed the *individual* Big Five dimensions in separate MLCFA models, an identical factor structure for the between-person and within-person levels was generally supported in all three studies (see Table S19 in the Supplemental Materials). However, when we analyzed all five traits in a combined model to test the five-factor structure with MLCFA, the fit of the models was unsatisfactory in all three studies (see Table S20 in the Supplemental

Materials; for similar results, see Grosz, 2020). MLCFA simultaneously tests the factor structure on the between-person *and* within-person levels, so misfit could result from either level. We thus conducted separate tests of the five-factor structure for both the between-person and within-person levels. On the between-person level, the model fit was not satisfactory, whereas on the within-person level, the fit was overall better, albeit not entirely satisfactory (see Table S20 in the Supplemental Materials). In particular, analyzing the levels separately did not markedly improve the fit. This indicates that the source of model specification lies within the levels (particularly in the between-person level) rather than in differences between the levels.

These results, indicating a lack of strong support for the Big Five factor structure, are not very surprising given that previous confirmatory tests of the Big Five factor structure in cross-sectional data typically failed (Marsh et al., 2010). However, the finding that the Big Five factor structure seemed to fit relatively better on the within-person level was unexpected given that these personality traits have been derived to describe differences *between* individuals (e.g., Beck & Jackson, 2020; John, 2021). Taken together, to the extent that the Big Five are considered a suitable description of *interindividual* differences, they may be considered an at least equally suitable description of *intraindividual* differences, at least in the context of our three studies.

Transparency and Openness

All analyses were performed in R (Version 4.1.2) using the RStudio environment (Version 2022.02.0+443). For FEM, we used the plm package (Version 2.6-0; Croissant & Millo, 2008); and for LSEM, we used the sirt package (Version 3.12-41; Robitzsch, 2022). A comprehensive list of all the R packages we used can be found on the Open Science Framework (OSF) at <https://osf.io/8rjex/>.

All analysis scripts for all studies are publicly available on the OSF. Due to the data protection policies of the responsible institutions, we are not allowed to share the data sets we used in our studies. However, all data sets (i.e., the HILDA, SOEP, and LISS data) can be requested directly from the responsible institutions (for details on data access, see the OSF). Our analysis scripts include all the steps needed to prepare the data, and thus, any researchers with access to the original data sets should be able to reproduce our findings. Further, for each study, we share the relevant questionnaires as study materials on the OSF.

Our analyses were not preregistered, but to replicate our findings, the same methodology was applied across three independent data sets, severely restricting room for any “favorable” result-driven analytical decisions (whatever might be considered “favorable” in the case of primarily descriptive age trends in personality). Adhering to the Journal Article Reporting Standards (JARS; Appelbaum et al., 2018; Kazak, 2018), we report how we determined our sample size, all data exclusions (if any), all manipulations (not applicable), and all measures in the studies.

Study 1: Household, Income and Labour Dynamics in Australia (HILDA) Survey

Method

Design and Participants

The HILDA Survey is an Australian panel study covering a broad range of individual, social, and economic topics. The first wave of measurement was initiated in 2001 with a large national probability sample of Australian households, and additional waves were conducted annually. In 2011 (11th measurement wave), the original sample of households was replenished with a top-up sample of additional households (Watson & Wooden, 2013). In 2005, 2009, 2013, and 2017, each household member who was at least 15 years of age was asked to fill out a

personality questionnaire. For a general introduction to the HILDA Survey, see the article by Watson and Wooden (2012; see also Summerfield et al., 2018; Watson & Wooden, 2021).

The HILDA data are available for research purposes—and have frequently been used in various research fields. In personality psychology, several studies have already examined the mean-level development of personality across the life span with HILDA data (Lucas & Donnellan, 2009; Wortman et al., 2012; see also Cobb-Clark & Schurer, 2012). However, these studies included only one or two waves of personality data and did not use FEM as the data-analytical approach. In the present study, we used General Release 17.0 of the HILDA data (Department of Social Services & Melbourne Institute of Applied Economic and Social Research, 2018). A list of publications using the HILDA data is available at <https://melbourneinstitute.unimelb.edu.au/hilda/publications>.

Overall, our analyses included $N = 15,268$ participants. An average of 3.02 measurement points of personality ($SD = 0.89$) were included per participant. Participants' mean age (across the measurement waves included for the respective participant) was 45.22 years ($SD = 18.07$). Based on our selection criteria, the youngest age we included was 15, and the oldest was 91. The proportion of female participants was 53.35%. As we performed the analyses separately for the Big Five, there was some variation in sample size across traits ranging from 15,009 to 15,168 (see Table 1; for trait-specific sample descriptions, see Table S2 in the Supplemental Materials).

Measures

In 2005, 2009, 2013, and 2017, the self-completion questionnaire contained an identical list of 36 adjectives to assess personality with items measuring the Big Five traits. Thirty of these items were a selection of Saucier's (1994) Mini-Markers, which, in turn, are a subset of Goldberg's (1992) items (for more details, see Losoncz, 2009). Seven items each were intended

to measure Neuroticism, Conscientiousness, Agreeableness, and Openness; eight to measure Extraversion. Participants were asked: “How well do the following words describe you?” and could respond on a 7-point scale ranging from 1 (*does not describe me at all*) to 7 (*describes me very well*).

However, the full list of adjectives frequently exhibited unsatisfactory psychometric properties, and a reduction in the number of items has been recommended and frequently used (e.g., Losoncz, 2009; Seifert et al., 2022; Summerfield et al., 2018; Wortman et al., 2012). Accordingly, for the present analyses, we used a subset of items consisting of four items for Extraversion; five each for Conscientiousness, Agreeableness, and Openness; and six for Neuroticism. This subset is identical to Seifert et al. (2022) where semantic (e.g., avoiding redundant items) and psychometric criteria were applied to reduce the items (for a detailed description of the selection procedure, see Seifert et al., 2022, Supplemental Materials). For transparency and comprehensiveness, we additionally present results for the excluded items in the Supplemental Materials in Table S4 and Figures S1 to S5.

In the present sample, satisfactory internal consistencies were found across waves; for Neuroticism ($\alpha = .76$ to $.77$; $\omega = .80$ to $.82$), Extraversion ($\alpha = .67$ to $.70$; $\omega = .69$ to $.72$), Conscientiousness ($\alpha = .78$; $\omega = .80$ to $.81$), Agreeableness ($\alpha = .75$ to $.76$; $\omega = .81$ to $.84$), and Openness ($\alpha = .70$ to $.73$; $\omega = .74$ to $.75$; for wave-specific results, see Table S3 in the Supplemental Materials).

Results

Measurement Invariance

We applied the LSEM framework to test for measurement invariance across years of measurement and years of birth (because both dimensions define years of age). Such invariance is a prerequisite for valid interpretations of age trends in personality mean scores. The results of these analyses are presented in Table 1.

In general, model fit was only negligibly impaired when the factor loadings were set to equality (i.e., assuming metric invariance) and, across the Big Five, ΔCFI was consistently below .01. Accordingly, the results provided evidence for metric invariance in personality in HILDA. However, additionally imposing equality constraints on the item intercepts (i.e., assuming scalar invariance) yielded a much more marked loss in model fit. The fit worsened, especially for Openness and Agreeableness (with a ΔCFI of around .02), and to a lesser extent, for Extraversion and Neuroticism (with a ΔCFI of around .01). Hence, unlike metric invariance, scalar invariance could not be established in HILDA.

This result implies that different items may exhibit different mean-level age trends; and this developmental heterogeneity may be missed when focusing solely on the age trend in the mean score. Thus, in the following FEM analyses, we contrasted the results for the mean scores with the results for the single items.

Table 1*Fit Indices for Testing for Measurement Invariance (MI) for the Big Five in the Household,**Income and Labour Dynamics in Australia (HILDA) Survey*

Model	<i>n</i>	CFI	RMSEA	SRMR
Neuroticism	15,083			
Baseline		.977	.027	.033
Metric MI		.973	.028	.040
Scalar MI		.960	.033	.044
Extraversion	15,014			
Baseline		.988	.030	.025
Metric MI		.986	.030	.031
Scalar MI		.976	.037	.043
Conscientiousness	15,075			
Baseline		.968	.038	.037
Metric MI		.966	.037	.041
Scalar MI		.959	.039	.042
Agreeableness	15,168			
Baseline		.981	.029	.037
Metric MI		.979	.029	.042
Scalar MI		.959	.038	.052
Openness	15,009			
Baseline		.982	.030	.030
Metric MI		.980	.030	.035
Scalar MI		.956	.042	.048

Note. Personality was assessed in 2005, 2009, 2013, and 2017. For each trait, focal points

(representing birth years) ranged from 1926 to 1998. Metric MI requires the same factor loadings across both years of measurement and years of birth. Scalar MI additionally requires the same item intercepts across both years of measurement and years of birth. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

Age Trajectories

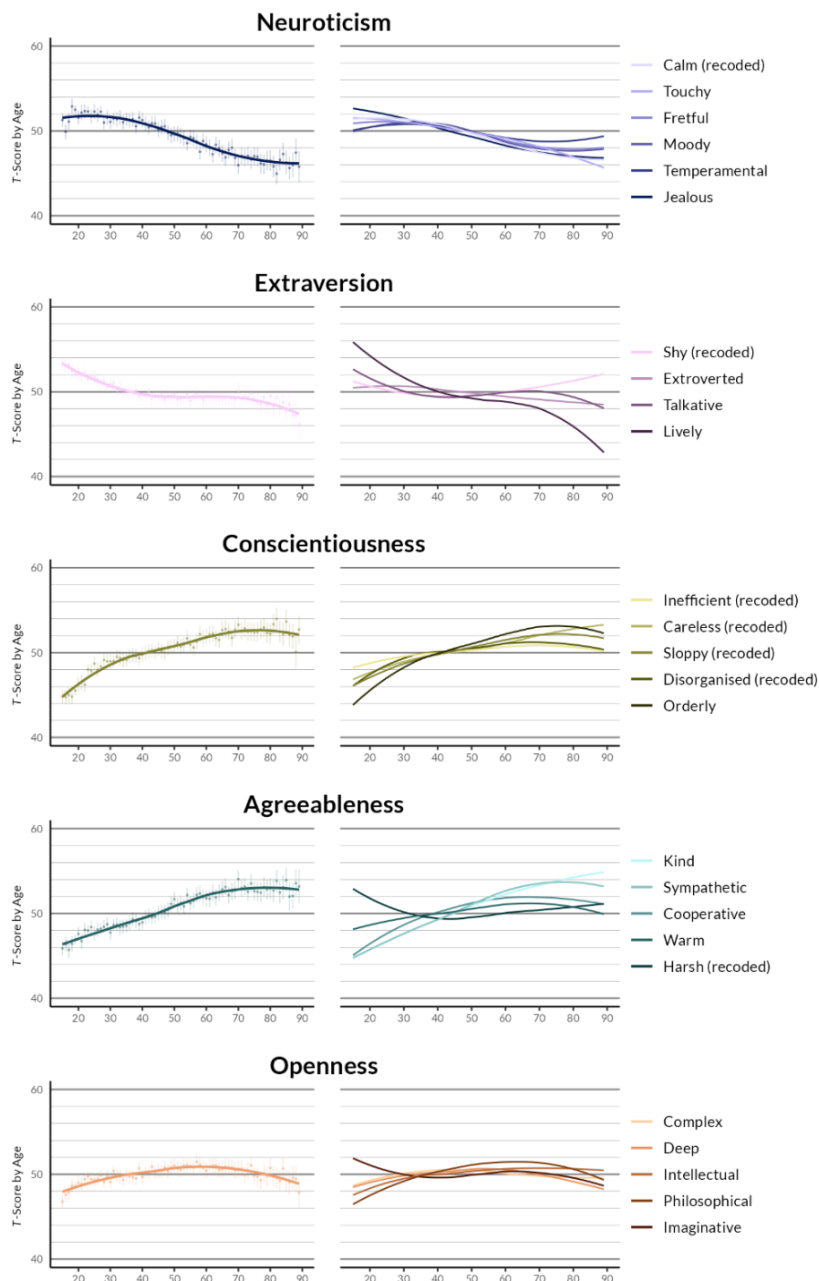
We used FEM to estimate Big Five mean-level trajectories that are only informed by within-person changes. To control for a potential initial elevation bias, we included a dummy variable indicating the wave in which a participant filled out the personality questionnaire for the first time. In the present sample, at their first measurement point, participants tended to respond to an item with stronger agreement, notwithstanding whether agreement on an item indicated higher or lower construct expression (i.e., results indicated a negative effect for recoded items and a positive effect for nonrecoded items; for more detailed results, see Table S4 in the Supplemental Materials). The impact of the first wave was small but consistent, supporting the inclusion of the first-wave dummy to remove any potential bias that may have been present. However, in a robustness check in which we removed the dummy, the developmental trends did not change substantively (for a contrast of the age trajectories, see Figures S1 to S5 in the Supplemental Materials).

Figure 1 depicts the resulting age trajectories for the mean scores (left panel) and the corresponding underlying items (right panel). Note that the depicted starting point of the trajectories is a weighted mean of model-implied intercepts.³

³ As only within-person changes are considered in FEM, the intercept is arbitrary and, therefore, typically dropped. However, for the visualization of the age trajectories, we computed an intercept replacement representing the level from which the changes occur (in our case: the youngest age as the reference group): The model-implied intercepts for each individual (the so-called individual fixed effects) were averaged with weights according to the number of waves in which an individual appeared in the analyses. For a consistent display of results across studies, age trajectories are plotted up to age 89 (for the full age range and unsmoothed trajectories, see the Supplemental Materials).

Figure 1

Within-Person Mean-Level Age Trajectories for the Big Five Scores (Left Panel) and the Corresponding Items (Right Panel) in the Household, Income and Labour Dynamics in Australia (HILDA) Survey



Note. Trajectories were smoothed with locally estimated scatterplot smoothing (LOESS).

Vertical bars represent the panel-robust standard error.

First, considering the Big Five mean scores (see the left panel in Figure 1), we found distinct age-related changes throughout the adult life span: a decrease in Neuroticism and Extraversion (the latter plateaued roughly between ages 40 and 70), an increase in Conscientiousness and Agreeableness, as well as a rather flat inverted U-shaped pattern in Openness (peaking between ages 50 and 60). Changes in personality across the life span were moderate to large, ranging in magnitude from about 6 to 8 *T*-score units.

Second, these developmental patterns in the mean scores were only partly mirrored on the item level (see the right panel in Figure 1; for the excluded items, see Figures S1 to S5 in the Supplemental Materials). For example, for Agreeableness, the item “Harsh” (recoded) distinctively declined until age 40 (meaning that up to age 40, people increasingly *agreed* that harsh described them well), whereas the remaining items for Agreeableness increased until age 40. Similarly, in the case of Openness, “Imaginative” decreased, whereas all other items initially increased, followed by a plateau or a small decrease. Regarding Extraversion, “Lively” exhibited a strong age-related decline, but the trajectories for the other items were rather weak. For Neuroticism and Conscientiousness, their development with age was comparably homogenous across items.

Summary

We used an Australian sample to examine how the Big Five traits change with age, taking into account only within-person information. Thereby, with decreases in Neuroticism and increases in Conscientiousness and Agreeableness, we found evidence for the maturity principle of personality development (e.g., Roberts & Nickel, 2017). However, both measurement invariance analyses and FEM showed that the age trends for the individual items that were subsumed under the same trait differed, sometimes even considerably. Aggregating the items

into mean scores may thus discard important between-item heterogeneity in development.

Furthermore, the results imply that estimated age trajectories for the Big Five scores may depend on the sample of items, as items can apparently not be arbitrarily exchanged.

Study 2: German Socio-Economic Panel (SOEP)

Method

Design and Participants

The SOEP is a household-based panel study from Germany with annual measurement waves. Focusing on life course trends, it covers a broad range of objective and subjective information—including numerous psychological variables, such as personality. The Big Five were measured in 2005, 2009, 2013, 2017, and 2019 for every household member who was at least 16 years of age. The survey began in 1984 with a random sample of households, and over time, several additional samples have been added to the survey. Of these samples, all were included in our analyses for which at least one 4-year retest interval of personality was covered.⁴ More general information on the SOEP is provided by Goebel et al. (2019; see also Giesselmann et al., 2019; Schröder et al., 2020).

The SOEP data can be accessed for scientific purposes and have been widely used (Goebel et al., 2019)—including several studies on personality development across the life span (Donnellan & Lucas, 2008; Graham et al., 2020; Lucas & Donnellan, 2011; Specht et al., 2011; see also Fitzenberger et al., 2022). However, none of these studies have included the five waves of personality data that are available to date, and none of them implemented FEM. In the present study, we used Version 36 (EU Edition) of the SOEP data (Liebig et al., 2021). For a list of

⁴ Three samples were excluded, as the relevant data were provided only for adolescents in these samples.

publications using the SOEP data, see

https://www.diw.de/en/diw_01.c.789503.en/publications_based_on_soep_data__soeplit.html.

A total of $N = 22,833$ participants were included in our analyses. On average, 3.27 measurement points ($SD = 1.16$) were used to measure personality per participant. Averaging participants' age across the respective waves in which they took part, participants had a mean age of 50.28 years ($SD = 18.06$). Following our inclusion criteria, the youngest age for which we considered personality measurements was 16, the oldest 93. The proportion of female participants was 52.74%. Conducting the analyses separately for the Big Five resulted in slightly varying sample sizes across traits (see Table 2; for trait-specific sample descriptions, see Table S5 in the Supplemental Materials).

Measures

The Big Five were measured in 2005, 2009, 2013, 2017, and 2019. Participants were presented with several statements that followed the phrase “I am someone who ...” (e.g., “I am someone who works thoroughly”). Participants were asked to rate how well these items described them on a 7-point scale ranging from 1 (*does not apply at all*) to 7 (*applies perfectly*). Each personality dimension was assessed with three items that were identical across waves. In 2009, a fourth item was added to measure Openness. We removed this additional item from the mean score to ensure comparable interpretations of the factors across time (but results for this item are reported in the Supplemental Materials Table S8 and Figure S10). The items were selected from the Big Five Inventory (John & Srivastava, 1999; see also Benet-Martínez & John, 1998) and were translated into German (for the original wording, see Table S6 in the Supplemental Materials). For more details on the Big Five measure that was used in the SOEP,

see Gerlitz and Schupp (2005; see also Dehne & Schupp, 2007; Hahn et al., 2012; Richter et al., 2017).

Due to the few and deliberately heterogeneous items (Gerlitz & Schupp, 2005), internal consistencies for the Big Five measure that was used in the SOEP were low (e.g., Seifert et al., 2022; Wagner et al., 2019). Internal consistencies were also low in our data analytic sample; for Neuroticism ($\alpha = .60$ to $.66$; $\omega = .63$ to $.68$), Extraversion ($\alpha = .67$ to $.69$; $\omega = .69$ to $.72$), Conscientiousness ($\alpha = .63$ to $.66$; $\omega = .66$ to $.68$), Agreeableness ($\alpha = .51$ to $.54$; $\omega = .55$ to $.59$), and Openness ($\alpha = .61$ to $.63$; $\omega = .62$ to $.64$; for wave-specific results, see Table S7 in the Supplemental Materials). Different assessment modes were used to collect the personality data—from self-completion questionnaires to telephone and (partly computer-assisted) in-person interviews—and for 19.20% of the included participants, the mode changed between waves. Because susceptibility to social desirability may be more pronounced in interviews (Richman et al., 1999; see also Ausmees et al., 2022), the assessment mode has the potential to lead to biased estimates of personality change among participants for whom the mode changed. To account for this possibility, we controlled for assessment mode in our analyses.

Results

Measurement Invariance

To test the Big Five for measurement invariance, we applied LSEM. The resulting fit indices are reported in Table 2. Residual variances of each item were allowed to covary across waves, but for Agreeableness, we had to fix the residual covariances of the item “Is considerate and kind to others” to 0 to avoid nonpositive definite residual covariance matrices (the same constraint was applied in the analyses by Seifert et al., 2022; see also Lucas & Donnellan, 2011).

Across traits, the model fit was only slightly impaired in general when the factor loadings were constrained to equality (i.e., assuming metric invariance), and ΔCFI was consistently below .01. Hence, as in Study 1, metric invariance was supported for personality. Further setting the item intercepts to equality (i.e., assuming scalar invariance) was followed by a pronounced loss in model fit (for each trait, ΔCFI was above .01), especially for Neuroticism, Openness and Conscientiousness. Thus, in accordance with Study 1, scalar invariance could not be established.

The measurement invariance analyses suggested that the single items in the SOEP might not develop in a uniform manner across age; and, thus, the age trend seen in the mean score might not be representative of the underlying items. Consequently, we present developmental trajectories on both the mean-score and item levels.

Table 2

Fit Indices for Testing for Measurement Invariance (MI) for the Big Five in the Socio-Economic Panel (SOEP)

Model	<i>n</i>	CFI	RMSEA	SRMR
Neuroticism	22,727			
Baseline		.994	.020	.015
Metric MI		.991	.022	.022
Scalar MI		.962	.041	.037
Extraversion	22,722			
Baseline		.993	.022	.024
Metric MI		.987	.029	.040
Scalar MI		.975	.037	.047
Conscientiousness	22,639			
Baseline		.990	.024	.027
Metric MI		.985	.027	.033
Scalar MI		.954	.043	.046
Agreeableness ^a	22,716			
Baseline		.986	.023	.021
Metric MI		.984	.023	.023
Scalar MI		.966	.031	.031
Openness	22,593			
Baseline		.996	.016	.012
Metric MI		.995	.016	.017
Scalar MI		.974	.036	.037

Note. Personality was assessed in 2005, 2009, 2013, 2017, and 2019. Focal points (representing birth years) ranged from 1923 to 2000 for Neuroticism, Conscientiousness, Agreeableness, and Openness; and from 1920 to 2000 for Extraversion. Metric MI requires the same factor loadings across both years of measurement and years of birth. Scalar MI additionally requires the same item intercepts across both years of measurement and years of birth. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

^a Residual covariances of one item across waves were fixed to 0 to avoid nonpositive definite residual covariance matrices.

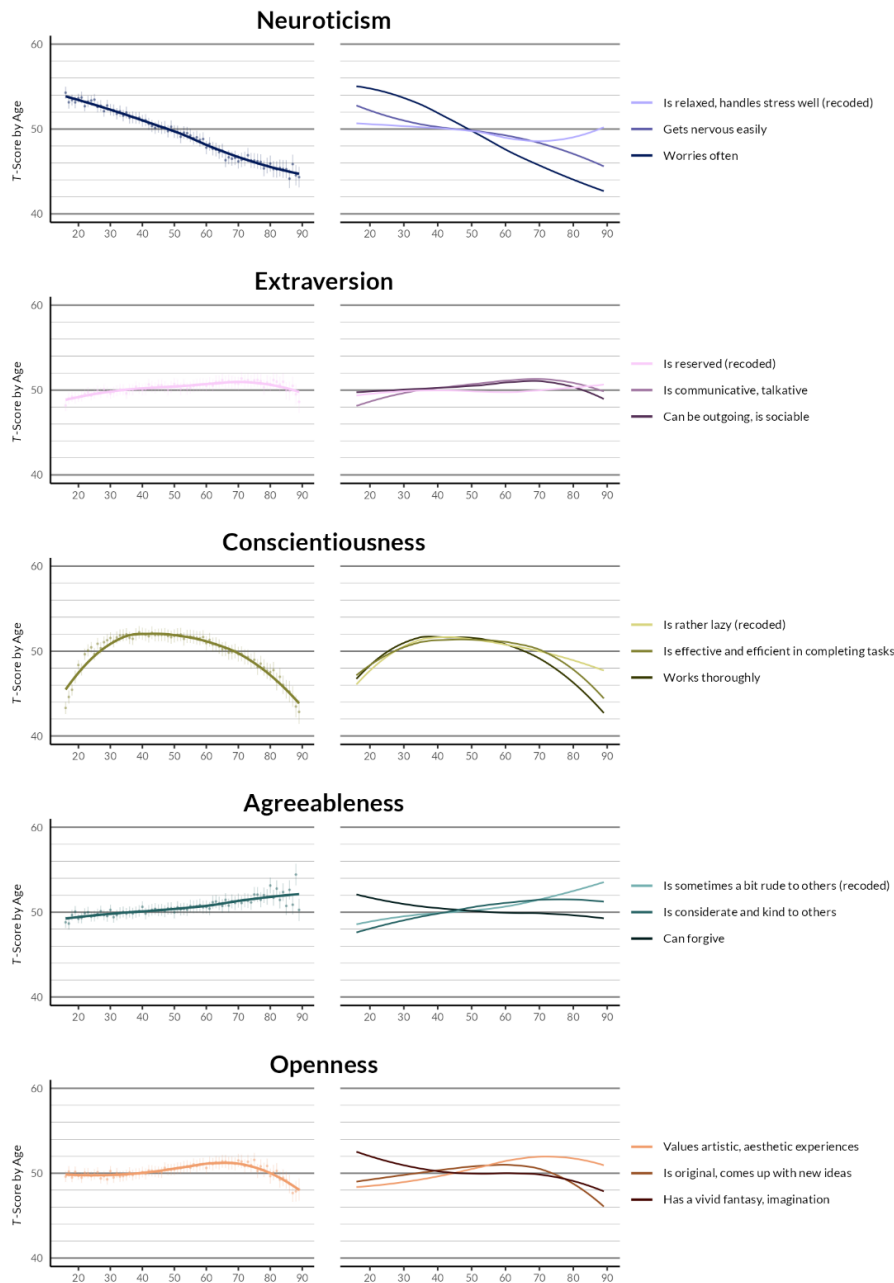
Age Trajectories

Mean-level age trends in the Big Five were estimated with FEM so that only changes within individuals were considered. Parallel to Study 1, we controlled for the first wave to account for a potential initial elevation bias. To account for mode changes in the SOEP, we additionally included a dummy variable that indicated whether the personality data in each wave were collected by means of a self-report questionnaire or an interview.

Across traits, we found that participants tended to indicate higher construct expression when they provided personality data for the first time. That is, in the first wave, respondents agreed more strongly with items that were aligned with the underlying trait, and they disagreed more strongly with items that were worded in the opposite direction (contrasting Study 1, in which respondents agreed more strongly regardless of the direction of the item; for more detailed results, see Table S8 in the Supplemental Materials). Furthermore, when participants filled out a self-report questionnaire instead of being interviewed, they tended to indicate higher values on Neuroticism and lower values on the remaining Big Five traits (this trend held for both items and mean scores; for more details, see Table S8 in the Supplemental Materials). These effects were small to moderate, supporting our decision to control for the first wave and the mode (for a demonstration of how these controls affected the age trajectories, see Figures S6 to S10 in the Supplemental Materials). The resulting age trajectories for the mean scores (left panel) and the corresponding items (right panel) are presented in Figure 2.

Figure 2

Within-Person Mean-Level Age Trajectories for the Big Five Scores (Left Panel) and the Corresponding Items (Right Panel) in the Socio-Economic Panel (SOEP)



Note. Trajectories were smoothed with locally estimated scatterplot smoothing (LOESS).

Vertical bars represent the panel-robust standard error.

On the level of mean scores (see the left panel in Figure 2), we found different life span trends for the Big Five: Neuroticism decreased, Extraversion and Agreeableness increased slightly (with some deviation from this pattern after age 80), Conscientiousness followed an asymmetrical inverted U-shaped pattern (i.e., the increase was steeper than the decrease) with a peak between ages 40 and 50, Openness first increased but then markedly decreased starting around age 70. Age effects were comparatively large for Neuroticism (which showed a change across the life course of around 8 *T*-score units) and Conscientiousness (which showed increases and decreases of around 8 *T*-score units), whereas the effects for the other three dimensions were much smaller, with changes of about 3 to 4 *T*-score units.

However, these trends were only partly present on the item level, and item-specific trends emerged (see the right panel in Figure 2). For example, regarding Neuroticism, the item “Worries often” indeed drove the mean-score decline, which was much less pronounced for the remaining items. The (recoded) responses to “Is relaxed, handles stress well” even increased in old age. Thus, despite the mean-level decrease in Neuroticism, individuals reported being *less* relaxed and *less* able to handle stress well with age. In the case of Agreeableness, “Can forgive” declined, whereas the other items increased. The Openness item “Values artistic, aesthetic experiences” increased with age, but the opposite occurred for “Has a vivid fantasy, imagination.” Regarding Conscientiousness, the item “Is rather lazy” (recoded) declined somewhat less strongly in old age than the remaining items. Thus, in old age, individuals reported substantially lower levels of Conscientiousness overall, but the increase in self-reported laziness was quite weak. For Extraversion, the items developed relatively homogeneously with age, with very little change overall.

Summary

Using data from Germany, we examined the robustness of the results from Study 1. We were able to replicate some of the age trends for the mean scores: Neuroticism decreased, Agreeableness increased, Openness increased and then decreased. Again, especially in young adulthood, we found evidence for the maturity principle of personality development (e.g., Bleidorn et al., 2021). But as in Study 1, our measurement invariance and FEM analyses showed that these mean-score trends did not consistently represent the item trends. Again, the results underscore the idea that the age trajectories for the Big Five scores depend on the underlying items.

Study 3: Dutch Longitudinal Internet Studies for the Social Sciences (LISS) Panel

Method

Design and Participants

The LISS Panel is a survey from the Netherlands that collects information on diverse topics including several psychological variables. It is based on a probability sample of households drawn from the population register. We included a total of 12 waves in the present study: Personality was measured eight times in the full sample (i.e., in 2008, 2009, 2011, 2013, 2014, 2017, 2019, and 2020) and four times for any participants who had not completed their participation in the respective previous year (i.e., in 2010, 2012, 2015, and 2018). Each panel member at least 16 years of age could provide personality data. Across the years, several refreshment samples of additional households were added to the initial sample. For more details on the LISS, see Scherpenzeel (2011; see also Scherpenzeel & Das, 2011).

The LISS data can be accessed for scientific purposes and have already been used for research in personality psychology (e.g., Beck & Jackson, 2022; Denissen et al., 2019; Schwaba

et al., 2018)—but studies have yet to investigate the mean-level development of personality across the life span with these data. Personality data in the LISS are released wave-wise (e.g., Marchand, 2018, 2019, 2020); for transparency and reproducibility, we provide information about the file versions we used on the OSF. A list of publications using the LISS data can be found at <https://www.dataarchive.lissdata.nl/publications>.

A total of $N = 10,163$ participants were included in the analyses. The sample size did not vary across traits (potentially due to not allowing missing responses on items in the online questionnaire). Individuals were excluded from the analyses if they provided missing or unclear information on relevant variables (e.g., varying year of birth, which affected $n = 111$ of the approximately 28,000 individuals who are listed in the LISS data overall). On average, 4.50 measurement points ($SD = 2.21$) were included per participant for personality. Averaging participants' age across the respective waves in which they took part, participants had a mean age of 47.44 years ($SD = 18.10$). Based on our inclusion criteria, the youngest age for which we considered personality measurement was 16, the oldest 89. The proportion of female participants was 54.67% (for $n = 19$ of the included participants [0.19%], information on gender was ambiguous).

Measures

In an online self-report survey, personality was measured in each wave with the 50-item International Personality Item Pool (IPIP) Version of the Goldberg (1992) markers for the Big Five personality traits. Each trait was intended to be measured with 10 items. Items consisted of short statements (e.g., “Feel little concern for others”), and participants rated how well these statements described themselves on a 5-point scale ranging from 1 (*very inaccurate*) to 5 (*very*

accurate). For an overview of the items, including the original Dutch wording, see Table S9 in the Supplemental Materials.

However, when we tested a unidimensional model, we found unsatisfactory properties for each trait when we used the full version of the personality questionnaire (for more details, see Tables S10 to S14 in the Supplemental Materials), the same as previous studies that used these data (Schwaba & Bleidorn, 2018; see also Schwaba et al., 2018). Therefore, we included only six items for each trait. The process of item selection with detailed psychometric results is reported in the Supplemental Materials. We furthermore present results for the excluded items in the Supplemental Materials (see Table S17 and Figures S11 to S15).

For the subset of included items, we found satisfactory internal consistencies in the present sample across waves for Neuroticism ($\alpha = .75$ to $.85$; $\omega = .79$ to $.87$), Extraversion ($\alpha = .77$ to $.82$; $\omega = .80$ to $.85$), Conscientiousness ($\alpha = .68$ to $.73$; $\omega = .71$ to $.78$), Agreeableness ($\alpha = .71$ to $.78$; $\omega = .73$ to $.81$), and Openness ($\alpha = .62$ to $.69$; $\omega = .67$ to $.74$; for wave-specific results, see Table S15 in the Supplemental Materials).

Results

Measurement Invariance

We again used LSEM to test for measurement invariance but had to make adjustments because of the planned missingness design of the LISS. To avoid estimation issues, we restricted the measurement invariance analyses to the eight waves in which personality was assessed in the full sample (leading to a somewhat lower sample size of $n = 9,512$ participants for each trait). The resulting fit indices from the measurement invariance analyses are reported in Table 3.⁵

⁵ As larger model sizes might impair the sensitivity to detect measurement invariance (Cao & Liang, 2022), we conducted additional analyses with a subset of five waves (i.e., 2008, 2011, 2014, 2017, and 2020), thereby increasing comparability with Studies 1 and 2 (which included four and five waves, respectively). These analyses

Table 3*Fit Indices for Testing for Measurement Invariance (MI) for the Big Five in the Longitudinal**Internet Studies for the Social Sciences (LISS) Panel*

Model	<i>n</i>	CFI	RMSEA	SRMR
Neuroticism	9,512			
Baseline		.962	.025	.032
Metric MI		.960	.025	.040
Scalar MI		.953	.026	.041
Extraversion	9,512			
Baseline		.970	.024	.032
Metric MI		.968	.024	.041
Scalar MI		.959	.027	.044
Conscientiousness	9,512			
Baseline		.957	.024	.036
Metric MI		.955	.024	.041
Scalar MI		.943	.027	.046
Agreeableness	9,512			
Baseline		.954	.024	.029
Metric MI		.952	.024	.035
Scalar MI		.942	.026	.039
Openness	9,512			
Baseline		.958	.023	.036
Metric MI		.956	.023	.039
Scalar MI		.950	.024	.041

Note. Personality assessments were included for 2008, 2009, 2011, 2013, 2014, 2017, 2019, and 2020. For each trait, focal points (representing birth years) initially ranged from 1930 to 2002 but were restricted to 1933–2002 for Conscientiousness and to 1931–1994 for Openness to avoid nonpositive definite factor covariance matrices. Metric MI requires the same factor loadings across both years of measurement and years of birth. Scalar MI additionally requires the same item intercepts across both years of measurement and years of birth. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

led in general to similar conclusions as the analyses that were based on eight waves (for more detailed results, see Table S16 in the Supplemental Materials).

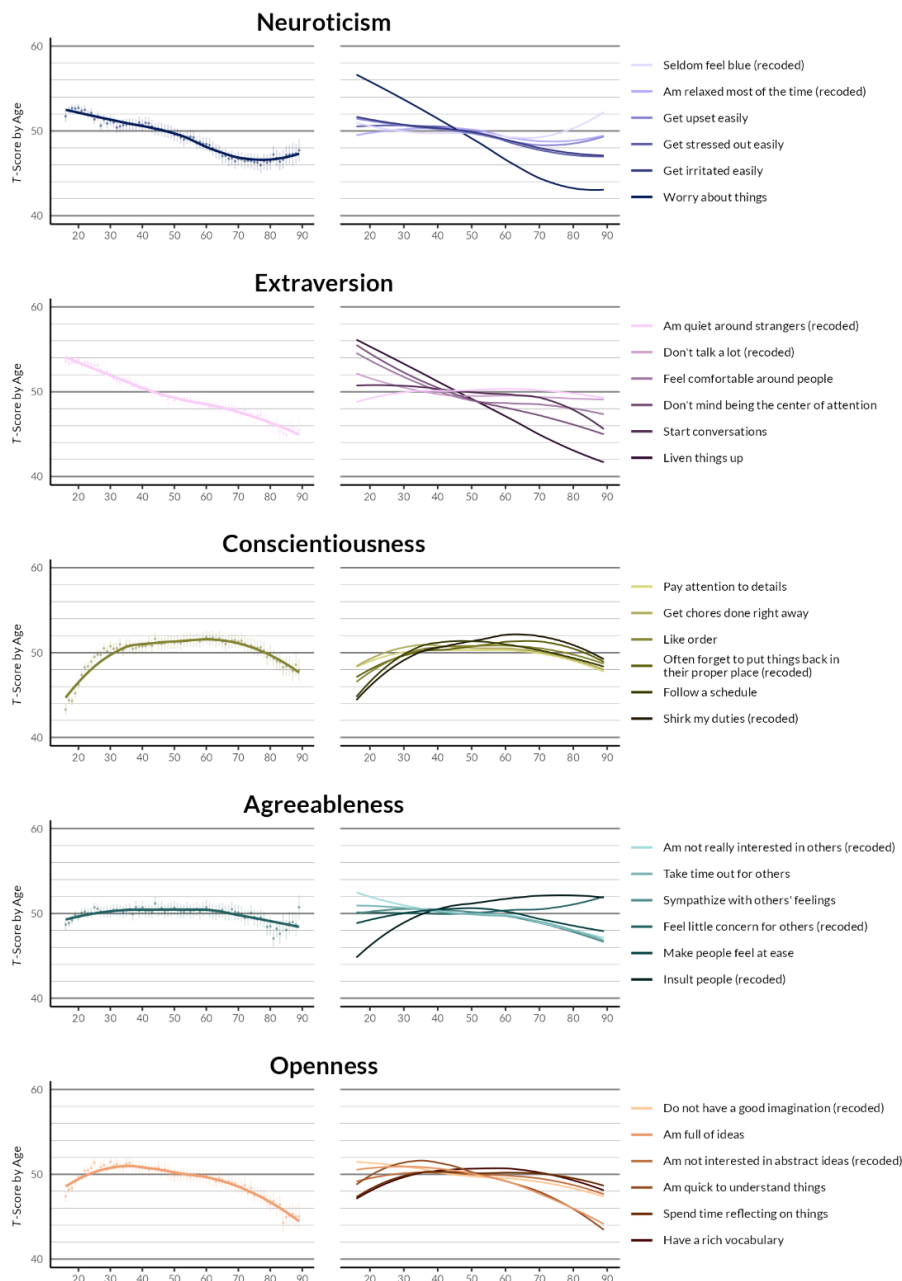
For each trait, requiring the factor loadings to be equal (i.e., assuming metric invariance) caused at best a slight impairment in model fit. Hence, resembling Studies 1 and 2, metric invariance was supported for personality. Additionally imposing equality constraints on the item intercepts (i.e., assuming scalar invariance) went along with a more marked loss in model fit that was relatively comparable across traits. However, as these decreases in model fit were weaker than in Studies 1 and 2, and mostly below the criterion of $\Delta\text{CFI} = .01$ suggested by Cheung and Rensvold (2002), we interpret this result as an ambiguous finding on scalar measurement invariance. To ensure that we did not potentially miss divergent patterns, and to ensure that this study was parallel to Studies 1 and 2, in the subsequent FEM analyses, we thus again additionally present results for the item level.

Age Trajectories

Personality age trajectories were estimated with FEM. All 12 waves of personality data were used for the analyses (thus, $N = 10,163$ participants were included). As in Studies 1 and 2, we controlled for the first wave to account for a potential initial elevation bias in personality; and, similar to Study 2, we found that participants tended to report higher construct expression if they provided personality data for the first time (for more detailed results, see Table S17 in the Supplemental Materials). Due to a small but consistent effect of the first wave on personality, which had the potential to bias the age trajectories, we controlled for the first-wave dummy. However, removing the dummy did not alter the developmental trends substantively (for a contrast of the age trajectories, see Figures S11 to S15 in the Supplemental Materials). Age trajectories for the mean scores (left panel) and the subordinate items (right panel) are presented in Figure 3.

Figure 3

Within-Person Mean-Level Age Trajectories for the Big Five Scores (Left Panel) and the Corresponding Items (Right Panel) in the Longitudinal Internet Studies for the Social Sciences (LISS) Panel



Note. Trajectories were smoothed with locally estimated scatterplot smoothing (LOESS).

Vertical bars represent the panel-robust standard error.

Regarding trends that occurred on the mean-score level throughout the adult life span (see the left panel in Figure 3), Neuroticism and Extraversion decreased by approximately 8 *T*-score units (with minor deviations from this general pattern at older ages for Neuroticism); Conscientiousness developed in an inverted U-shaped fashion but held fairly steady between ages 30 and 70 (increasing by roughly 8 *T*-score units and decreasing by about 4 *T*-score units); Agreeableness exhibited no pronounced age-related changes besides a slight increase at the youngest ages we observed and a dip at around 80 years (each covering approximately 2 *T*-score units); Openness increased until age 25 by around 4 *T*-score units and declined afterwards by roughly 6 *T*-score units.

For Conscientiousness, the items followed the mean-score trend quite homogeneously, but for all other traits, the individual items diverged (see the right panel in Figure 3; for the excluded items, see Figures S11 to S15 in the Supplemental Materials). Regarding Neuroticism, the item “Worry about things” had a steeper age-related decline than the remaining items (which further showed some developmental heterogeneity in old age). A similar pattern emerged for Extraversion, for which the amount of decline varied considerably across items with the most pronounced age trend for “Liven things up” and a rather flat trajectory for “Am quiet around strangers” (recoded). For Agreeableness, the recoded item “Insult people” increased, whereas the remaining items tended to decrease with age. For Openness, “Am full of ideas” and “Am quick to understand things” declined to a greater degree after midlife than the other items.

Summary

In Study 3, we were able to replicate some of the general mean-score age trends that we found in both Studies 1 and 2 (i.e., a decrease in Neuroticism and an inverted U-shape for Openness), or in one of them (i.e., the decline in Extraversion as in Study 1 and the inverted U-shape for Conscientiousness as in Study 2). Whereas decreasing Neuroticism and (at least in young adulthood) increasing Conscientiousness are in line with the maturity principle of personality development (e.g., Caspi et al., 2005), we found comparatively weak evidence for such changes in Agreeableness. Additionally, Study 3 replicated a general result that was also present in Studies 1 and 2: the heterogeneous age trends in the items.

Additional Analyses: Comparing Fixed Effects Modeling With Standard Multilevel Modeling

With FEM as our analytical approach, we relied exclusively on within-person changes to estimate age trends in personality. To better understand the extent to which separating within-person change from between-person differences is important for substantive conclusions, we reanalyzed our data for the Big Five scores with standard multilevel modeling (MLM), where estimates incorporate both within-person *and* between-person information (Curran & Bauer, 2011).

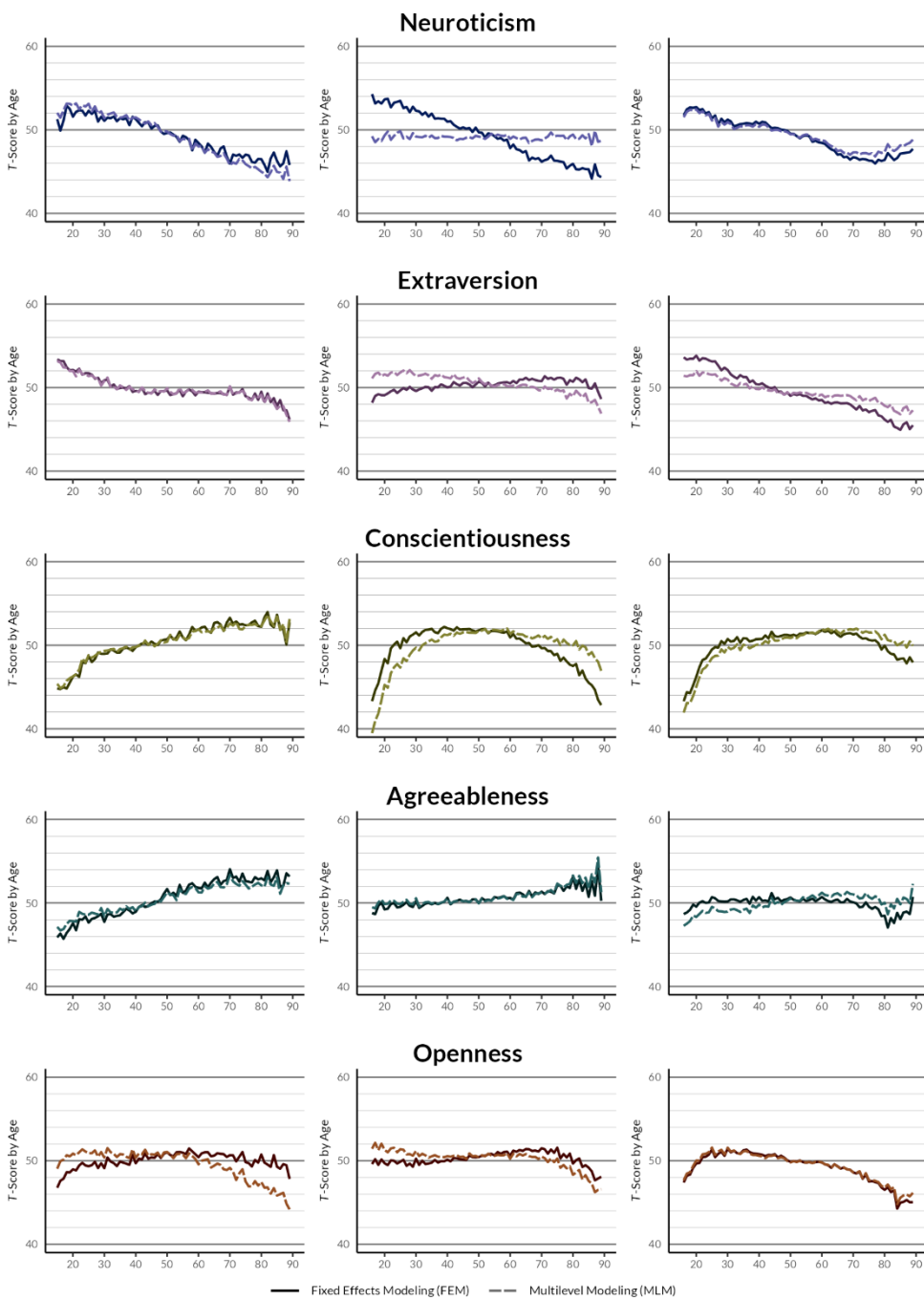
The resulting age trajectories from MLM for the Big Five scores are presented in Figure 4. For ease of comparison, Figure 4 furthermore contains the FEM-based score trends from Figures 1 to 3. In some cases, age trends from MLM and FEM were similar, but in other cases, the two approaches diverged. Most strikingly, in Study 2, MLM indicated nearly no age-related changes for Neuroticism, whereas FEM showed a pronounced decline across the life span. Relevant differences between the developmental trajectories from MLM and FEM were also

present for the remaining Big Five traits; for Extraversion (Studies 2 and 3), Conscientiousness (Study 2), Agreeableness (Study 3), and Openness (Studies 1 and 2). The additional inclusion of participants with only one personality assessment—participants who had to be excluded from the FEM analyses—did not affect the age trajectories that resulted from the MLM analyses (see Table S18 and Figure S19 in the Supplemental Materials).

Discrepancies between the results from MLM and FEM can be traced back to differences in the sources of information the approaches use to estimate the personality age trajectories. For illustration, imagine the following case: Individuals from earlier born cohorts exhibit higher levels of Neuroticism, so that purely from a between-person perspective, an *increase* appears with age. At the same time, however, there is a *decrease* in the trait within individuals as they get older. These contradictory between-person and within-person age trends would be mixed in analyses with MLM, resulting in a rather flat developmental trajectory for Neuroticism. By contrast, FEM would indicate a decrease with age, as here, only the within-person changes are considered. A very similar case might be present for Neuroticism in Study 2 (see the middle panel in Figure 4), where MLM and FEM are in stark conflict.

Figure 4

Mean-Level Age Trajectories for the Big Five Scores Based on Multilevel Modeling (MLM) and Fixed Effects Modeling (FEM) in the Household, Income and Labour Dynamics in Australia (HILDA) Survey (Left Panel), the Socio-Economic Panel (SOEP; Middle Panel), and the Longitudinal Internet Studies for the Social Sciences (LISS) Panel (Right Panel)



— Fixed Effects Modeling (FEM) - - - Multilevel Modeling (MLM)

In general, any between-person differences in personality that are not due to true age effects might explain why trajectories based on MLM differ from FEM estimates. Thereby, discrepancies between the two modeling approaches grow with the size of the non-age-related between-person effects in the data. For further comparison, we computed additional analyses to estimate personality age trajectories that consider *only* between-person information (for more details, see the Supplemental Materials). As expected, the trajectories from MLM generally fell in the middle of between-person and FEM trends. In Studies 1 and 2, the trajectories from MLM were somewhat closer to between-person trends than to FEM, whereas in Study 3, the trajectories from MLM and FEM tended to resemble each other (see Figures S16 to S18 in the Supplemental Materials). This difference might be due to the fact that Study 3 had a larger number of assessments per individual, and consequently, more within-person information was available than in Studies 1 and 2.

Taken together, age trends derived from FEM and MLM do not necessarily converge. Researchers interested in developmental processes should be aware that if they analyze longitudinal data by using the standard method of implementing MLM, relevant within-person effects will be inseparably blended with potentially confounding between-person effects. As the actual age trajectory of interest will be obscured, such an approach might result in misleading conclusions. Instead, we recommend FEM as a straightforward approach for extracting unblended within-person effects.

General Discussion

The numerous studies examining the mean-level development of personality with age have failed to draw a conclusive picture, as substantial heterogeneity in the reported trajectories has been observed. These inconsistencies may have occurred in part because, in previous

research, within-person changes were not strictly separated from between-person differences in personality; thus, reported age effects may have been confounded with cohort effects. This issue cannot automatically be resolved by relying on longitudinal data, as common statistical models for such data in psychology do not clearly separate between- from within-person information or may lead to biased estimates if developmental patterns vary between individuals.

By contrast, in our studies, age trajectories were informed only by within-person changes in personality, and they can correctly recover average patterns even in the presence of heterogeneous trajectories. We analyzed panel data from Australia (HILDA; Study 1), Germany (SOEP; Study 2), and the Netherlands (LISS; Study 3), so we were able to gauge the robustness and generalizability of the results across independent samples.

Personality Age Trajectories Across Studies

Much in line with the maturity principle of personality development (e.g., Roberts & Nickel, 2017), when we compared the age trajectories across the three samples, we consistently found that Neuroticism declined throughout adulthood. Previous studies have typically reported a similar developmental pattern for young and middle adulthood (e.g., Bleidorn et al., 2022; cf. e.g., Specht et al., 2011) but have not been conclusive about whether the decline continues in older ages (e.g., Wortman et al., 2012) or ceases (e.g., Graham et al., 2020). Further, we found that the decline in Neuroticism was generally mirrored on the item level. Here, items measuring the tendency to worry (included in Studies 2 and 3) showed the most pronounced downward trends with age; and quantifying the similarity of individual item trajectories with the Fréchet distance indicated that worry items were more similar to each other than to other items from the same trait in the same study (see Figure S20 in the Supplemental Materials). Thus, if studies

include items that tap into the general tendency to worry, they might find steeper declines in Neuroticism with age.

Extraversion in adulthood showed inconsistent trends across studies: A slight increase (Study 2) versus a clear decline (Study 3) versus a pronounced decrease followed by a plateau (roughly around age 40; Study 1). However, in all studies, we found a decline in the oldest ages, starting at about age 80. Unsystematic age patterns for Extraversion have been reported before (e.g., Marsh et al., 2013; Soto et al., 2011) with different trends for different aspects of the trait (Roberts et al., 2006). Thus, heterogeneity in developmental patterns might be explained by differences in questionnaire content; and our data supported this idea. We consistently found marked age-related declines in Extraversion items assessing vitality (e.g., “Lively” in Study 1 and “Liven things up” in Study 3) and comparatively weak trends in the other items. This finding was corroborated by our Fréchet distance analyses, which showed that the developmental trends of the vitality items were typically more similar to each other than to the remaining Extraversion items within a study (see Figure S21 in the Supplemental Materials). Study 2 did not contain any items that captured this aspect of vitality, perhaps explaining why no decline was observed in this sample. Intriguingly, our consistent and marked age-related declines in vitality differed from meta-analytic results by Roberts et al. (2006), where age effects for the Social Vitality facet of Extraversion were comparatively weak.

For Conscientiousness, in each sample, we found a steep increase until about age 30; a result that was in line with previous evidence (e.g., Bleidorn et al., 2022; cf. Graham et al., 2020) and the maturity principle of personality development (e.g., Bleidorn et al., 2021). This very clear pattern was also consistently reflected on the item level. In middle adulthood, trait levels either remained stable (Study 2) or continued to increase (but somewhat less strongly than in

young adulthood; Studies 1 and 3). At older ages, Conscientiousness sharply decreased in Study 2 (see, e.g., Marsh et al., 2013), but the decrease was much less pronounced in Study 3 and fairly nonexistent in Study 1. Item-level trends at these higher ages also became less consistent with the exception of items pertaining to work (e.g., “Works thoroughly” in Study 2 and “Am exacting in my work” in Study 3; the latter item not contributing to the mean score), which showed pronounced declines at older ages. The Fréchet distances indicated that the work-related items developed in a similar manner, while they exhibited different developmental patterns compared to the work-*un*related items (the latter also exhibited similar developmental trends; see Figure S22 in the Supplemental Materials). The work items were included in the mean score only in Study 2 and this may explain why Study 2 showed a much more noticeable decline in Conscientiousness in old age. This trend also suggests that whether a study finds a decline in Conscientiousness in old age depends on whether the items are linked to the work context (which becomes less relevant after retirement) or whether they refer to Conscientiousness in a more general manner (more applicable to life beyond work). We recommend that researchers be careful about using contextualized personality measures for studies that include participants from across the age span. Changes in life circumstances may affect item responses and, as pointed out in the section on measurement invariance, this allows for different interpretations. For example, one could either conceptualize a subfactor of Conscientiousness that declines with retirement and measure it intentionally, or one could assume that the effects of life circumstances on those items are independent of Conscientiousness and are thus a nuisance to be avoided.

In each sample, in accordance with the maturity principle once again (e.g., Caspi et al., 2005), increases in Agreeableness were found until roughly age 30 (e.g., Bleidorn et al., 2022; cf. Lucas & Donnellan, 2011). Beyond age 30, Agreeableness continued to increase in Study 1 and

to a lesser extent in Study 2, whereas it did not change in Study 3. On the item level, we consistently observed that the items with the strongest age increase could be summarized under the kindness facet (Study 1: “Kind,” “Sympathetic”; Study 2: “Is considerate and kind to others,” “Is sometimes a bit rude to others” [recoded]; Study 3: “Insult people” [recoded]). The remaining items showed less pronounced age trends, and many of them even decreased with age (e.g., Study 1: “Harsh” [recoded]; Study 2: “Can forgive”; Study 3: “Take time out for others”; for a measure of similarity of the item trends, see Figure S23 in the Supplemental Materials). These findings suggest that, for Agreeableness in particular, item choice will strongly determine whether one finds a pronounced increase with age (Study 1) or whether patterns cancel each other out and result in much flatter trajectories (Studies 2 and 3). It may even be possible to select Agreeableness items in a manner that leads to a decrease with age. These different patterns for things that are usually subsumed under the “Agreeableness” label may also be substantively interesting. For example, one could speculate that general kindness is mostly adaptive (or, at least socially desirable) and is thus indeed a sign of a “mature” personality, whereas forgiving others and taking time out for others indiscriminately can backfire.

In line with previous research, the Openness age trajectories tended to diverge across the three samples (e.g., Soto et al., 2011; Wortman et al., 2012). In Study 1, Openness increased until roughly age 60, followed by a decline. In Study 2, Openness initially did not change but started to increase from age 30 to age 70 and then sharply declined. In Study 3, Openness initially sharply increased, peaked between ages 30 and 40, and declined afterwards. One *could* describe the patterns as an inverted U-shape, but such a description would obscure the fact that the studies completely disagreed about the timing of the change, except for the decline in oldest age (which has varied somewhat in size across studies; see also Bleidorn et al., 2022). The

specific item content may partly explain these divergences, which is particularly suggestive given the generally somewhat vague conceptualization of Openness (Costa et al., 2019; John, 2021; Schwaba, 2019). First, we did find that items covering the aspect of being imaginative consistently declined with age (e.g., “Has a vivid fantasy, imagination” in Study 2 and “Have a vivid imagination” in Study 3; the latter item not contributing to the mean score). Second, Study 3 actually included items that may tap into self-reported aspects of intelligence—“Am quick to understand things” may reflect processing speed and indeed showed an age trajectory that mirrored fluid intelligence (increase in young adulthood followed by a continuous decline); “Have a rich vocabulary” may reflect crystallized intelligence and indeed showed the matching age trajectory (increase until midlife followed by a rather stable plateau; e.g., McArdle et al., 2002; Salthouse, 2019). For the remaining Openness items, developmental changes were typically rather small in magnitude (for a measure of similarity of the item trends, see Figure S24 in the Supplemental Materials).

To summarize, we did find clear evidence for the maturity principle of personality development (e.g., Caspi et al., 2005; Roberts & Nickel, 2017). Most prominently, Neuroticism decreased throughout adulthood in all samples, regardless of the choice of items. Furthermore, Conscientiousness increased until age 30 in all samples, once again regardless of the items. Agreeableness also increased until age 30 across samples, mostly driven by the items that tapped into kindness. But how large these changes are (e.g., how strongly Agreeableness increases), how the other Big Five traits (i.e., Extraversion and Openness) behave, and what happens beyond middle age is a more complex story (see, e.g., Bleidorn et al., 2022; Roberts et al., 2006). All in all, the changes that were observed in personality across the entire life span were generally

moderate to large, a finding that is in line with meta-analytic estimates (e.g., Bleidorn et al., 2022; Roberts et al., 2006).

Our findings suggest that the apparent heterogeneity (see also Bleidorn et al., 2022) does not (only) reflect sample idiosyncrasies or irreducible differences between countries but can (also) be explained by differences in item choice: Single items show developmental patterns that differ from the trait to which they supposedly belong. Dissimilar item trajectories within a trait were also reflected by the Fréchet distance, a quantitative measure of similarity (for more details, see Figures S20 to S24 in the Supplemental Materials). At the same time, high similarity was found with the Fréchet distance for items with similar content across studies (e.g., items tapping into the tendency to worry). However, the replication of differential item patterns across studies was not possible for all cases. Either particular items were not surveyed across different studies (e.g., something like “being quick to understand things” was assessed only in Study 3), or these items showed no empirically consistent pattern across different studies, although they were similar in content (e.g., “Talkative” in Study 1 and “Is communicative, talkative” in Study 2).

Altogether, a narrow focus on trait mean scores might lead researchers to miss important parts of the picture, and our results support calls to examine personality development below domain levels (e.g., Möttus & Rozgonjuk, 2021). Thus, the present findings are in line with the rising interest in so-called “personality nuances,” which refer to individual items, or groups of very similar items, as the most specific units in the personality hierarchy (McCrae, 2015; McCrae & Möttus, 2019). For example, studies have found that personality nuances are stable and heritable, that they show interrater agreement (Möttus et al., 2017, 2019), and that they are valid predictors of life outcomes beyond traits and facets (Stewart et al., 2022). Nevertheless, from a theoretical point of view, it would be preferable to investigate age trajectories on the facet

level, where multiple personality nuances are integrated into a particular aspect of a superordinate and more general Big Five trait (McCrae, 2015; McCrae & Mõttus, 2019). Compared with individual items, facets have the advantage that they can be measured with less error and can also be modeled as a latent variable. Numerous studies have demonstrated the utility of personality facets (e.g., Anglim et al., 2020; Paunonen & Ashton, 2001) and, indeed, different facets of the same Big Five trait have shown unique developmental trends (e.g., Schwaba et al., 2022; Soto et al., 2011). However, it should be ensured that the items constituting a personality facet develop homogeneously with age too (e.g., by testing for measurement invariance on the facet level; see, e.g., Olaru et al., 2022); and it requires a more detailed assessment of personality as typically conducted for the kinds of large panel studies that we used in the present investigation.

Interpreting the Trajectories as Age Effects

The developmental trends for the Big Five personality traits presented in our studies were informed only by changes that occurred within persons across time. Thus, the estimated age effects were controlled for all *constant* confounding variables, including cohort effects on personality (e.g., Jokela et al., 2017). Of course, this is an important advantage over cross-sectional studies (e.g., Soto et al., 2011) but also over studies that analyzed longitudinal data in a way that includes differences within *and* between individuals (e.g., Graham et al., 2020; see also Curran & Bauer, 2011); in both cases, observed age trajectories may be confounded with cohort effects.

However, in our analyses that were based only on within-person information, time-varying variables may also still bias the results. We thus controlled for two variables that could potentially induce associations between age and reported personality that do not reflect

associations with actual personality: changes in response mode and initial elevation bias (Anvari et al., 2022; Arslan et al., 2021; Shrout et al., 2018). In Study 2, the only one of our studies in which a change in response mode was possible, we found systematic effects of the assessment mode on personality levels. This effect could be explained by a stronger susceptibility to socially desirable responding in interviews than in self-report questionnaires (Richman et al., 1999; see also Ausmees et al., 2022). Concerning initial elevation bias, response patterns seemed to be systematically affected by the first wave in general, but the specific effects varied by study: higher item agreement in Study 1 and higher construct expression in Studies 2 and 3 (leaving the exact nature of the bias somewhat obscure; see also Cerino et al., 2022). In any case, by controlling for changes in response mode and initial elevation bias, we were able to determine that our age trajectories were not biased by these variables.

However, other time-varying variables may still bias results, and period effects are especially relevant in the context of age trajectories. Period effects cannot be controlled for as easily as changes in response modes and initial elevation bias because age, cohort, and period effects cannot be captured simultaneously (the so-called “identification problem”; e.g., A. Bell, 2020). If two of the three variables are given, the third is logically determined as $\text{age} = \text{period} - \text{cohort}$. Whether period effects are problematic for analyses depends on the assumed shape of such effects. For example, a linear period effect—everybody becomes more conscientious from year to year—would not be distinguishable from a linear age effect, even in within-person data. But one could also imagine more shock-like period effects, such as economic crises that affect personality or at least change how individuals describe their personality. Such period effects would tend to be less problematic, as we observed multiple cohorts of individuals—thus, the period effect will affect personality ratings of people at different ages,

avoiding the perfect confounding of age and period. However, period effects in general have received comparably little attention in the study of personality development.

Furthermore, the longitudinal time span of the analyzed panel data was shorter than a life span, such that aging patterns had to be estimated across different cohorts. To actually piece together within-person changes as a cohesive life span trajectory, we must assume that a person's cohort does not interact with their age. Cohort may affect a person's general *level* of personality but not age-related *changes*, such that they can be meaningfully integrated into a common trajectory (what is called "linkage" or "convergence"; R. Q. Bell, 1953; Miyazaki & Raudenbush, 2000; Sliwinski et al., 2010; see also Mirowsky & Kim, 2007). If, instead, age effects systematically vary across cohorts, merging the individual pattern results in a trajectory that represents a (by-cohort) weighted average of the different age effects. To comprehensively test whether cohort modifies the effects of age, longitudinal data spanning longer time intervals would be necessary; these could again be analyzed within the FEM framework by examining the interaction between year of birth and age (McNeish & Kelley, 2019). Fortunately, recent research drawing on such more extensive longitudinal data suggests that aging patterns in personality may be comparable across cohorts (Brandt et al., 2022) so that cohort effects may be only a minor issue in the identification of unbiased age effects.

Limitations and Outlook

We investigated personality development based on the rather general framework of Big Five traits, and the measures included in the studies were rather short. Thus, our sample of personality items is certainly not comprehensive in covering all facets of the Big Five, let alone of personality in the general sense. The data analytic approach we champion could be applied to more extensive Big Five measures and to measures capturing other domains of interindividual

differences (e.g., values, goals, and interests; Kandler et al., 2014) to paint a more comprehensive picture of personality development. Likewise, even though there were items with similar content in the panel studies we included, it would be insightful to see if age trajectories are similar across samples if *identical* personality questionnaires are used.

Furthermore, one could combine the different Big Five measures from multiple panel studies (e.g., HILDA, SOEP, and LISS) and assess them in an age-heterogenous sample, preferably in a longitudinal design. These data would allow researchers to quantify the extent to which variance is shared between items from different personality questionnaires and how this overlap leads to convergence (or divergence) in the age-related item trends.

In addition, we relied on self-report measures, which may provide a limited perspective. In particular, age differences in personality seem to be contingent on social desirability to some extent (Ausmees et al., 2022), and other-reports might be less susceptible to such response tendencies (Richman et al., 1999). Thus, to move from a rather descriptive to a more mechanistic understanding of the mechanisms that underlie developmental changes in personality, it would be fruitful to contrast developmental patterns of personality across different sources of information.

Lastly, like so many studies in the field, we relied on samples from countries that are typically described as “western, educated, industrialized, rich, and democratic” (“WEIRD”; Henrich et al., 2010a, 2010b; for a critical examination of the term, see Clancy & Davis, 2019), which is a general issue in psychological research (Arnett, 2008; Thalmayer et al., 2021; see also Lin & Li, 2023). Thus, we cannot speak to the question of whether the developmental patterns we observed can be considered universal.

Conclusion

The present study investigated age trajectories for the Big Five traits using a statistical approach that was exclusively based on within-person changes. By using such an approach, we were able to avoid between-person confounding (e.g., cohort effects), leading to the better identification and a more reliable estimation of age effects. With decreases in Neuroticism and increases in Conscientiousness as well as in Agreeableness, the results confirmed the so-called personality maturation in younger adulthood but also showed that single items can show different developmental trends even when they pertain to the same personality dimension. This tendency may provide a partial explanation for the heterogeneous findings in previous studies.

To better recover age effects, we recommend that future research should routinely investigate personality development by analyzing longitudinal data with statistical models that rely on only within-person changes. In addition, age trajectories should be examined not only at the level of broad personality dimensions but also at the level of facets and items, as these may exhibit distinctly different aging patterns. We believe that these two straightforward steps could move the field toward a more cumulative mode of science—and broaden and refine the field's understanding of personality development.

References

- Allison, P. D. (2009). *Fixed effects regression models*. SAGE Publications.
- Allison, P. D., Williams, R., & Moral-Benito, E. (2017). Maximum likelihood for cross-lagged panel models with fixed effects. *Socius: Sociological Research for a Dynamic World*, 3, 1–17. <https://doi.org/10.1177/2378023117710578>
- Alt, H., & Godau, M. (1995). Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(1–2), 75–91. <https://doi.org/10.1142/S0218195995000064>
- Andersen, H. K. (2022). A closer look at random and fixed effects panel regression in structural equation modeling using lavaan. *Structural Equation Modeling: A Multidisciplinary Journal*, 29(3), 476–486. <https://doi.org/10.1080/10705511.2021.1963255>
- Anglim, J., Horwood, S., Smillie, L. D., Marrero, R. J., & Wood, J. K. (2020). Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychological Bulletin*, 146(4), 279–323. <https://doi.org/10.1037/bul0000226>
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. <https://doi.org/10.1515/9781400829828>
- Anusic, I., Lucas, R. E., & Donnellan, M. B. (2012). Cross-sectional age differences in personality: Evidence from nationally representative samples from Switzerland and the United States. *Journal of Research in Personality*, 46(1), 116–120. <https://doi.org/10.1016/j.jrp.2011.11.002>
- Anvari, F., Efendić, E., Olsen, J., Arslan, R. C., Elson, M., & Schneider, I. K. (2022). Bias in self-reports: An initial elevation phenomenon. *Social Psychological and Personality Science*. Advance online publication. <https://doi.org/10.1177/19485506221129160>

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018).

Journal Article Reporting Standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. <https://doi.org/10.1037/amp0000191>

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford*

Bulletin of Economics and Statistics, *49*(4), 431–434. <https://doi.org/10.1111/j.1468-0084.1987.mp49004006.x>

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less

American. *American Psychologist*, *63*(7), 602–614. <https://doi.org/10.1037/0003-066X.63.7.602>

Arslan, R. C., Reitz, A. K., Driebe, J. C., Gerlach, T. M., & Penke, L. (2021). Routinely

randomize potential sources of measurement reactivity to estimate and adjust for biases in subjective reports. *Psychological Methods*, *26*(2), 175–185.

<https://doi.org/10.1037/met0000294>

Ashton, M. C., & Lee, K. (2016). Age trends in HEXACO-PI-R self-reports. *Journal of*

Research in Personality, *64*, 102–111. <https://doi.org/10.1016/j.jrp.2016.08.008>

Asselmann, E., & Specht, J. (2021). Testing the social investment principle around childbirth:

Little evidence for personality maturation before and after becoming a parent. *European Journal of Personality*, *35*(1), 85–102. <https://doi.org/10.1002/per.2269>

Ausmees, L., Kandler, C., Realo, A., Allik, J., Borkenau, P., Hřebíčková, M., & Mõttus, R.

(2022). Age differences in personality traits and social desirability: A multi-rater multi-sample study. *Journal of Research in Personality*, *99*, Article 104245.

<https://doi.org/10.1016/j.jrp.2022.104245>

Beck, E. D., & Jackson, J. J. (2020). Idiographic traits: A return to Allportian approaches to personality. *Current Directions in Psychological Science*, 29(3), 301–308.

<https://doi.org/10.1177/0963721420915860>

Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, 122(3), 523–553.

<https://doi.org/10.1037/pspp0000386>

Bell, A. (2020). Age period cohort analysis: A review of what we should and shouldn't do. *Annals of Human Biology*, 47(2), 208–217.

<https://doi.org/10.1080/03014460.2019.1707872>

Bell, A., & Jones, K. (2014). Current practice in the modelling of age, period and cohort effects with panel data: A commentary on Tawfik et al. (2012), Clarke et al. (2009), and McCulloch (2012). *Quality & Quantity*, 48(4), 2089–2095.

<https://doi.org/10.1007/s11135-013-9881-x>

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3(1), 133–153.

<https://doi.org/10.1017/psrm.2014.7>

Bell, R. Q. (1953). Convergence: An accelerated longitudinal approach. *Child Development*, 24(2), 145–152. <https://doi.org/10.2307/1126345>

Benet-Martínez, V., & John, O. P. (1998). *Los Cinco Grandes* across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, 75(3), 729–750. <https://doi.org/10.1037/0022-3514.75.3.729>

- Bleidorn, W., Hopwood, C. J., Back, M. D., Denissen, J. J. A., Hennecke, M., Hill, P. L., Jokela, M., Kandler, C., Lucas, R. E., Luhmann, M., Orth, U., Roberts, B. W., Wagner, J., Wrzus, C., & Zimmermann, J. (2021). Personality trait stability and change. *Personality Science*, 2, Article e6009. <https://doi.org/10.5964/ps.6009>
- Bleidorn, W., Hopwood, C. J., & Lucas, R. E. (2018). Life events and personality trait change. *Journal of Personality*, 86(1), 83–96. <https://doi.org/10.1111/jopy.12286>
- Bleidorn, W., Klimstra, T. A., Denissen, J. J. A., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2013). Personality maturation around the world: A cross-cultural examination of social-investment theory. *Psychological Science*, 24(12), 2530–2540. <https://doi.org/10.1177/0956797613498396>
- Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 148(7–8), 588–619. <https://doi.org/10.1037/bul0000365>
- Borsboom, D. (2023). Psychological constructs as organizing principles. In L. A. van der Ark, W. H. M. Emons, & R. R. Meijer (Eds.), *Essays on contemporary psychometrics* (pp. 89–108). Springer. https://doi.org/10.1007/978-3-031-10370-4_5
- Boyce, C. J., Wood, A. M., & Powdthavee, N. (2013). Is personality fixed? Personality changes as much as "variable" economic factors and more strongly predicts changes to life satisfaction. *Social Indicators Research*, 111(1), 287–305. <https://doi.org/10.1007/s11205-012-0006-z>
- Brandt, N. D., Becker, M., Tetzner, J., Brunner, M., Kuhl, P., & Maaz, K. (2020). Personality across the lifespan: Exploring measurement invariance of a short Big Five inventory from

- ages 11 to 84. *European Journal of Psychological Assessment*, 36(1), 162–173.
<https://doi.org/10.1027/1015-5759/a000490>
- Brandt, N. D., Drewelies, J., Willis, S. L., Schaie, K. W., Ram, N., Gerstorf, D., & Wagner, J. (2022). Acting like a Baby Boomer? Birth-cohort differences in adults' personality trajectories during the last half a century. *Psychological Science*, 33(3), 382–396.
<https://doi.org/10.1177/09567976211037971>
- Brüderl, J., & Ludwig, V. (2015). Fixed-effects panel regression. In H. Best & C. Wolf (Eds.), *The SAGE handbook of regression analysis and causal inference* (pp. 327–358). SAGE Publications. <https://doi.org/10.4135/9781446288146.n15>
- Cameron, C. A., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *The Journal of Human Resources*, 50(2), 317–372. <https://doi.org/10.3368/jhr.50.2.317>
- Cao, C., & Liang, X. (2022). The impact of model size on the sensitivity of fit measures in measurement invariance testing. *Structural Equation Modeling: A Multidisciplinary Journal*, 29(5), 744–754. <https://doi.org/10.1080/10705511.2022.2056893>
- Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology*, 56(1), 453–484.
<https://doi.org/10.1146/annurev.psych.55.090902.141913>
- Cerino, E. S., Schneider, S., Stone, A. A., Sliwinski, M. J., Mogle, J., & Smyth, J. M. (2022). Little evidence for consistent initial elevation bias in self-reported momentary affect: A coordinated analysis of ecological momentary assessment studies. *Psychological Assessment*, 34(5), 467–482. <https://doi.org/10.1037/pas0001108>

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance.

Structural Equation Modeling: A Multidisciplinary Journal, 14(3), 464–504.

<https://doi.org/10.1080/10705510701301834>

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing

measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*,

9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Clancy, K. B., & Davis, J. L. (2019). Soylent is people, and WEIRD is white: Biological

anthropology, whiteness, and the limits of the WEIRD. *Annual Review of Anthropology*,

48(1), 169–186. <https://doi.org/10.1146/annurev-anthro-102218-011133>

Cobb-Clark, D. A., & Schurer, S. (2012). The stability of Big-Five personality traits. *Economics*

Letters, 115(1), 11–15. <https://doi.org/10.1016/j.econlet.2011.11.015>

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum

Associates. <https://doi.org/10.4324/9780203771587>

Costa, P. T., Jr., McCrae, R. R., & Löckenhoff, C. E. (2019). Personality across the life span.

Annual Review of Psychology, 70(1), 423–448. [https://doi.org/10.1146/annurev-psych-](https://doi.org/10.1146/annurev-psych-010418-103244)

010418-103244

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of*

Statistical Software, 27(2), 1–43. <https://doi.org/10.18637/jss.v027.i02>

Curran, P. J. (2003). Have multilevel models been structural equation models all along?

Multivariate Behavioral Research, 38(4), 529–569.

https://doi.org/10.1207/s15327906mbr3804_5

- Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, *62*(1), 583–619. <https://doi.org/10.1146/annurev.psych.093008.100356>
- Dehne, M., & Schupp, J. (2007). *Persönlichkeitsmerkmale im Sozio-oekonomischen Panel (SOEP)—Konzept, Umsetzung und empirische Eigenschaften* [Personality traits in the socio-economic panel (SOEP)—Concept, implementation, and empirical properties] (Research Notes 26). DIW Berlin.
- den Boer, L., Klimstra, T. A., Branje, S. J., Meeus, W. H., & Denissen, J. J. (2019). Personality maturation during the transition to working life: Associations with commitment as a possible indicator of social investment. *European Journal of Personality*, *33*(4), 456–467. <https://doi.org/10.1002/per.2218>
- Denissen, J. J. A., Luhmann, M., Chung, J. M., & Bleidorn, W. (2019). Transactions between life events and personality traits across the adult lifespan. *Journal of Personality and Social Psychology*, *116*(4), 612–633. <https://doi.org/10.1037/pspp0000196>
- Department of Social Services & Melbourne Institute of Applied Economic and Social Research. (2018). *The Household, Income and Labour Dynamics in Australia (HILDA) Survey: General Release 17 (Waves 1–17)* [Data set]. Australian Data Archive. <https://doi.org/10.26193/ptklyp>
- Deventer, J., Lüdtke, O., Nagy, G., Retelsdorf, J., & Wagner, J. (2019). Against all odds—Is a more differentiated view of personality development in emerging adulthood needed? The case of young apprentices. *British Journal of Psychology*, *110*(1), 60–86. <https://doi.org/10.1111/bjop.12336>

- Donnellan, M. B., & Lucas, R. E. (2008). Age differences in the Big Five across the life span: Evidence from two national samples. *Psychology and Aging, 23*(3), 558–566.
<https://doi.org/10.1037/a0012897>
- Fitzenberger, B., Mena, G., Nimczik, J., & Sunde, U. (2022). Personality traits across the life cycle: Disentangling age, period and cohort effects. *The Economic Journal, 132*(646), 2141–2172. <https://doi.org/10.1093/ej/ueab093>
- Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29–51. <https://doi.org/10.1037//0033-2909.95.1.29>
- Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171–191. <https://doi.org/10.1037//0033-2909.101.2.171>
- Fréchet, M. M. (1906). Sur quelques points du calcul fonctionnel [On some points of the functional calculus]. *Rendiconti Del Circolo Matematico Di Palermo, 22*(1), 1–72.
<https://doi.org/10.1007/BF03018603>
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168.
<https://doi.org/10.1177/2515245919847202>
- Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91.
<https://doi.org/10.1037/a0032138>
- Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S., & Subtil, F. (2016). kmlShape: An efficient method to cluster longitudinal data (time-series) according to their shapes. *PLoS ONE, 11*(6), Article e0150738. <https://doi.org/10.1371/journal.pone.0150738>

- Gerlitz, J.-Y., & Schupp, J. (2005). *Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP* [The measurement of the Big Five personality traits in the SOEP] (Research Notes 4). DIW Berlin.
- Giesselmann, M., Bohmann, S., Goebel, J., Krause, P., Liebau, E., Richter, D., Schacht, D., Schröder, C., Schupp, J., & Liebig, S. (2019). The individual in context(s): Research potentials of the Socio-Economic Panel Study (SOEP) in sociology. *European Sociological Review*, 35(5), 738–755. <https://doi.org/10.1093/esr/jcz029>
- Glenn, N. D. (2003). Distinguishing age, period, and cohort effects. In J. T. Mortimer & M. J. Shanahan (Eds.), *Handbook of the life course* (pp. 465–476). Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-0-306-48247-2_21
- Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment*, 27(2), 404–418. <https://doi.org/10.1177/1073191117746503>
- Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Journal of Economics and Statistics*, 239(2), 345–360. <https://doi.org/10.1515/jbnst-2018-0022>
- Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1), 26–42. <https://doi.org/10.1037/1040-3590.4.1.26>
- Graham, E. K., Weston, S. J., Gerstorf, D., Yoneda, T. B., Booth, T., Beam, C. R., Petkus, A. J., Drewelies, J., Hall, A. N., Bastarache, E. D., Estabrook, R., Katz, M. J., Turiano, N. A., Lindenberger, U., Smith, J., Wagner, G. G., Pedersen, N. L., Allemand, M., Spiro, A., III, . . . Mroczek, D. K. (2020). Trajectories of Big Five personality traits: A coordinated

- analysis of 16 longitudinal samples. *European Journal of Personality*, *34*(3), 301–321.
<https://doi.org/10.1002/per.2259>
- Grosz, M. P. (2020). *The factor structure of Big Five personality trait measures at the between- and within-person levels*. PsyArXiv. <https://doi.org/10.31234/osf.io/k6r7g>
- Hahn, E., Gottschling, J., & Spinath, F. M. (2012). Short measurements of personality—Validity and reliability of the GSOEP Big Five Inventory (BFI-S). *Journal of Research in Personality*, *46*(3), 355–359. <https://doi.org/10.1016/j.jrp.2012.03.008>
- Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379.
<https://doi.org/10.1037/met0000239>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature*, *466*(7302), 29. <https://doi.org/10.1038/466029a>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.
<https://doi.org/10.1017/S0140525X0999152X>
- 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*(2–3), 257–278.
<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*(2), 87–102.

- Hilgert, L., Kroh, M., & Richter, D. (2016). The effect of face-to-face interviewing on personality measurement. *Journal of Research in Personality, 63*, 133–136.
<https://doi.org/10.1016/j.jrp.2016.05.006>
- Hsu, H.-Y., Kwok, O.-M., Lin, J. H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A Monte Carlo study. *Multivariate Behavioral Research, 50*(2), 197–215. <https://doi.org/10.1080/00273171.2014.977429>
- 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), 1–55. <https://doi.org/10.1080/10705519909540118>
- Hülür, G. (2017). Cohort differences in personality. In J. Specht (Ed.), *Personality development across the lifespan* (pp. 519–536). Academic Press. <https://doi.org/10.1016/B978-0-12-804674-6.00031-4>
- Ion, A., Gunnesch-Luca, G., Petre, D., & Iliescu, D. (2022). Secular changes in personality: An age-period-cohort analysis. *Journal of Research in Personality, 100*, Article 104280.
<https://doi.org/10.1016/j.jrp.2022.104280>
- Jackson, J. J., Bogg, T., Walton, K. E., Wood, D., Harms, P. D., Lodi-Smith, J., Edmonds, G. W., & 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*(2), 446–459.
<https://doi.org/10.1037/a0014156>
- John, O. P. (2021). History, measurement, and conceptual elaboration of the Big-Five trait taxonomy: The paradigm matures. In O. P. John & R. W. Robins (Eds.), *Handbook of personality: Theory and research* (4th ed., pp. 35–82). Guilford Press.

- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). Guilford Press.
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). Guilford Press.
- Jokela, M., Pekkarinen, T., Sarvimäki, M., Terviö, M., & Uusitalo, R. (2017). Secular rise in economically valuable personality traits. *Proceedings of the National Academy of Sciences*, *114*(25), 6527–6532. <https://doi.org/10.1073/pnas.1609994114>
- Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling* (Version 0.5-5) [R package]. <https://CRAN.R-project.org/package=semTools>
- Kandler, C., Zimmermann, J., & McAdams, D. P. (2014). Core and surface characteristics for the description and theory of personality differences and development. *European Journal of Personality*, *28*(3), 231–243. <https://doi.org/10.1002/per.1952>
- Kazak, A. E. (2018). Editorial: Journal Article Reporting Standards. *American Psychologist*, *73*(1), 1–2. <https://doi.org/10.1037/amp0000263>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford Press.
- Kornadt, A. E., Hagemeyer, B., Neyer, F. J., & Kandler, C. (2018). Sound body, sound mind? The interrelation between health change and personality change in old age. *European Journal of Personality*, *32*(1), 30–45. <https://doi.org/10.1002/per.2135>

- Krämer, M. D., van Scheppingen, M. A., Chopik, W. J., & Richter, D. (2022). The transition to grandparenthood: No consistent evidence for change in the Big Five personality traits and life satisfaction. *European Journal of Personality*. Advance online publication. <https://doi.org/10.1177/08902070221118443>
- Kratz, F., & Brüderl, J. (2021). *The age trajectory of happiness: How lack of causal reasoning has produced the myth of a U-shaped age-happiness trajectory*. PsyArXiv. <https://doi.org/10.31234/osf.io/d8f2z>
- Lang, F. R., John, D., Lüdtke, O., Schupp, J., & Wagner, G. G. (2011). Short assessment of the Big Five: Robust across survey methods except telephone interviewing. *Behavior Research Methods*, 43(2), 548–567. <https://doi.org/10.3758/s13428-011-0066-z>
- Liebig, S., Goebel, J., Grabka, M., Schröder, C., Zinn, S., Bartels, C., Fedorets, A., Franken, A., Gerike, M., Griese, F., Jacobsen, J., Kara, S., König, J., Krause, P., Kröger, H., Liebau, E., Metzinger, M., Nebelin, J., Petrenz, M., . . . Zimmermann, S. (2021). *Socio-Economic Panel (SOEP): Data for years 1984–2019: SOEP-Core* (Version 36, EU Edition) [Data set]. DIW Berlin. <https://doi.org/10.5684/soep.core.v36eu>
- Lin, Z., & Li, N. (2023). Global diversity of authors, editors, and journal ownership across subdisciplines of psychology: Current state and policy implications. *Perspectives on Psychological Science*, 18(2), 358–377. <https://doi.org/10.1177/17456916221091831>
- Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.
- Lodi-Smith, J., & Roberts, B. W. (2007). Social investment and personality: A meta-analysis of the relationship of personality traits to investment in work, family, religion, and volunteerism. *Personality and Social Psychology Review*, 11(1), 68–86. <https://doi.org/10.1177/1088868306294590>

- Losoncz, I. (2009). Personality traits in HILDA. *Australian Social Policy*, 8, 169–198.
- Lucas, R. E., & Donnellan, M. B. (2009). Age differences in personality: Evidence from a nationally representative Australian sample. *Developmental Psychology*, 45(5), 1353–1363. <https://doi.org/10.1037/a0013914>
- Lucas, R. E., & Donnellan, M. B. (2011). Personality development across the life span: Longitudinal analyses with a national sample from Germany. *Journal of Personality and Social Psychology*, 101(4), 847–861. <https://doi.org/10.1037/a0024298>
- Luo, J., Zhang, B., Estabrook, R., Graham, E. K., Driver, C. C., Schalet, B. D., Turiano, N. A., Spiro, A., III, & Mroczek, D. K. (2022). Personality and health: Disentangling their between-person and within-person relationship in three longitudinal studies. *Journal of Personality and Social Psychology*, 122(3), 493–522. <https://doi.org/10.1037/pspp0000399>
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1), 19–40. <https://doi.org/10.1037//1082-989X.7.1.19>
- Marchand, M. (2018). *LISS Core Study: Personality: Wave 10* (Version 1.0) [Data set]. CentERdata. <https://doi.org/10.17026/dans-xsm-uayn>
- Marchand, M. (2019). *LISS Core Study: Personality: Wave 11* (Version 1.0) [Data set]. CentERdata. <https://doi.org/10.17026/dans-zmj-5egt>
- Marchand, M. (2020). *LISS Core Study: Personality: Wave 12* (Version 1.0) [Data set]. CentERdata. <https://doi.org/10.17026/dans-z2c-fzcd>
- 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(3), 471–491.
<https://doi.org/10.1037/a0019227>
- Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of Big-Five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, 49(6), 1194–1218.
<https://doi.org/10.1037/a0026913>
- McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, 38(1), 115–142.
<https://doi.org/10.1037//0012-1649.38.1.115>
- McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, 19(2), 97–112.
<https://doi.org/10.1177/1088868314541857>
- McCrae, R. R., Costa, P. T., Jr., Ostendorf, F., Angleitner, A., Hřebíčková, M., Avia, M. D., Sanz, J., Sánchez-Bernardos, M. L., Kusdil, M. E., Woodfield, R., Saunders, P. R., & Smith, P. B. (2000). Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology*, 78(1), 173–186.
<https://doi.org/10.1037/0022-3514.78.1.173>
- McCrae, R. R., & Mõttus, R. (2019). What personality scales measure: A new psychometrics and its implications for theory and assessment. *Current Directions in Psychological Science*, 28(4), 415–420. <https://doi.org/10.1177/0963721419849559>
- McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making

- recommendations. *Psychological Methods*, 24(1), 20–35.
<https://doi.org/10.1037/met0000182>
- McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, 28(1), 61–88. <https://doi.org/10.1037/met0000425>
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93(3), 568–592. <https://doi.org/10.1037/0021-9010.93.3.568>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203–240). American Psychological Association. <https://doi.org/10.1037/10409-007>
- Milojev, P., & Sibley, C. G. (2017). Normative personality trait development in adulthood: A 6-year cohort-sequential growth model. *Journal of Personality and Social Psychology*, 112(3), 510–526. <https://doi.org/10.1037/pspp0000121>
- Mirowsky, J., & Kim, J. (2007). Graphing age trajectories: Vector graphs, synthetic and virtual cohort projections, and cross-sectional profiles of depression. *Sociological Methods & Research*, 35(4), 497–541. <https://doi.org/10.1177/0049124106296015>
- Miyazaki, Y., & Raudenbush, S. W. (2000). Tests for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods*, 5(1), 44–63.
<https://doi.org/10.1037/1082-989X.5.1.44>
- Möttus, 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*(3), 474–490.
<https://doi.org/10.1037/pspp0000100>
- Möttus, R., Realo, A., Allik, J., Esko, T., Metspalu, A., & Johnson, W. (2015). Within-trait heterogeneity in age group differences in personality domains and facets: Implications for the development and coherence of personality traits. *PLoS ONE*, *10*(3), Article e0119667.
<https://doi.org/10.1371/journal.pone.0119667>
- Möttus, R., & Rozgonjuk, D. (2021). Development is in the details: Age differences in the Big Five domains, facets, and nuances. *Journal of Personality and Social Psychology*, *120*(4), 1035–1048. <https://doi.org/10.1037/pspp0000276>
- Möttus, R., Sinick, J., Terracciano, A., Hřebíčková, M., Kandler, C., Ando, J., Mortensen, E. L., Colodro-Conde, L., & Jang, K. L. (2019). 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*, *117*(4), e35–e50. <https://doi.org/10.1037/pspp0000202>
- Mueller, S., Wagner, J., Smith, J., Voelkle, M. C., & Gerstorf, D. (2018). The interplay of personality and functional health in old and very old age: Dynamic within-person interrelations across up to 13 years. *Journal of Personality and Social Psychology*, *115*(6), 1127–1147. <https://doi.org/10.1037/pspp0000173>
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398. <https://doi.org/10.1177/0049124194022003006>
- 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*(4), 473–492. <https://doi.org/10.1111/jopy.12173>

- Olaru, G., & Allemand, M. (2022). Correlated personality change across time and age. *European Journal of Personality, 36*(5), 729–749. <https://doi.org/10.1177/08902070211014054>
- Olaru, G., Schroeders, U., Hartung, J., & Wilhelm, O. (2019). Ant colony optimization and local weighted structural equation modeling: A tutorial on novel item and person sampling procedures for personality research. *European Journal of Personality, 33*(3), 400–419. <https://doi.org/10.1002/per.2195>
- 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*(3), 264–278. <https://doi.org/10.1002/per.2185>
- Olaru, G., Stieger, M., Rügger, D., Kowatsch, T., Flückiger, C., Roberts, B. W., & Allemand, M. (2022). Personality change through a digital-coaching intervention: Using measurement invariance testing to distinguish between trait domain, facet, and nuance change. *European Journal of Personality*. Advance online publication. <https://doi.org/10.1177/08902070221145088>
- Paulewicz, B., & Blaut, A. (2022). *The general causal cumulative model of ordinal response*. PsyArXiv. <https://doi.org/10.31234/osf.io/e7a3x>
- Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81*(3), 524–539. <https://doi.org/10.1037/0022-3514.81.3.524>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*(5), 754–775. <https://doi.org/10.1037/0021-9010.84.5.754>

Richter, D., Rohrer, J., Metzger, M., Nestler, W., Weinhardt, M., & Schupp, J. (2017). *SOEP scales manual (updated for SOEP-Core v32.1)* (SOEP Survey Papers: Series C 423). DIW Berlin. <http://hdl.handle.net/10419/156115>

Roberts, B. W., & Nickel, L. B. (2017). A critical evaluation of the Neo-Socioanalytic Model of personality. In J. Specht (Ed.), *Personality development across the lifespan* (pp. 157–177). Academic Press. <https://doi.org/10.1016/B978-0-12-804674-6.00011-9>

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), 1–25. <https://doi.org/10.1037/0033-2909.132.1.1>

Roberts, B. W., Wood, D., & Caspi, A. (2008). The development of personality traits in adulthood. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 375–398). Guilford Press.

Roberts, B. W., Wood, D., & Smith, J. L. (2005). Evaluating Five Factor Theory and social investment perspectives on personality trait development. *Journal of Research in Personality, 39*(1), 166–184. <https://doi.org/10.1016/j.jrp.2004.08.002>

Robitzsch, A. (2022). *sirt: Supplementary item response theory models* (Version 3.12-41) [R package]. <https://CRAN.R-project.org/package=sirt>

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>

- Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. <https://doi.org/10.1177/0013164413498257>
- Rüttenauer, T., & Ludwig, V. (2023). Fixed effects individual slopes: Accounting and testing for heterogeneous effects in panel data or other multilevel models. *Sociological Methods & Research, 52*(1), 43–84. <https://doi.org/10.1177/0049124120926211>
- Salthouse, T. A. (2019). Trajectories of normal cognitive aging. *Psychology and Aging, 34*(1), 17–24. <https://doi.org/10.1037/pag0000288>
- Saucier, G. (1992). Openness versus Intellect: Much ado about nothing? *European Journal of Personality, 6*(5), 381–386. <https://doi.org/10.1002/per.2410060506>
- Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment, 63*(3), 506–516. https://doi.org/10.1207/s15327752jpa6303_8
- Scherpenzeel, A. (2011). Data collection in a probability-based internet panel: How the LISS Panel was built and how it can be used. *Bulletin of Sociological Methodology/Bulletin De Méthodologie Sociologique, 109*(1), 56–61. <https://doi.org/10.1177/0759106310387713>
- Scherpenzeel, A., & Das, M. (2011). "True" longitudinal and probability-based internet panels: Evidence from the Netherlands. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 77–104). Routledge.
- Schröder, C., König, J., Fedorets, A., Goebel, J., Grabka, M. M., Lüthen, H., Metzinger, M., Schikora, F., & Liebig, S. (2020). The economic research potentials of the German

Socio-Economic Panel study. *German Economic Review*, 21(3), 335–371.

<https://doi.org/10.1515/ger-2020-0033>

Schwaba, T. (2019). The structure, measurement, and development of Openness to Experience across adulthood. In D. P. McAdams, R. L. Shiner, & J. L. Tackett (Eds.), *The handbook of personality development* (pp. 185–200). Guilford Press.

Schwaba, T., & Bleidorn, W. (2018). Individual differences in personality change across the adult life span. *Journal of Personality*, 86(3), 450–464.

<https://doi.org/10.1111/jopy.12327>

Schwaba, T., Bleidorn, W., Hopwood, C. J., Manuck, S. B., & Wright, A. G. C. (2022). Refining the maturity principle of personality development by examining facets, close others, and comaturation. *Journal of Personality and Social Psychology*, 122(5), 942–958.

<https://doi.org/10.1037/pspp0000400>

Schwaba, T., Luhmann, M., Denissen, J. J. A., Chung, J. M., & Bleidorn, W. (2018). Openness to experience and culture-openness transactions across the lifespan. *Journal of Personality and Social Psychology*, 115(1), 118–136.

<https://doi.org/10.1037/pspp0000150>

Seifert, I. S., Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2022). The development of the rank-order stability of the Big Five across the life span. *Journal of Personality and Social Psychology*, 122(5), 920–941. <https://doi.org/10.1037/pspp0000398>

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences*, 115(1), E15–E23.

<https://doi.org/10.1073/pnas.1712277115>

- Sliwinski, M., Hoffman, L., & Hofer, S. M. (2010). Evaluating convergence of within-person change and between-person age differences in age-heterogeneous longitudinal studies. *Research in Human Development, 7*(1), 45–60.
<https://doi.org/10.1080/15427600903578169>
- Smits, I. A. M., Dolan, C. V., Vorst, H. C. M., Wicherts, J. M., & Timmerman, M. E. (2011). Cohort differences in Big Five personality factors over a period of 25 years. *Journal of Personality and Social Psychology, 100*(6), 1124–1138.
<https://doi.org/10.1037/a0022874>
- Soto, C. J., & John, O. P. (2012). Development of Big Five domains and facets in adulthood: Mean-level age trends and broadly versus narrowly acting mechanisms. *Journal of Personality, 80*(4), 881–914. <https://doi.org/10.1111/j.1467-6494.2011.00752.x>
- 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*(2), 330–348. <https://doi.org/10.1037/a0021717>
- Specht, J., Bleidorn, W., Denissen, J. J. A., Hennecke, M., Hutteman, R., Kandler, C., Luhmann, M., Orth, U., Reitz, A. K., & Zimmermann, J. (2014). What drives adult personality development? A comparison of theoretical perspectives and empirical evidence. *European Journal of Personality, 28*(3), 216–230.
<https://doi.org/10.1002/per.1966>
- 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*(4), 862–882.
<https://doi.org/10.1037/a0024950>

- 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*(5), 1041–1053. <https://doi.org/10.1037/0022-3514.84.5.1041>
- Stewart, R. D., Möttus, R., Seeboth, A., Soto, C. J., & Johnson, W. (2022). The finer details? The predictability of life outcomes from Big Five domains, facets, and nuances. *Journal of Personality*, *90*(2), 167–182. <https://doi.org/10.1111/jopy.12660>
- Summerfield, M., Bevitt, A., Fok, K., Hahn, M., La, N., Macalalad, N., O’Shea, M., Watson, N., Wilkins, R., & Wooden, M. (2018). *HILDA user manual—Release 17*. Melbourne Institute: Applied Economic and Social Research.
- Terracciano, A., McCrae, R. R., Brant, L. J., & Costa, P. T., Jr. (2005). Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging*, *20*(3), 493–506. <https://doi.org/10.1037/0882-7974.20.3.493>
- Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American? *The American Psychologist*, *76*(1), 116–129. <https://doi.org/10.1037/amp0000622>
- Townsend, Z., Buckley, J., Harada, M., & Scott, M. A. (2013). The choice between fixed and random effects. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *The SAGE handbook of multilevel modeling* (pp. 73–88). SAGE Publications. <https://doi.org/10.4135/9781446247600.n5>
- van Scheppingen, M. A., Jackson, J. J., Specht, J., Hutteman, R., Denissen, J. J. A., & Bleidorn, W. (2016). Personality trait development during the transition to parenthood: A test of social investment theory. *Social Psychological and Personality Science*, *7*(5), 452–462. <https://doi.org/10.1177/1948550616630032>

- 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*(1), 4–70. <https://doi.org/10.1177/109442810031002>
- Wagner, J., Lüdtke, O., & Robitzsch, A. (2019). Does personality become more stable with age? Disentangling state and trait effects for the Big Five across the life span using local structural equation modeling. *Journal of Personality and Social Psychology, 116*(4), 666–680. <https://doi.org/10.1037/pspp0000203>
- Wang, L., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods, 20*(1), 63–83. <https://doi.org/10.1037/met0000030>
- Watson, N., & Wooden, M. (2012). The HILDA Survey: A case study in the design and development of a successful household panel survey. *Longitudinal and Life Course Studies, 3*(3), 369–381. <https://doi.org/10.14301/lcs.v3i3.208>
- Watson, N., & Wooden, M. (2013). Adding a top-up sample to the Household, Income and Labour Dynamics in Australia Survey. *The Australian Economic Review, 46*(4), 489–498. <https://doi.org/10.1111/1467-8462.12027>
- Watson, N., & Wooden, M. (2021). The Household, Income and Labour Dynamics in Australia (HILDA) Survey. *Journal of Economics and Statistics, 241*(1), 131–141. <https://doi.org/10.1515/jbnst-2020-0029>
- 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*(3), 39–47. <https://doi.org/10.1111/j.1745-3992.2010.00182.x>

- Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4*(1), 10–18. <https://doi.org/10.1111/j.1750-8606.2009.00110.x>
- Wortman, J., Lucas, R. E., & Donnellan, M. B. (2012). Stability and change in the Big Five personality domains: Evidence from a longitudinal study of Australians. *Psychology and Aging, 27*(4), 867–874. <https://doi.org/10.1037/a0029322>
- Wu, W., & Lang, K. M. (2016). Proportionality assumption in latent basis curve models: A cautionary note. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(1), 140–154. <https://doi.org/10.1080/10705511.2014.938578>
- Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E. (2020). From data to causes II: Comparing approaches to panel data analysis. *Organizational Research Methods, 23*(4), 688–716. <https://doi.org/10.1177/1094428119847280>