

Seifert, Ingo S.; Rohrer, Julia M.; Egloff, Boris; Schmukle, Stefan Christian

Working Paper

The development of the rank-order stability of the Big Five across the life span

SOEPpapers on Multidisciplinary Panel Data Research, No. 1156

Provided in Cooperation with:

German Institute for Economic Research (DIW Berlin)

Suggested Citation: Seifert, Ingo S.; Rohrer, Julia M.; Egloff, Boris; Schmukle, Stefan Christian (2021) : The development of the rank-order stability of the Big Five across the life span, SOEPpapers on Multidisciplinary Panel Data Research, No. 1156, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/249156>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

1156²⁰²¹

SOEP papers
on Multidisciplinary Panel Data Research

The Development of the Rank-Order Stability of the Big Five Across the Life Span

Ingo S. Seifert, Julia M. Rohrer, Boris Egloff, 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:

Jan **Goebel** (Spatial Economics)
Stefan **Liebig** (Sociology)
David **Richter** (Psychology)
Carsten **Schröder** (Public Economics)
Jürgen **Schupp** (Sociology)
Sabine **Zinn** (Statistics)

Conchita **D'Ambrosio** (Public Economics, DIW Research Fellow)
Denis **Gerstorff** (Psychology, DIW Research Fellow)
Katharina **Wrohlich** (Gender Economics)
Martin **Kroh** (Political Science, Survey Methodology)
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)

ISSN: 1864-6689 (online)

German Socio-Economic Panel (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: soeppapers@diw.de



The Development of the Rank-Order Stability of the Big Five Across the Life Span

Ingo S. Seifert¹, Julia M. Rohrer¹, Boris Egloff², and Stefan C. Schmukle¹

¹ Department of Psychology, Leipzig University

² Department of Psychology, Johannes Gutenberg-University Mainz

Date of submission: January 4, 2021

Date first revision submitted: June 18, 2021

Date second revision submitted: July 12, 2021


Accepted: July 18, 2021

© 2021, 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/pspp0000398


Seifert, I. S., Rohrer, J. M., Egloff, B., & Schmukle, S. C. (in press). The development of the rank-order stability of the Big Five across the life span. *Journal of Personality and Social Psychology*.


Supplemental material: <https://osf.io/th3ja/>

Author Note

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

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

Boris Egloff  <https://orcid.org/0000-0002-5736-9912>

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/rzfqm/>. On the OSF, we share all the analytic methods and code necessary to reproduce our results directly from the original data sets without any additional steps of data preparation. On the OSF, we further share the research material for Studies 1 and 2. We are not allowed to make the data from Studies 1 and 2 publicly available, but on the OSF, we provide information about how researchers can request these data for research purposes.

This paper uses unit record data from Household, Income and Labour Dynamics in Australia (HILDA) Survey conducted by the Australian Government Department of Social Services (DSS). The Socio-Economic Panel (SOEP) data were made available by the German Institute for Economic Research (DIW). We thank these institutions for providing these data sets. However, 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, or DIW. This research was supported by funding for doctoral candidates from Leipzig University to Ingo Seifert. We thank the Center for Information Services and High Performance Computing (ZIH) at TU Dresden for generous allocations of computer time. Computations for this work were done (in part) using resources from the Leipzig University Computing Centre. We are grateful to Alexander Robitzsch for his helpful support in the analyses with sirt.

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

Abstract

Several studies have suggested that the rank-order stability of personality increases until midlife and declines later in old age. However, this inverted U-shaped pattern has not consistently emerged in previous research; in particular, a recent investigation implementing several methodological advances failed to support it. To resolve the matter, we analyzed data from two representative panel studies and investigated how certain methodological decisions affect conclusions regarding the age trajectories of stability. The data came from Australia ($N = 15,465$; Study 1) and Germany ($N = 21,777$; Study 2), and each study included four waves of personality assessment. We investigated the life span development of the rank-order stability of the Big Five for 4-, 8-, and 12-year intervals. Whereas Study 1 provided strong evidence for an inverted U-shape with rank-order stability declining past age 50, Study 2 provided more mixed results that nonetheless generally supported the inverted U-shape. This developmental trend held for single personality traits as well as for the overall pattern across traits; and it held for all three retest intervals—both descriptively and in formal tests. Additionally, we found evidence that health-related changes accounted for the decline in rank-order stability in older age. This suggests that if analyses are implicitly conditioned on health (e.g., by excluding participants with missing data on later waves), the decline in stability in old age will be underestimated or even missed. Our results provide further evidence for the inverted U-shaped age pattern in personality stability development but also extend knowledge about the underlying processes.

Keywords: personality development, rank-order stability, Big Five, panel studies, local structural equation modeling

The Development of the Rank-Order Stability of the Big Five Across the Life Span

In recent decades, research has yielded broad consensus that the development of personality across the life span is characterized by both stability *and* change (for reviews, see Caspi et al., 2005; Costa et al., 2019; McAdams & Olson, 2010). Personality development is a faceted process, with several indices referring to various concepts of stability and change (Asendorpf, 2010). For example, past studies have extensively scrutinized consistency and change in the rank order (i.e., individuals' relative position to each other on a trait; e.g., Roberts & DelVecchio, 2000), mean level (i.e., average trait level; e.g., Roberts et al., 2006), structure (i.e., relations among traits; e.g., Lüdtke et al., 2009), and ipsative aspects (i.e., within-person trait patterns; e.g., De Fruyt et al., 2006) of personality.

The development of rank-order stability is particularly important for personality psychology, as rank-order stability captures the extent to which interindividual differences persist over time. This aspect of consistency is therefore also called differential stability, and it is usually assessed by computing the correlation of a given trait across two time points. If everyone's trait value changed by the same amount in the same direction, relative ordering would be preserved, and rank-order stability would be perfect (i.e., $r = 1$), despite any mean-level changes. By contrast, if everyone's trait value changed in a nonsystematic manner, the relative ordering could be completely destroyed, resulting in no rank-order stability (i.e., $r = 0$), even if the mean trait level remained perfectly stable. In other words, rank-order changes are independent of mean-level changes (Block, 1971). As the present article focuses on rank-order stability, we will use the term stability to refer to this concept unless stated otherwise.

Rank-Order Stability of Personality Across Time and Age

The stability of personality tends to be generally high but not perfect (e.g., Anusic & Schimmack, 2016), and estimated stability coefficients decline as the length of the time interval between measurement points increases (Ardelt, 2000; Conley, 1984; Ferguson, 2010; Roberts & DelVecchio, 2000), approaching a nonzero asymptote (Anusic & Schimmack, 2016; Fraley & Roberts, 2005). This highlights the essential trait-like character of personality, which was further corroborated by Damian et al.'s (2019) findings, which indicated that even across 50 years, a modest degree of rank-order stability in personality can be found (cf. Harris et al., 2016).

It is a common empirical finding that rank-order stability increases up to a certain point in midlife (the so-called principle of cumulative continuity; Roberts & Nickel, 2017), without ever reaching perfect stability (Anusic & Schimmack, 2016; Ferguson, 2010; Roberts & DelVecchio, 2000). Meta-analytic work (Ferguson, 2010; Roberts & DelVecchio, 2000) has provided strong evidence for this general pattern but has also highlighted disagreement about the age at which stability ceases to increase (see Costa et al., 2019), with estimates ranging from 30 (Ferguson, 2010; see also Briley & Tucker-Drob, 2014) to 50 years (Roberts & DelVecchio, 2000).

Beyond this increase, meta-analyses have suggested that stability remains at a high level throughout late adulthood (Briley & Tucker-Drob, 2014; Ferguson, 2010; Roberts & DelVecchio, 2000). However, these age trends later in life suffer from a loss of precision due to the rather sparse number of older participants who have been sampled. Indeed, more recent studies relying on age-representative samples found a decline in stability past age 50 (Lucas & Donnellan, 2011; Milojev & Sibley, 2014; Specht et al., 2011; Wortman et al., 2012; see also Ardelt, 2000), resulting in a life span trajectory resembling an inverted U. This decline in

stability has been found not only for all Big Five personality traits (Lucas & Donnellan, 2011; Specht et al., 2011; Wortman et al., 2012) but also for the Honesty-Humility factor from the HEXACO personality framework (Milojev & Sibley, 2014).

What Accounts for the Life Span Trends in Personality Stability?

In their meta-analysis, Briley and Tucker-Drob (2014) found that genetic contributions to stability remain fairly constant across age, and thus, the observed increase in personality stability until adulthood seems to be predominantly explained by increases in environmental stability. Several mechanisms may lead to an increasingly stable environment: For example, personality traits may lead an individual to create certain environments, which, in turn, conserve the respective personality traits (the so-called niche-picking principle; Roberts & Nickel, 2017). As an example, a greater fit with a university environment was found to go along with more stability in students' personality (Roberts & Robins, 2004). Similarly, identity-related processes may increase stability. As people get older, they gain a clearer perception of their self and increasingly select environments (e.g., social roles) that fit with and reinforce their personality dispositions (Roberts & DeVecchio, 2000).

The downward trend in stability in older age could be explained by significant changes in biological, cognitive, and social domains (Ardelt, 2000; Lucas & Donnellan, 2011; Wortman et al., 2012). Major family- and work-related life events (e.g., children moving out, retirement) as well as declines in cognitive functioning and health might *differentially* affect individuals and, hence, impact the stability of interindividual differences in personality. Whereas this explanation is conceptually convincing, *direct* empirical evidence is currently missing. Life events could lead to decreases in stability in old age, first, if their effects are heterogeneous (i.e., leading to a decrease in stability in the group of people experiencing the event) or, second, if their occurrence

or timing varies across respondents, leading to a decrease in stability in the overall sample even if effects are homogeneous. Concerning the first point, Specht et al. (2011) found no conclusive effect of major life events on personality stability; concerning the second, we know of no research that has systematically investigated how variability in the onset of life events affects personality stability. Furthermore, whereas it has been shown that changes in cognitive abilities (e.g., Wettstein et al., 2017) and changes in health (e.g., Kornadt et al., 2018; Mueller et al., 2018) seem to go along with changes in personality, research has yet to directly address whether processes of differential senescence explain the decrease in personality stability in older age.

Recent Advancements in Personality Stability Development

The latest contribution to the age development of personality stability by J. Wagner et al. (2019) enriched the discourse with several refinements but also challenged key findings. J. Wagner et al. aimed to provide “a more nuanced differentiation of stability” (p. 676) by applying the trait-state-occasion (TSO) model (Cole et al., 2005, 2015) to decompose stability into a component that is due to a time-invariant trait and a component that is due to autoregressive paths between occasion-specific factors. Their results generally indicated that stability in personality might be primarily traced back to the effects of the trait component. J. Wagner et al. further investigated how age-related changes in stability reflect age-related changes in these two components of the TSO model: The contributions of both components to stability fluctuated rather unsystematically across the life span with fairly different trajectories across personality factors and samples with, overall, only small effects of age and no consistent support for an increase or decrease in personality stability.

A sound interpretation of effects of age on the time-invariant trait component is, in our view, questionable. In the TSO model, by definition, the time-invariant trait indicates the

component that is truly stable across the whole life and is thus independent of age. In line with this argument, in a similar decomposition based on the Stable Trait, AutoRegressive Trait, and State (STARTS) model (Kenny & Zautra, 1995, 2001), age-related changes in the stability of self-esteem were fully attributed to age-related changes in autoregressive effects, rather than effects of age on the stable trait factor (Donnellan et al., 2012).

However, empirically, the trait component only captures individual differences that are perfectly stable *within the investigated time frame* (Cole et al., 2005, 2015). Consequently, if the time frame of the study is shorter than the duration of a lifetime, then this component will differ between participants of different ages (who differ in respect to personality stability). If the time frame of the study were increased, the trait component would decrease, approaching an asymptote that captures truly stable factors of personality across the life course (Cole, 2015; see also Anusic & Schimmack, 2016).

So how can we interpret the age trajectories for the contribution of the time-invariant trait to stability reported by J. Wagner et al. (2019)? They result from the limited length of the time interval covered by the study in combination with the specification of the statistical model. Due to the limited time interval, the “trait” component will partially capture aspects of personality that are not trait-like. Age-related changes in stability are thus split between the two model components that are allowed to change with age (i.e., autoregressive effects and the time-invariant trait), resulting in age trajectories that can hardly be interpreted in a meaningful way. In particular, the age trajectory of the contribution of the “trait” will chiefly depend on the length of the study; and the age trajectory of the autoregressive component will reflect whatever is left over after accounting for the “trait”.

Apart from age trends in the components of the TSO model that we find hard to interpret, J. Wagner et al. (2019) also investigated the age development of rank-order stability. Here, their study provided multiple improvements over previous research, as it did not rely on age groups (Lucas & Donnellan, 2011; Wortman et al., 2012) or on prespecified developmental patterns (Ardelt, 2000; Milojevic & Sibley, 2014; Specht et al., 2011). The resulting fine-grained age trajectories provided a further more nuanced view on personality stability development with challenging results: In two large age-representative samples, overall, “test–retest correlations only partially support the life span trend of the inverted U-shape but rather suggested some increase in stability in young adulthood but less variation later in life” (J. Wagner et al., 2019, p. 676). The results also indicated that different traits might follow different developmental trends (for similar suggestions, see Milojevic & Sibley, 2014; Specht et al., 2011; see also Costa et al., 2019) with some traits resembling an inverted U-shape pattern, while others showed no age-related changes in stability at all.

Despite the methodological improvements implemented by J. Wagner et al. (2019), they made some decisions that raise methodological concerns: First, item loadings were not constrained to be equal across age, and, thus, the analyses did not ensure that personality was measured invariantly across age. Second, their modeling approach assumed that stability differed only cross-sectionally between individuals of different ages but not within individuals as they got older, which partially precluded change in personality stability—the very topic of the investigation. Third, J. Wagner et al. included only individuals with complete data on all three measurement waves over the 8-year period of the surveys. This may have led to the selective exclusion of participants who did not participate in later waves due to, for example, health concerns or death. The resulting sample of complete cases may thus be biased toward higher

health and longevity, and this, in turn, could bias estimates of personality stability upwards, especially in older ages when changes in health go along with changes in personality (Kornadt et al., 2018; Mueller et al., 2018). Thus, excluding participants who did not participate in all waves may have distorted the actual trajectories among older participants.¹

Hence, the question of how to interpret the diverging results reported by J. Wagner et al. (2019) has yet to be answered: Were the results caused by the methodological issues outlined above, or do they actually reflect a more accurate picture of personality stability development thanks to the methodological advancement of more finely grained age trajectories? Despite the large body of studies on personality development, the actual age trajectory of the rank-order stability of the Big Five across the life span is still an open question.

The Present Studies

To address this issue, we relied on the same two age-diverse and nationally representative large panel studies that were analyzed by J. Wagner et al. (2019): the Household, Income and Labour Dynamics in Australia (HILDA; Study 1) Survey and the German Socio-Economic Panel (SOEP; Study 2). In contrast to J. Wagner et al., we additionally included both studies' fourth waves, which had become available in the meantime. Our analyses thus covered 12 years of personality development with 4-year retest intervals. Statistically, we used a two-step approach: In a first set of descriptive analyses, we used local structural equation modeling (LSEM; Hildebrandt et al., 2016) to obtain latent-level estimations of fine-grained age effects on personality stability which are not distorted by measurement error. These analyses were similar to the ones reported by J. Wagner et al. (2019) but avoided the limitations noted above. The aim

¹ Both the estimation of age-invariant loadings and a more elaborate handling of missing data were not possible for J. Wagner et al. (2019) due to a lack of availability in the then-current Version 1.14-10 of the respective R package *sirt*.

was to portray the age trend in rank-order stability as accurately as possible in an exploratory manner. In a second set of analyses, we then formally tested the observed age trajectories by statistically modeling the descriptive age curves. This allowed us to stringently test different theoretically possible age trajectories against each other.

In total, we analyzed data from more than 35,000 individuals. Relying on samples from two countries allowed us to assess the robustness of our findings. In both studies, we extended previous research by examining whether the age trends reported for stability development could be generalized to time spans of 8 and 12 years. Indeed, previous studies have usually covered rather short retest intervals (for 2-year stabilities, see Milojev & Sibley, 2014; for 4-year stabilities, see Lucas & Donnellan, 2011; Specht et al., 2011; J. Wagner et al., 2019; Wortman et al., 2012; for about 7 years as the average interval in meta-analyses, see Ferguson, 2010; Roberts & DelVecchio, 2000). Furthermore, the Big Five traits (i.e., Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness) were considered both in isolation (to take possible differences between traits into account; Milojev & Sibley, 2014; Specht et al., 2011; J. Wagner et al., 2019) and jointly to allow for the most precise characterization of the overall pattern.

Additionally, to explicitly account for potential disparities between our and J. Wagner et al.'s (2019) results, we investigated how different changes to the modeling strategy affected the resulting stability trajectories. In particular, we investigated the hypothesis that relying on complete cases would result in an unrepresentatively healthy sample that would bias the age trajectories of personality. We directly examined the role of health on personality stability development, which is further of special interest as health-related processes have been suggested to be a key factor in the development of personality stability in older age (Ardelt, 2000; Lucas &

Donnellan, 2011; Wortman et al., 2012), with little direct empirical evidence so far. We report these additional analyses after Study 2.

General Method

The HILDA and SOEP studies share essential design-related features (e.g., both are panel studies; both contain Big Five measures in 2005, 2009, 2013, and 2017), so we kept the analysis strategies for the two data sets as parallel as possible. We first describe our general methodological approach and later fill in the study-specific details. Our research did not require ethical approval because we analyzed existing and fully anonymized data; informed consent was obtained from participants by the respective institutions.

Inclusion Criteria

Both studies are ongoing longitudinal panels with assessment waves that are conducted annually. The sampling units are households, and all the members above the respective age of eligibility provided personality data. However, not all individuals' data were available in all waves. First, new households were included from time to time. Second, new individuals moved into already included households, or existing members reached the age of eligibility. Third, individuals might not have participated in single waves for various reasons; or they may have dropped out of the study completely (e.g., because they died). Thus, including only the participants who took part in all four personality assessments would have led to a substantial decrease in the sample size. More importantly, as reasoned above, excluding participants with incomplete data over the 12-year period of the survey may have resulted in a sample that was not representative with respect to health and longevity, which, in turn, may have resulted in the overestimation of rank-order stability. This issue would likely have particularly affected

estimates in older age when health tends to change along with personality (Kornadt et al., 2018; Mueller et al., 2018).

Thus, instead of restricting ourselves to complete cases, we set the inclusion criteria for our analyses to be as permissive as possible. Inclusion was determined on a trait basis (i.e., sample sizes could vary between traits), and respondents were included if they had answered at least one item characterizing the respective Big Five trait in at least two waves. We used full information maximum likelihood (FIML) estimation in all further analyses, which allowed us to include the participants with incomplete data (e.g., Newman, 2014). This approach recovers the correct estimates when values are missing at random (MAR) conditional on variables included in the model (Newman, 2014). We believe that this assumption is likely to be violated in older age when changes in health are closely linked to both changes in personality (e.g., Mueller et al., 2018) and dropout. The types of panel data available to us were too “coarse” to include a set of variables that could fulfill the assumption of MAR—we lacked data between the annual assessments, and health can change rapidly within a year; so, our data were most likely missing not at random (MNAR) by design. Rather than considering missingness as merely cumbersome during parameter estimation, in additional analyses, we explicitly explored how it may have affected our conclusions and potentially explained contradictory findings in the literature. It should be noted that listwise deletion (i.e., the reliance on complete cases) rests on the even stronger assumption that data are missing completely at random (MCAR; Newman, 2014). To examine how this (implausible) assumption could have biased our conclusions, and to allow for comparisons with previous studies, we additionally report results on the basis of only complete cases.

Estimation of Rank-Order Stabilities

Rank-order stability is computed as the correlation of the same personality factor across two measurement points. This autocorrelation is usually estimated on a latent level within a structural equation modeling (SEM) framework (e.g., Lucas & Donnellan, 2011; Specht et al., 2011; J. Wagner et al., 2019; Wortman et al., 2012). As latent modeling takes measurement error into account, the resulting estimates are more precise. The question of how personality stability develops over the life span can be addressed by investigating how latent retest correlations are moderated by age.

Approaches to Moderation in Structural Equation Modeling

A variety of methods for implementing moderation in SEM exist (Kline, 2016). For example, in latent moderated structural equations (LMS; developed by Klein and Moosbrugger, 2000; see also Maslowsky et al., 2015), the latent interactions of continuous variables can be modeled (e.g., Specht et al., 2011). Using this technique, the moderation pattern (e.g., linear, quadratic, or cubic) needs to be specified a priori. Thus, it does not allow researchers to describe a moderation pattern without specifying a functional form (see Costa et al., 2019).

An alternative approach is multiple-group structural equation modeling (MGSEM), which does not require such a prespecified pattern. MGSEM simultaneously estimates one SEM for every value that the moderator variable takes on. This approach is typically used for categorical moderators, and it entails methodological difficulties if the moderator is a continuous variable (e.g., age; see Brandt et al., 2020; Lucas & Donnellan, 2011; Wortman et al., 2012). Researchers routinely transform their continuous moderator by arbitrarily lumping values together (e.g., 4-year or 5-year age groups; Lucas & Donnellan, 2011; Wortman et al., 2012). Such a categorization is problematic for various reasons: (a) results depend on which and how

many groups are defined; (b) the categorization leads to a loss of information (MacCallum et al., 2002); (c) dividing the sample into distinct groups can lead to very small sample sizes in some of these subgroups, which, in turn, can result in unstable estimates (Wolf et al., 2013).

A recently developed extension of MGSEM can be applied to overcome these issues. In local structural equation modeling (LSEM; Hildebrandt et al., 2009, 2016; Olaru et al., 2019), parameter estimates for a particular value of a moderator (e.g., the stability estimate for a particular year of age) are also informed by respondents whose values on the moderator do not exactly match (e.g., by respondents who are slightly younger and older). The information is weighted according to the distance of the respondent's moderator value from the focal value (e.g., the estimate at age 30 will be strongly influenced by respondents who are 31 years of age but barely influenced by respondents who are 61).

The weighting procedure leads to more robust and precise estimates as well as higher power than MGSEM (Olaru et al., 2019). Further, LSEM outperforms its antecedent by incorporating a continuous moderator in its actual continuous form. In other words, instead of arbitrarily grouping values, each value of the moderator (e.g., each year of age) is modeled separately as a so-called focal point, resulting in smooth trajectories. An additional advantage is that LSEM is a nonparametric approach for the moderation of latent models: The shape of the moderation effect does not have to be specified explicitly as it must be, for example, in LMS. Because of these distinct advantages, we chose LSEM to estimate the moderating effect of age on the latent rank-order stability of personality (for further applications of LSEM, see, e.g., Hartung et al., 2018; J. Wagner et al., 2019; Zheng et al., 2019).

Implementation of Local Structural Equation Modeling in the Present Studies

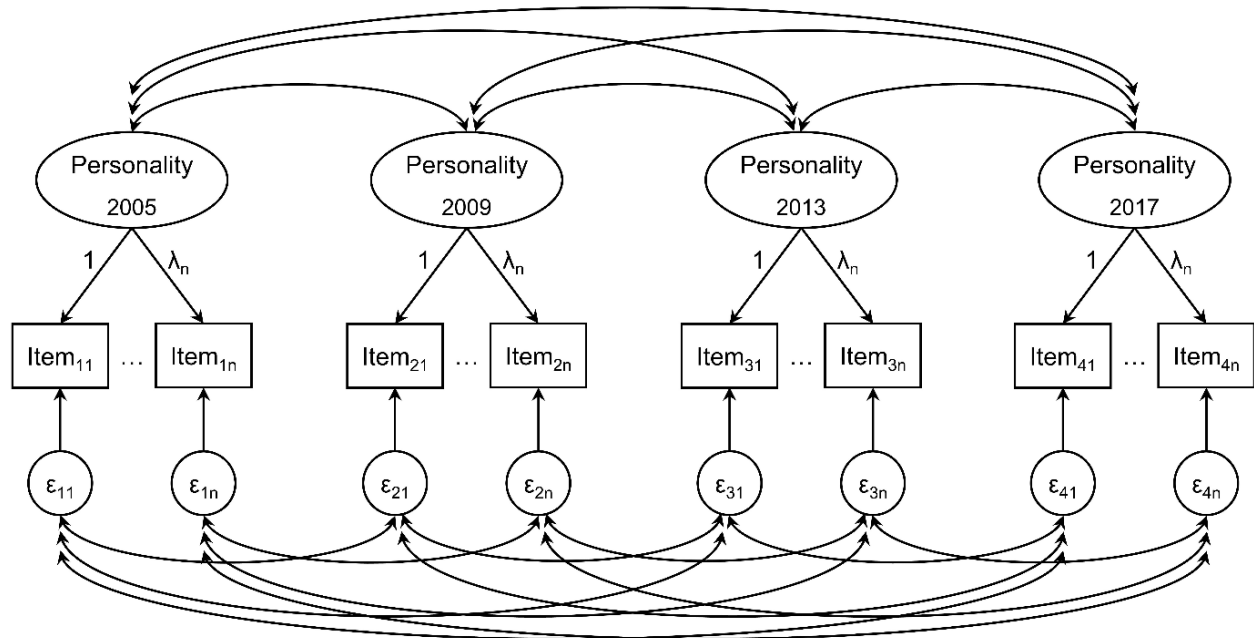
Statistical Model. Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness were analyzed separately. As illustrated in Figure 1, we modeled each of the four waves of measurement as a separate latent factor. The number of manifest indicators per factor ranged from three to six items, depending on study and trait. All four latent factors were allowed to covary with each other. Thus, three 4-year stabilities (2005–2009, 2009–2013, and 2013–2017), two 8-year stabilities (2005–2013 and 2009–2017), and one 12-year stability (2005–2017) were estimated simultaneously. We allowed the residual variances of each item to covary with the same item across time (see Little, 2013).

Focal Points and Weighting Parameters. To estimate the moderating effect of age on rank-order stability, we used LSEM to estimate the described statistical model for each focal point (i.e., each year of age). In our analyses, the focal points were defined as the participants' year of age at the first wave of measurement in 2005. We restricted the range of focal points such that each focal point had at least $n = 10$ observations, which was necessary to avoid nonpositive definite covariance matrices and Heywood cases (for a restriction of focal points, see also Hartung et al., 2018; for more detailed information on how we restricted the range of the focal points, see the supplemental material).

In LSEM, the whole sample is used to estimate a single SEM. That is, observations outside the range of focal points still contributed to the estimation. However, the sample is iteratively weighted for these estimations: Participants with a matching focal point get the largest weight (i.e., 1), and the more participants deviate from that focal point, the less weight they receive. For the weighting procedure, a Gaussian kernel function is most frequently used (see Gnambs & Schroeders, 2020; Hartung et al., 2018, 2020; Hildebrandt et al., 2016; Olaru et al., 2019; J. Wagner et al., 2019). As the term *Gaussian* suggests, the resulting weights follow a normal distribution around each focal point, and the standard deviation of this normal distribution (the so-called bandwidth) determines the degree of influence that the data points surrounding the focal point have. The size of the bandwidth can be set by the bandwidth factor h . Following both recommendations (Hildebrandt et al., 2016; Olaru et al., 2019) and common practice (e.g., J. Wagner et al., 2019), we chose a Gaussian kernel function with a bandwidth factor of $h = 2$.

Figure 1

The Age-Moderated Statistical Model for Estimating Rank-Order Stabilities



Note. The first subscript indicates the measurement point; the second subscript indicates the item. To ensure measurement invariance, factor loadings were set to be equal across measurement points and ages.

Measurement Invariance. To be able to draw valid conclusions, it is crucial to ensure that the same personality construct was measured at every assessment point for all ages (i.e., measurement invariance on two dimensions: wave and age). Our baseline model assumes configural invariance (i.e., the same pattern of factor loadings) across waves and ages. Next, to test for metric measurement invariance, we estimated LSEMs with added constraints on the factor loadings and compared the fit statistics: First, we examined whether the same construct was measured *across the four waves* by constraining the factor loadings of the same item to be equal across the four measurement points. Then, we additionally established metric measurement

invariance *across age* by restricting the factor loading of an item to be equal across all ages (for the introduction of invariance constraints across moderator values in LSEM, see Olaru et al., 2020). The model with metric invariance across both time points and ages was used to estimate the rank-order stabilities.

Model Fit. We used a joint estimation approach for the LSEM analyses to obtain common fit indices across focal points (comparative fit index [CFI], root mean square error of approximation [RMSEA], standardized root mean square residual [SRMR]). Conventional cutoffs (Hu & Bentler, 1999) are frequently used to evaluate these indices ($CFI \geq .95$, $RMSEA \leq .06$, $SRMR \leq .08$). To test for metric measurement invariance, differences in the fit indices were evaluated stepwise. Rough guidelines (Chen, 2007; Cheung & Rensvold, 2002; Rutkowski & Svetina, 2014) for the evaluation of invariance were suggested ($\Delta CFI \leq .01$ to $.02$, $\Delta RMSEA \leq .015$ to $.030$), but again no clear consensus exists (Putnick & Bornstein, 2016).

Processing of Stability Estimates. As depicted in Figure 1, we obtained three 4-year stabilities, two 8-year stabilities, and one 12-year stability per personality factor. For the purpose of further analyses and display, these estimates were subsequently processed (for more details about each step, see the supplemental material). First, we standardized the rank-order estimates to obtain latent correlations. Second, because the statistical moderator (i.e., the focal point) was the age at the first measurement point, we recoded these values by adding the necessary number of years to represent respondents' actual year of age for the stability estimates 2009 to 2013, 2009 to 2017, and 2013 to 2017. Third, we set boundaries for age in the stability trajectories so that every age cell within the range included at least $n = 10$ participants in the relevant two measurement waves. Fourth, the rank-order stabilities of the same time interval (i.e., three 4-year stabilities and two 8-year stabilities) were averaged. Finally, to characterize the trajectory across

all traits, we also averaged the rank-order correlations across the Big Five. Standard errors were obtained by applying a bootstrapping procedure with 1,000 replicates each.

Test of Age Trajectories

Comparison of Regression Models

We used a regression-based approach to formally test the different hypotheses on how the rank-order stability of the Big Five develops with age. Based on our aforementioned analyses, a total of 18 age trajectories—resulting from three time intervals (4 years, 8 years, and 12 years) \times six domains (the Big Five plus the average across the five domains)—emerged. For the regression analyses, these 18 trajectories served as separate data sets, and we predicted the latent rank-order stability from age.

Different assumptions about the development of stability can be mathematically translated into several regression models: First, the most parsimonious model for an increase in stability throughout the life span (i.e., the cumulative continuity principle; Roberts et al., 2008) is a linear model; $stability_i = b_0 + b_1age_i + e_i$ (with $b_1 > 0$). Second, if personality becomes more stable until a certain age and then reaches a plateau (e.g., Ferguson, 2010; Roberts & DelVecchio, 2000), an exponential function can describe the trajectory; $stability_i = b_0 + b_1 \exp(b_2age_i) + e_i$ (with $b_1 < 0$ and $b_2 < 0$; b_0 represents the height of the plateau). Third, if stability increases at younger ages and decreases at older ages (e.g., Lucas & Donnellan, 2011; Wortman et al., 2012), a quadratic function can describe the trajectory; $stability_i = b_0 + b_1age_i + b_2age_i^2 + e_i$ (with $b_2 < 0$). Fourth, to consider the possibility that the stability of personality does not change systematically with age (see, e.g., J. Wagner et al., 2019), we also included an intercept-only model; $stability_i = b_0 + e_i$ (b_0 indicates the mean of rank-order stability across age).

These four models are not nested in a hierarchical order, and for nonlinear models (e.g., the exponential model), R^2 is not an adequate measure of model fit (Spiess & Neumeier, 2010). Instead, we used the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for model comparison, and we identified the model with the lowest value as the best-fitting model. As the AIC and BIC consistently indicated the same best-fitting model, we only report the BIC (for the AIC, see Tables S11 and S19 in the supplemental material).

Two-Lines Test

A quadratic regression is suitable for describing symmetrical inverted U-shapes. However, the term “inverted U-shape” is often used in a broader sense to refer to some increase followed by some decrease, regardless of symmetry. Capturing asymmetrical inverted U-shapes with quadratic regression functions may lead to misdescriptions of such patterns (Simonsohn, 2018a). We thus additionally conducted the two-lines test suggested by Simonsohn (2018a) to check for asymmetrical inverted U-shapes of rank-order stability development. The two-lines test approximates the sign flip implied by (inverted) U-shapes by fitting two straight regression lines: one representing the increase and the other representing the decrease. In the case of an inverted U-shaped function, the slope of the first line is expected to be significantly positive, and the slope of the second line is expected to be significantly negative. The lines are separated by the so-called break point, which is algorithmically set to balance statistical power between the two lines in order to increase the detectability of a significant slope in the statistically weaker one (allegorically, it has been referred to as the Robin Hood algorithm). On the basis of simulations, Simonsohn concluded that the two-lines test achieves a lower false-positive rate than other existing methods testing for (inverted) U-shaped patterns (for an application of the two-lines test, see, e.g., Brown et al., 2021).

Software, Analysis Scripts, and Data

All analyses were run in R (Version 3.5.0) using the RStudio environment (Version 1.2.1335). The R package *sirt* (Robitzsch, 2019) was used to estimate the LSEMs. For bootstrapping, we used *boot* (Canty & Ripley, 2017). The package *nls2* (Grothendieck, 2013) was used to fit the exponential models. In order to conduct the two-lines test, we adapted the R code Simonsohn (2018b) provided online. All analysis scripts (including more extensive results and the list of all R packages we used) are publicly available on the Open Science Framework (OSF) and can be accessed at <https://osf.io/rzfqm/>. As explained below, HILDA and SOEP data can be requested directly from the responsible institutions. All our analysis scripts are written to ensure that our findings can be directly reproduced from the original HILDA and SOEP data sets (with no need for any additional steps to prepare the data). This article reports the results of latent variable analyses. To gauge the robustness of our results, we ran comparable analyses on a manifest level, which largely confirmed the latent results. Details about the manifest analyses can be found in the supplemental material.

Study 1

Method

Design and Participants

The HILDA Survey is an ongoing Australian panel study that is collecting a wide range of information about economic, social, and individual issues. Data collection started in 2001, and new waves are conducted annually. In the first wave, a large national probability sample of Australian households was interviewed. In these selected households, every member aged 15 years and older was asked to fill out a self-completion questionnaire every year. In the 11th measurement wave (2011), the initial sample was replenished by adding new households.

Watson and Wooden (2012) provided a general introduction to HILDA, and Summerfield et al. (2018) provided a more technical description. Personality was assessed in 2005, 2009, 2013, and 2017.

HILDA data are available to researchers worldwide.² Because of the broad range of topics that are covered and the rigorous methodological standards the survey adheres to, the survey data have been frequently used for research in a variety of scientific disciplines, including psychology. Not surprisingly, the personality data have been used in previous studies to examine the development of the rank-order stability of the Big Five (J. Wagner et al., 2019; Wortman et al., 2012). However, these studies used only the first two or three waves of personality data in HILDA and differ from our study in the statistical analyses that were performed.

To estimate the rank-order stabilities of the Big Five, a total of $N = 15,465$ participants were included. To describe the age of the HILDA sample with respect to the longitudinal design, participants' age was determined by averaging across the waves in which an individual participated. According to these estimates, participants had a mean age of 45.45 years ($SD = 18.26$). The lowest age of a participant was 15, and the highest was 101. The proportion of female participants was 53%. Conducting the analyses separately for the Big Five resulted in slightly different sample sizes for Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness (see Table 1).

Measures

In 2005, 2009, 2013, and 2017, the same list of 36 adjectives was used as a measure of the Big Five (for the list, see Table S2 in the supplemental material). Each time, participants were asked “How well do the following words describe you?,” and their responses on a 7-point

² In our analyses, we used Release 17.0 of HILDA (Department of Social Services & Melbourne Institute of Applied Economic and Social Research, 2018).

scale could range from 1 (*Does not describe me at all*) to 7 (*Describes me very well*). Most items originated from the trait-descriptive adjectives presented by Saucier (1994), which, in turn, are a selection of the words listed by Goldberg (1992). For a more detailed report of the development and evaluation of the Big Five measure in HILDA, see Losoncz (2009).

However, unsatisfactory psychometric properties consistently arose when using all 36 adjectives in the sample (see Tables S5 to S9 in the supplemental material). We therefore excluded some items, resulting in six items for Neuroticism; five for Conscientiousness, Agreeableness, and Openness; and four for Extraversion. Previous reports also recommended or focused on a subset of the adjectives in HILDA (Losoncz, 2009; Summerfield et al., 2018; Wortman et al., 2012). In the supplemental material, we describe the rationale behind our selection of items and detailed results of the psychometric analyses. The subset of items that we used for our analyses generally showed satisfactory internal consistencies across the four waves for Neuroticism ($\alpha = .77$; $\omega = .78$), Extraversion ($\alpha = .65$ to $.69$; $\omega = .66$ to $.70$), Conscientiousness ($\alpha = .77$ to $.78$; $\omega = .78$ to $.79$), Agreeableness ($\alpha = .73$ to $.74$; $\omega = .74$), and Openness ($\alpha = .71$ to $.73$; $\omega = .71$ to $.74$).³ As the internal consistencies were far from perfect, a latent modeling approach in which we controlled for measurement error was indicated.

Results

Measurement Invariance

In a first step, we tested for the measurement invariance of our Big Five measures across age and measurement waves (see Table 1). As clearly indicated by all fit indices, the LSEM

³ Cronbach's alpha (α) requires an essentially tau-equivalent measurement model (i.e., equal factor loadings for the items in an SEM). The assumption is seldom tenable, and violations lead to a biased estimation of internal consistency (Graham, 2006). We therefore also report McDonald's omega (ω ; e.g., McNeish, 2018), which does not assume essential tau-equivalence. To set the eligibility criteria for participants in the psychometric analyses to be as inclusive as possible, we included all participants who answered at least one item from the respective Big Five factor within one wave. For sample characteristics, see Tables S3 and S4 in the supplemental material.

model that assumed the same pattern of factor loadings across time and age (configural invariance) fit the data well. Restricting the factor loadings of the same item to equality across the four measurement waves went along with only marginal differences in fit. Additionally, setting the factor loadings of the items to be identical across ages also resulted in no substantial loss in fit. Hence, metric invariance across time and age was established. For each personality dimension, the final model fit the data well.

Rank-Order Stability of Personality

As measurement invariance is vital for drawing valid conclusions about age differences in personality stability, we used the model with metric invariance across assessment points and participants' ages to estimate the latent rank-order stabilities of Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness. The corresponding latent rank-order stabilities of the Big Five are depicted in Figure 2.

Table 1

Fit Indices for Testing for Measurement Invariance (MI) Across Time and Age for the Big Five in the Household, Income and Labour Dynamics in Australia Survey

Model	<i>n</i>	CFI	RMSEA	SRMR
Neuroticism	15,433			
Configural MI		.987	.024	.036
Metric MI across time		.987	.023	.037
Metric MI across time and age		.985	.025	.043
Extraversion	15,438			
Configural MI		.989	.034	.030
Metric MI across time		.989	.033	.031
Metric MI across time and age		.988	.033	.035
Conscientiousness	15,422			
Configural MI		.973	.040	.043
Metric MI across time		.973	.038	.043
Metric MI across time and age		.973	.038	.046
Agreeableness	15,455			
Configural MI		.987	.028	.043
Metric MI across time		.987	.027	.044
Metric MI across time and age		.986	.028	.048
Openness	15,411			
Configural MI		.989	.028	.031
Metric MI across time		.989	.027	.032
Metric MI across time and age		.988	.027	.037

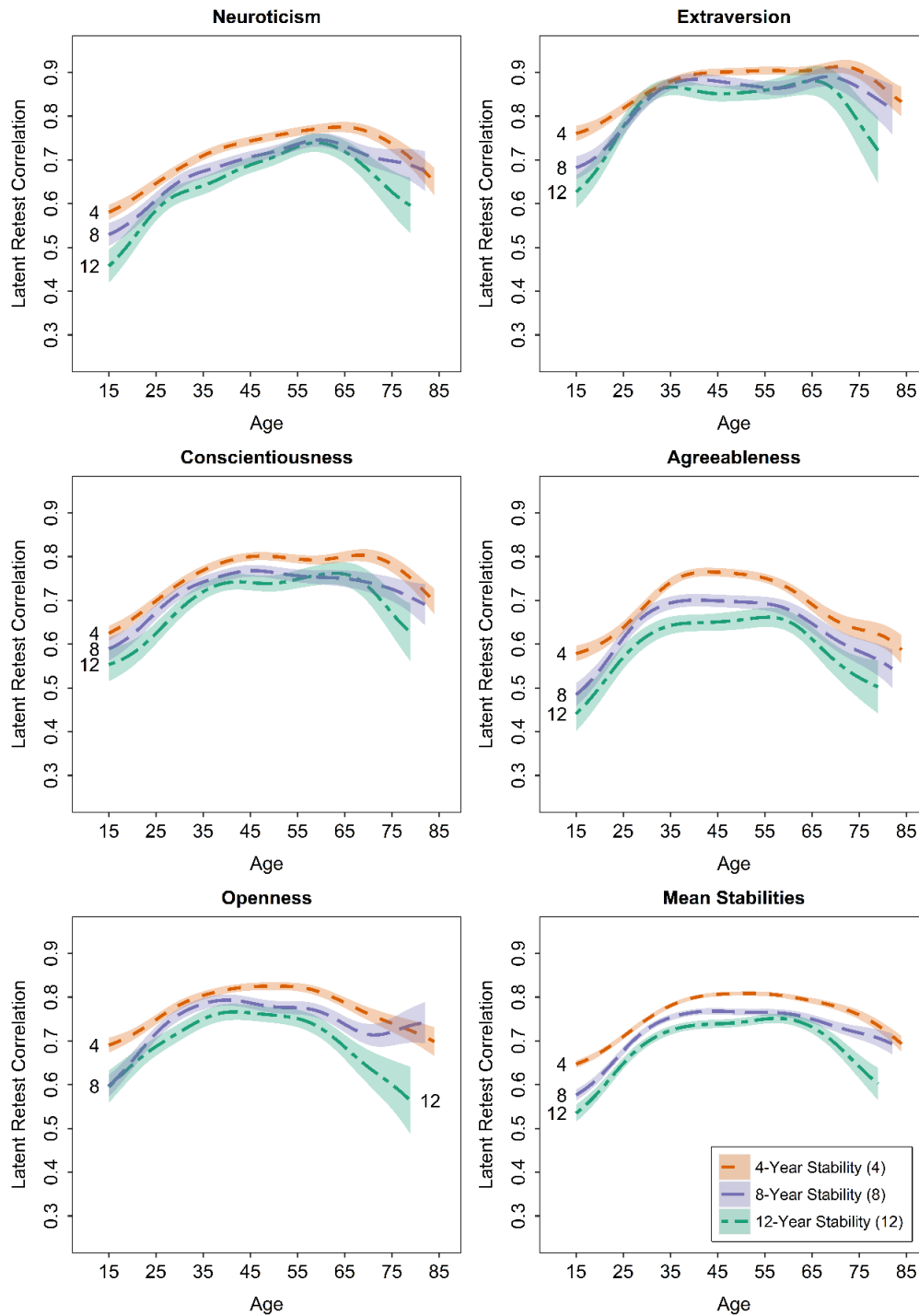
Note. Local structural equation modeling was used for the analysis. Configural MI assumes the same pattern of factor loadings across time and age. Metric MI assumes the same factor loadings across time (and age). CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

Our results reveal several findings: First, across all traits and ages, the stability estimates characterized personality as neither completely unstable ($r = 0$) nor perfectly stable ($r = 1$). Across the life course, the lowest rank-order correlations ranged from .44 to .76. The peak of the correlations ranged from .66 to .91. Second, the longer the time interval between two measurement points, the lower the rank-order stability. As can be seen in Figure 2, this was the case for all Big Five dimensions (Table S10 in the supplemental material presents descriptive statistics for the rank-order stabilities). The stabilities also varied across the Big Five: Extraversion had higher retest correlations than Openness and Conscientiousness, which, in turn, had higher retest correlations than Neuroticism and Agreeableness (see Figure 2 and Table S10 in the supplemental material).

Third and most importantly, the rank-order stabilities of the Big Five changed over the life span. The absolute difference between the lowest and highest correlations ranged from .14 to .28 (the standard deviation of the correlations within a trajectory ranged from .04 to .07), indicating non-negligible variation. Changes in rank-order stability were comparable in magnitude across trajectories. As can be seen in Figure 2, particularly when considering the stabilities averaged across the Big Five in the right bottom panel, stability began at a relatively low level at around 15 years, continuously increased with age, and then finally decreased after roughly 50 years of age. This inverted U-shape emerged across all Big Five and retest intervals. Intriguingly, the results for Extraversion, Conscientiousness, and Agreeableness suggested that rank-order stability plateaued in the middle-aged groups.

Figure 2

The Rank-Order Stabilities of the Big Five for the 4-, 8-, and 12-Year Intervals Across Age in the Household, Income and Labour Dynamics in Australia Survey



Note. Shaded areas represent the bootstrapped standard error.

Test of Age Trajectories

We tested the developmental shape of personality stability more formally in a regression framework. Per age trajectory (as they are depicted in Figure 2), we estimated an intercept-only, a linear, an exponential, and a quadratic model. These four age functions were compared with each other using the BIC. A lower BIC value indicates a better fit. Table 2 reports the results of the model comparisons (for the estimated regression parameters, see Table S12 in the supplemental material). The association between age and rank-order stability was best described by a quadratic function for all personality factors and retest intervals.

The respective quadratic regression functions for the stabilities averaged across the five personality dimensions are depicted in the left column of Figure 3. Figures S1 to S5 in the supplemental material display the quadratic functions for the individual Big Five dimensions. For each trajectory, the quadratic regression function followed an inverted U-shaped pattern. Thus, regression analyses confirmed the visual impression that the development of personality rank-order stability is characterized by an increase at younger ages and a decrease at older ages.

The two-lines test (Simonsohn, 2018a) confirmed this age pattern: All 18 trajectories could be described by a significant positive slope followed by a significant negative slope (for the model parameters of the two-lines test, see Table S13 in the supplemental material). The estimated two lines for the mean stabilities in HILDA are depicted in the right column of Figure 3 (for the individual Big Five dimensions, see Figures S1 to S5 in the supplemental material).

Table 2

Comparisons of Regression Models to Describe the Development of Rank-Order Stability With Age in the Household, Income and Labour Dynamics in Australia Survey

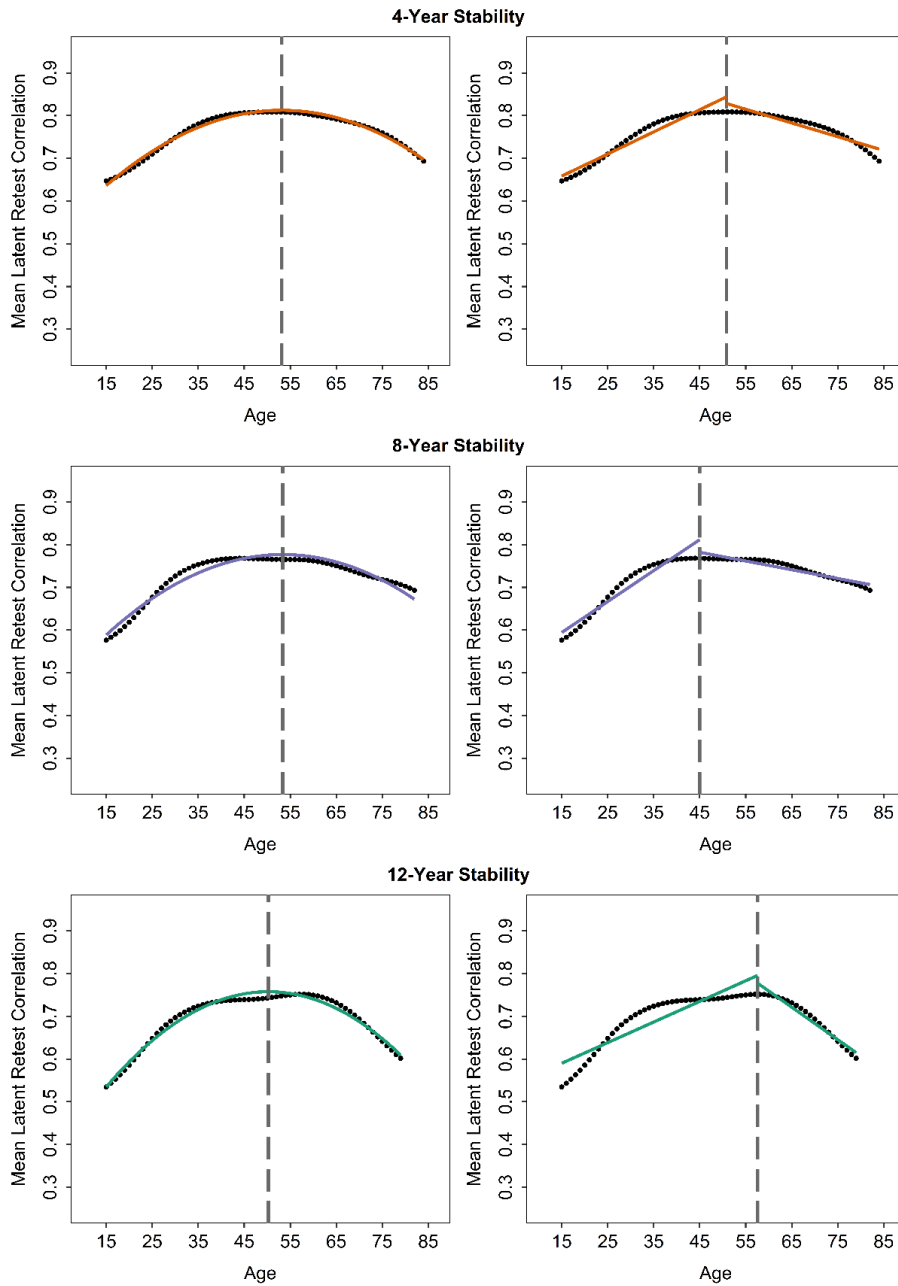
Model	BIC		
	4-year stability	8-year stability	12-year stability
Neuroticism			
Intercept	-195.756	-183.792	-146.829
Linear	-224.970	-234.266	-176.426
Exponential	-288.707	-335.889	-240.505
Quadratic	-430.641	-480.979	-349.190
Extraversion			
Intercept	-229.651	-184.250	-155.235
Linear	-263.334	-214.859	-165.872
Exponential	-350.625	-320.614	-235.145
Quadratic	-443.346	-328.031	-278.275
Conscientiousness			
Intercept	-207.577	-210.063	-166.012
Linear	-229.904	-228.258	-186.968
Exponential	-307.525	-322.047	-245.928
Quadratic	-431.910	-418.529	-356.473
Agreeableness			
Intercept	-180.546	-174.993	-165.412
Linear	-176.374	-170.875	-162.568
Exponential	-198.085	-208.643	-201.283
Quadratic	-353.737	-359.756	-390.327
Openness			
Intercept	-231.287	-212.624	-170.973
Linear	-227.044	-213.796	-169.725
Exponential	-251.914	-286.579	-205.291
Quadratic	-499.621	-298.585	-497.708
Mean stabilities			
Intercept	-218.182	-201.160	-170.146
Linear	-224.533	-214.053	-174.904
Exponential	-282.688	-306.251	-228.615
Quadratic	-526.240	-388.484	-408.165

Note. The relatively best-fitting model is displayed in bold. BIC = Bayesian information

criterion.

Figure 3

Quadratic Regression Model (Left Column) and Two-Lines Test (Right Column) for the Mean Rank-Order Stabilities in the Household, Income and Labour Dynamics in Australia Survey



Note. Each dot represents the rank-order stability estimate for an individual year of age. The dashed vertical line indicates the highest point in the quadratic function (left column) or the break point in the two-lines test (right column).

At what age does personality stability peak, turning from an increase into a decrease?

Both the quadratic regression and the two-lines test provided answers. With a quadratic regression function, the age at maximal stability can be determined analytically by the first derivative. With the two-lines test, a specific year of age is algorithmically set as a break point that separates the two lines. However, both values can only be interpreted as approximations as they are determined on the basis of empirical data, and differences between the two methods may arise at any point (see Simonsohn, 2018a).

According to the quadratic regression, averaged across the Big Five and retest intervals, personality stability was at its highest at 52.67 years of age ($SD = 3.93$, Range: 45.19 to 58.38). This age varied more across the five factors ($SD = 3.68$) than across the retest intervals ($SD = 1.93$). Averaged across the retest intervals, Openness ($M = 48.69$, $SD = 3.19$) and Agreeableness ($M = 48.75$, $SD = 0.49$) peaked somewhat earlier than Conscientiousness ($M = 54.31$, $SD = 1.42$), Extraversion ($M = 55.17$, $SD = 3.10$), and Neuroticism ($M = 56.43$, $SD = 2.13$).

The results of the two-lines test were largely similar. Across the Big Five and the retest intervals, the break point for stability was set at 55.64 years ($SD = 10.77$, Range: 39.86 to 69.93). Again, there was more variability between the personality dimensions ($SD = 9.78$) than between the retest intervals ($SD = 4.10$). Sorted by the mean break point across the retest intervals, the following order of the Big Five resulted: Openness ($M = 44.19$, $SD = 5.41$), Agreeableness ($M = 47.02$, $SD = 7.57$), Conscientiousness ($M = 58.62$, $SD = 11.80$), Neuroticism ($M = 60.86$, $SD = 2.17$). However, for Extraversion, the break point was considerably later than the peak estimated by the quadratic regression ($M = 67.52$, $SD = 2.69$), possibly because the underlying trajectory of stability was characterized by a plateau in middle age (see Figure S2 in the supplemental material), increasing the uncertainty surrounding the peak superimposed by the two methods.

Summary of Study 1

Our results confirmed several well-known findings. That is, the personality dimensions showed a considerable amount of stability across several years (e.g., Anusic & Schimmack, 2016). Further, the stability estimates decreased as the duration of the interval between measurement points increased (e.g., Conley, 1984). Supporting the cumulative continuity principle, the rank-order stability of the Big Five increased with age until middle adulthood (Roberts & Nickel, 2017). But in contrast to previous studies, we did not find that rank-order stability reached a plateau after middle adulthood (e.g., Roberts & DelVecchio, 2000). Rather, for all Big Five dimensions and across all time intervals between measurements, stability consistently decreased after roughly 50 years of age, supporting an inverted U-shaped development of stability across the life span (e.g., Lucas & Donnellan, 2011; Specht et al., 2011). To examine the robustness of these results, we performed parallel analyses on data from the SOEP in Study 2.

Study 2

Method

Design and Participants

The SOEP is an ongoing longitudinal annual survey of German households and their members. A broad range of objective and subjective indicators of well-being—as well as background information including psychological variables—are core topics of the survey. The first wave in 1984 included a representative selection of German households. Across the years, several refreshment samples of new households were added. All household members who had reached the age of eligibility were asked to complete individual questionnaires. The minimum age of participants changed over time: Whereas in 2005, all participants who were 17 years of

age were included in the data sets (and a single participant was even 16 years old), in later waves, all participants were at least 18. The Big Five were assessed in 2005, 2009, 2013, and 2017. For general information regarding the SOEP, see Goebel et al. (2019) and G. G. Wagner et al. (2007).

The SOEP data are available for scientific purposes.⁴ The SOEP data have been used across a broad range of research fields (Goebel et al., 2019), which is not surprising considering the wide array of information collected with strict methodological standards. Previous studies that examined the age-related development of the rank-order stability of the Big Five in the SOEP (Lucas & Donnellan, 2011; Specht et al., 2011; J. Wagner et al., 2019) only made use of the first two or three waves of personality data and differed from our study in the analyses they performed.

A total of $N = 21,777$ participants were included in our analyses. Averaging participants' age across the waves in which they took part, participants were on average 51.47 years old ($SD = 17.47$). The youngest participant was 16, and the oldest was 103 years old. The proportion of female individuals was 53%. Conducting the analyses separately for the Big Five resulted in slightly different sample sizes for Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness (see Table 3).

Measures

The Big Five were measured in 2005, 2009, 2013, and 2017. In the corresponding waves, participants were asked to rate several items, each of which completed the sentence "I am someone who ..." (e.g., "I am someone who works thoroughly"). On a 7-point scale, participants could indicate their agreement to the statements, ranging from 1 (*Does not apply at all*) to 7

⁴ In our analyses, we used Version 34 of the SOEP (Liebig et al., 2019).

(*Applies perfectly*). Each personality dimension was assessed with the same set of three items across waves. However, in 2009, a fourth item for assessing Openness was added. To keep the meaning of the factor comparable across time, we did not include this additional item in any of our analyses.

Items were based on the Big Five Inventory (Benet-Martínez & John, 1998; John & Srivastava, 1999) and were translated into German. For more information about the development and evaluation of the personality measure along with the German wording of the items, see Gerlitz and Schupp (2005; for a translation, see Table S14 in the supplemental material). An economic and efficient assessment (i.e., minimizing the number of items and maximizing validity) was prioritized over high reliabilities in the construction of the Big Five scales in the SOEP (Gerlitz & Schupp, 2005).

Consequently, as indicated by previous studies (e.g., Lucas & Donnellan, 2011; Specht et al., 2011; J. Wagner et al., 2019), the internal consistencies tended to be low. In our sample, we also obtained low internal consistencies across the four waves for Neuroticism ($\alpha = .59$ to $.62$; $\omega = .61$ to $.64$), Extraversion ($\alpha = .66$; $\omega = .67$), Conscientiousness ($\alpha = .58$ to $.62$; $\omega = .58$ to $.63$), Agreeableness ($\alpha = .48$ to $.51$; $\omega = .49$ to $.53$), and Openness ($\alpha = .60$ to $.63$; $\omega = .60$ to $.63$; for more details, see Tables S15 to S17 in the supplemental material). Accordingly, latent modeling to control for measurement error was especially called for.

Results

Measurement Invariance

We conducted several LSEMs to test the Big Five for measurement invariance. The corresponding fit indices are reported in Table 3. As indicated therein, configural invariance (i.e., the same pattern of factor loadings across time and age) was clearly supported. Residual

variances of the same item were allowed to freely covary across the four waves, but to avoid having nonpositive definite residual covariance matrices, we had to fix the residual covariances of one of the Agreeableness items to 0 (for a similar case, see Lucas & Donnellan, 2011); when freely estimated, these residual covariances were negative, which impeded model interpretation. To test for metric measurement invariance, we successively constrained the factor loadings to be equal across measurement waves and across age. In parallel with the results reported for HILDA, we were able to establish metric invariance across both time and age (see Table 3).

Rank-Order Stability of Personality

As before, we assumed measurement invariance across time and age to allow for valid conclusions regarding the development of the rank-order stability of personality with age. The latent stability correlations of the Big Five across the 4-, 8-, and 12-year intervals are depicted in Figure 4. To consistently display age-specific rank-order stabilities that are based on all four waves, we used 18 years as the lower bound for age (as described above, in 2005, some younger participants were also included in the sample).

Generally, the results were comparable to those of Study 1. First, personality was neither perfectly stable nor completely unstable. Rather, the Big Five showed moderate to high stabilities. This applied across ages and retest intervals, with the lowest rank-order correlations ranging from .34 to .66, and the highest ranging from .61 to .81. Second, the more time that passed between measurement occasions, the lower the rank-order correlations became, and this applied consistently to all Big Five traits (Table S18 in the supplemental material presents the descriptive statistics for the rank-order stabilities). In addition, the Big Five differed in their stabilities: Extraversion and Neuroticism were somewhat more stable than Openness and

Agreeableness. Conscientiousness had the lowest stability estimates (see Figure 4 and Table S18 in the supplemental material).

Table 3

Fit Indices for Testing for Measurement Invariance (MI) Across Time and Age for the Big Five in the Socio-Economic Panel

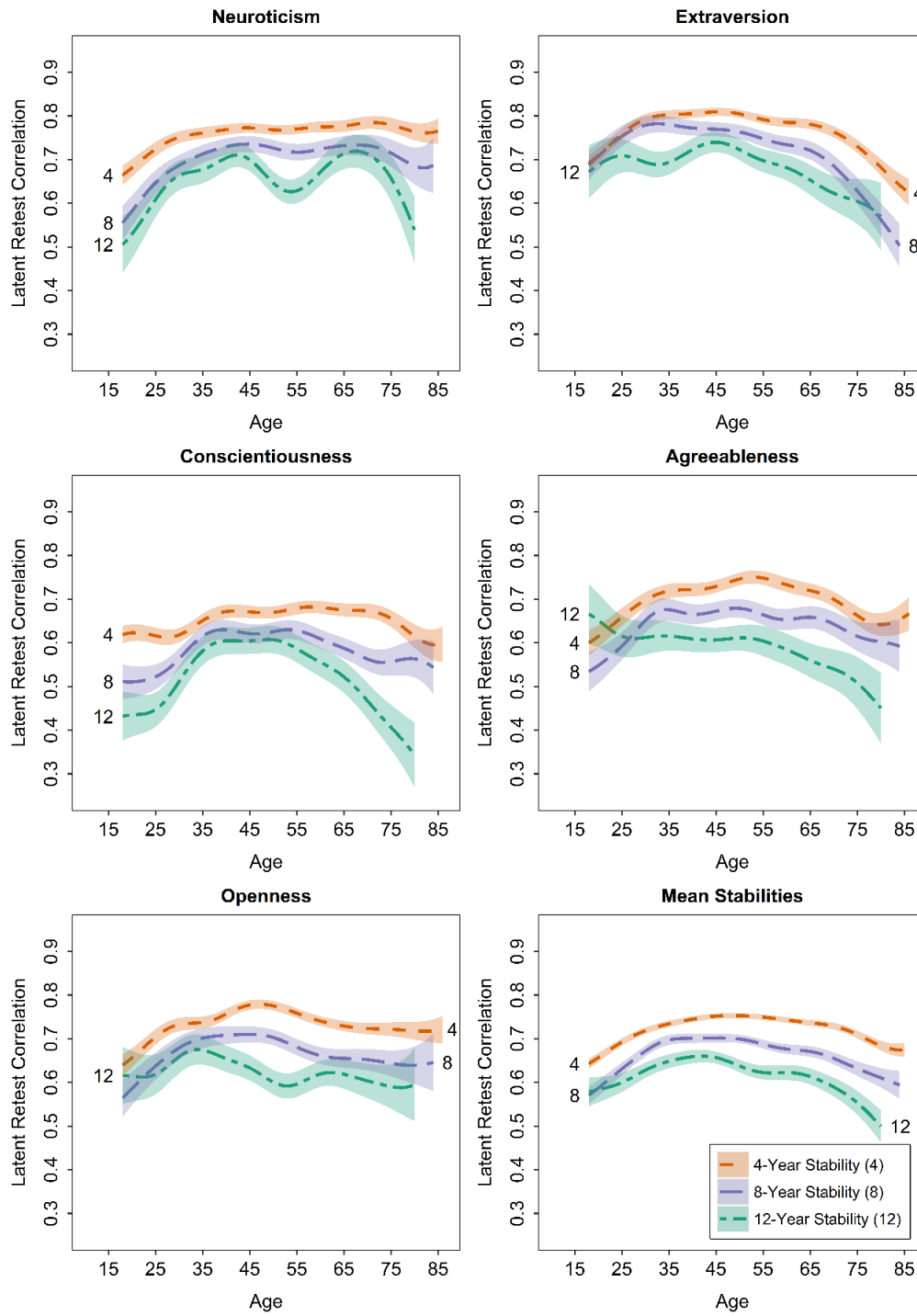
Model	<i>n</i>	CFI	RMSEA	SRMR
Neuroticism	21,768			
Configural MI		.997	.020	.014
Metric MI across time		.996	.020	.017
Metric MI across time and age		.995	.021	.020
Extraversion	21,764			
Configural MI		.994	.030	.025
Metric MI across time		.994	.028	.027
Metric MI across time and age		.990	.036	.038
Conscientiousness	21,760			
Configural MI		.993	.029	.024
Metric MI across time		.993	.027	.025
Metric MI across time and age		.990	.030	.029
Agreeableness ^a	21,772			
Configural MI		.991	.026	.019
Metric MI across time		.991	.024	.021
Metric MI across time and age		.991	.024	.022
Openness	21,759			
Configural MI		.999	.014	.012
Metric MI across time		.999	.012	.014
Metric MI across time and age		.999	.012	.016

Note. Local structural equation modeling was used for the analysis. Configural MI assumes the same pattern of factor loadings across time and age. Metric MI assumes the same factor loadings across time (and age). CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

^a To avoid having nonpositive definite residual covariance matrices, the residual covariances of one Agreeableness item were fixed to 0.

Figure 4

The Rank-Order Stabilities of the Big Five for the 4-, 8-, and 12-Year Intervals Across Age in the Socio-Economic Panel



Note. Shaded areas represent the bootstrapped standard error.

Third, the stability of the Big Five changed with age. Within the 18 estimated age trajectories, the rank-order stabilities often varied substantially: The distance from the lowest to the highest retest correlation ranged from .09 to .28 (the standard deviation of the rank-order estimates within a trajectory ranged from .03 to .08). Considering the trajectories in Figure 4, the rank-order stability of the Big Five seemed to develop systematically with age. From adulthood to roughly 50 years of age, the stability increased before it decreased at higher ages. However, this inverted U-shaped pattern was less pronounced compared with the results from Study 1, and some developmental courses in Figure 4 appeared to deviate from the overall trend. The trajectories of the 12-year rank-order stabilities for Neuroticism, Agreeableness, and Openness in particular showed different shapes. Averaging the age trajectories across the Big Five resulted in a clearly visible inverted U-shape (see the right bottom panel in Figure 4).

Test of Age Trajectories

Because the age trajectories in Figure 4 did not reveal unambiguous trends, a more formal test of the trajectory of stabilities was indicated. As in Study 1, we ran several regression models (intercept-only, linear, exponential, and quadratic) and compared the model fits using the BIC (see Table 4; for the estimated regression parameters, see Table S20 in the supplemental material). With three exceptions, the quadratic function again provided the comparatively best model fit. For the three exceptions (i.e., the 4- and 12-year stabilities for Neuroticism and the 12-year stability for Agreeableness), an exponential model had a better fit.

The left column of Figure 5 displays the age trajectories of the stabilities averaged across the Big Five dimensions along with the quadratic regression curves that were fitted. For all three time intervals, the quadratic regressions confirmed an inverted U-shaped age curve. Figures S6 to S10 in the supplemental material present the shapes of the best-fitting models for each

individual Big Five dimension. Whenever the quadratic model provided the best fit, the fitted model indicated an inverted U-shape—except for the 12-year stabilities for Openness. Regarding the three cases in which an exponential model fit best, the 4-year-stability of Neuroticism showed an increase reaching a plateau, whereas the 12-year rank-order stability of Agreeableness was best described by an exponential decline (i.e., stability was first stable on a plateau and then decreased). In the third case, for the 12-year stability of Neuroticism, the exponential model seemed to be misspecified because the underlying data showed a distinct inverted W-shaped pattern. Despite these three deviating cases, averaging the stability estimates across the five factors closely mirrored a quadratic pattern (see the left column of Figure 5). Taken together, the model comparisons provided further but not unequivocal evidence for an inverted U-shaped pattern for the development of personality stability.

In a similar vein, the two-lines test (Simonsohn, 2018a) also yielded support for an inverted U-shaped pattern of the development of personality stability. For all stabilities averaged across the five dimensions and for most of the 15 dimension-specific age trajectories, the two-lines test showed a significant positive slope for the first line and a significant negative slope for the second line (for the average stabilities, see the right column of Figure 5; for the individual Big Five dimensions, see Figures S6 to S10 in the supplemental material). Again, only a few stability trajectories of the single personality factors deviated from an inverted U-shaped pattern (for the model parameters, see Table S21 in the supplemental material).

Table 4

Comparisons of Regression Models to Describe the Development of Rank-Order Stability With Age in the Socio-Economic Panel

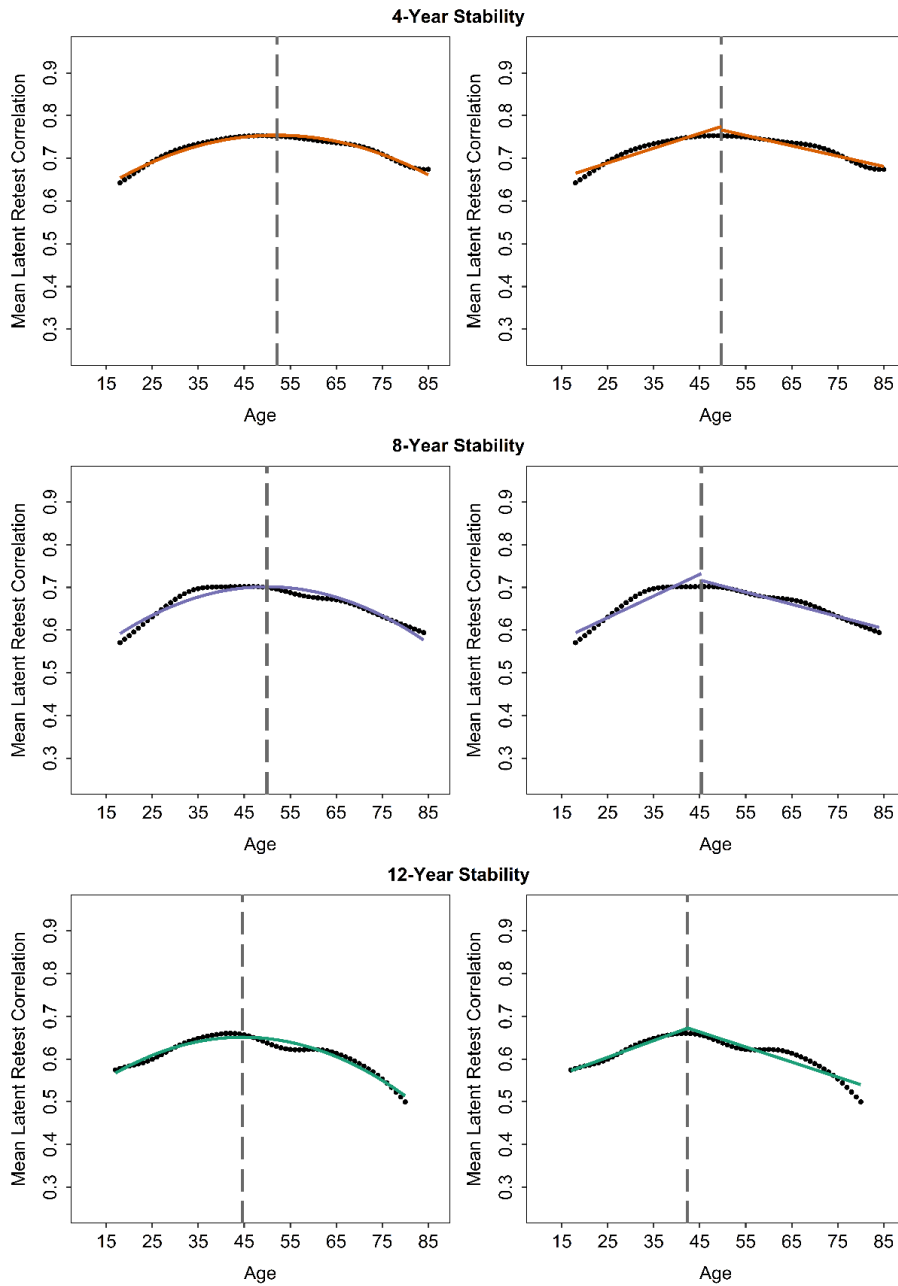
Model	BIC		
	4-year stability	8-year stability	12-year stability
Neuroticism			
Intercept	-290.270	-225.172	-176.338
Linear	-330.557	-242.277	-182.465
Exponential	-501.001	-357.817	-225.289
Quadratic	-421.831	-359.378	-224.882
Extraversion			
Intercept	-210.718	-153.926	-209.297
Linear	-222.809	-200.069	-245.236
Exponential	-291.129	-322.256	-310.385
Quadratic	-431.535	-391.580	-367.566
Conscientiousness			
Intercept	-288.602	-235.459	-136.383
Linear	-285.238	-232.412	-134.100
Exponential	-307.097	-266.045	-166.454
Quadratic	-400.052	-340.743	-335.395
Agreeableness			
Intercept	-237.165	-238.486	-200.485
Linear	-232.940	-235.022	-297.333
Exponential	-263.734	-284.395	-376.978
Quadratic	-390.532	-362.011	-338.289
Openness			
Intercept	-278.254	-251.928	-279.127
Linear	-276.109	-247.782	-299.577
Exponential	-333.167	-288.619	-301.774
Quadratic	-361.176	-315.488	-308.080
Mean stabilities			
Intercept	-275.324	-240.430	-228.852
Linear	-271.431	-237.153	-236.569
Exponential	-309.506	-262.382	-292.866
Quadratic	-523.500	-399.857	-412.353

Note. The relatively best-fitting model is displayed in bold. BIC = Bayesian information

criterion.

Figure 5

Quadratic Regression Model (Left Column) and Two-Lines Test (Right Column) for the Mean Rank-Order Stabilities in the Socio-Economic Panel



Note. Each dot represents the rank-order stability estimate for an individual year of age. The dashed vertical line indicates the highest point in the quadratic function (left column) or the break point in the two-lines test (right column).

As in Study 1, we estimated the age at which personality was most stable by identifying the maximum of the quadratic regression function and the break point in the two-lines test. We focused on the three trajectories of the stabilities averaged across the Big Five because, for these trajectories, an inverted U-shape was most apparent (see Figure 5). According to the quadratic functions, averaged across the three retest intervals, personality stability reached a maximum at 48.90 years of age ($SD = 3.84$). The results of the two-lines test were similar: Averaged across retest intervals, the mean break point was at 45.80 years of age ($SD = 3.73$).

Summary of Study 2

We conducted Study 2 to test the robustness of the results we found in Study 1. In general, the results for the German sample (SOEP; Study 2) were similar to the results in the Australian sample (HILDA; Study 1). That is, in the SOEP, we found that the Big Five showed a considerable amount of stability across several years (e.g., Anusic & Schimmack, 2016). Further, the more time that passed between the personality measurements, the lower the stability became (e.g., Conley, 1984). Again, we found some evidence in support of the cumulative continuity principle (Roberts & Nickel, 2017): The rank-order stability of the Big Five increased with age until middle adulthood. Looking at the trajectory of stability after middle adulthood, the data again indicated a decrease after age 50 (e.g., Wortman et al., 2012), which could be seen most clearly for the mean stability across the Big Five. But whereas Study 1 delivered strong support for the inverted U-shape, the findings in Study 2 were somewhat less consistent such that single trajectories deviated from this age trajectory.

Additional Analyses

Explaining Discrepancies Between Our and Recent Findings

Our approach considerably overlapped with J. Wagner et al.'s (2019) contribution, yet our results contrasted with theirs as we found much stronger evidence for systematic age-related changes in rank-order stability. Several of our methodological decisions differed and may plausibly explain these discrepancies. To pin down which decisions substantially affected the conclusions, we made small-step adjustments to move from the analyses by J. Wagner et al. to our analyses and kept track of the resulting stability trajectories (for more details, see the supplemental material).

In a preliminary step, using the same data as J. Wagner et al. (2019; i.e., the first three waves), and applying their data-analytic decisions, we successfully reproduced the developmental trajectories for the single Big Five personality traits they reported in their work (see Figure S13 in the supplemental material). In a next step, we averaged the 4-year stabilities (which were the only coefficients reported by J. Wagner et al., 2019) across the Big Five traits to arrive at mean stability estimates, which allowed us to compare their findings to our main results. Then, step by step, we changed the analyses and recalculated the stability trajectories; Figure 6 contrasts all different sets of results. First, we added the fourth wave of personality data. Second, we made several adjustments to the model. Most notably, we set the factor loadings equal across age to ensure that personality was measured invariantly across age, and we modeled single items rather than item parcels in the HILDA data set (for more details about the second step, see Table S23 in the supplemental material). Third, we freed the stabilities to be noninvariant across time to allow for changes in stability *within* participants as they aged. Fourth, instead of including only cases with complete personality data on all four waves, we included all informative cases

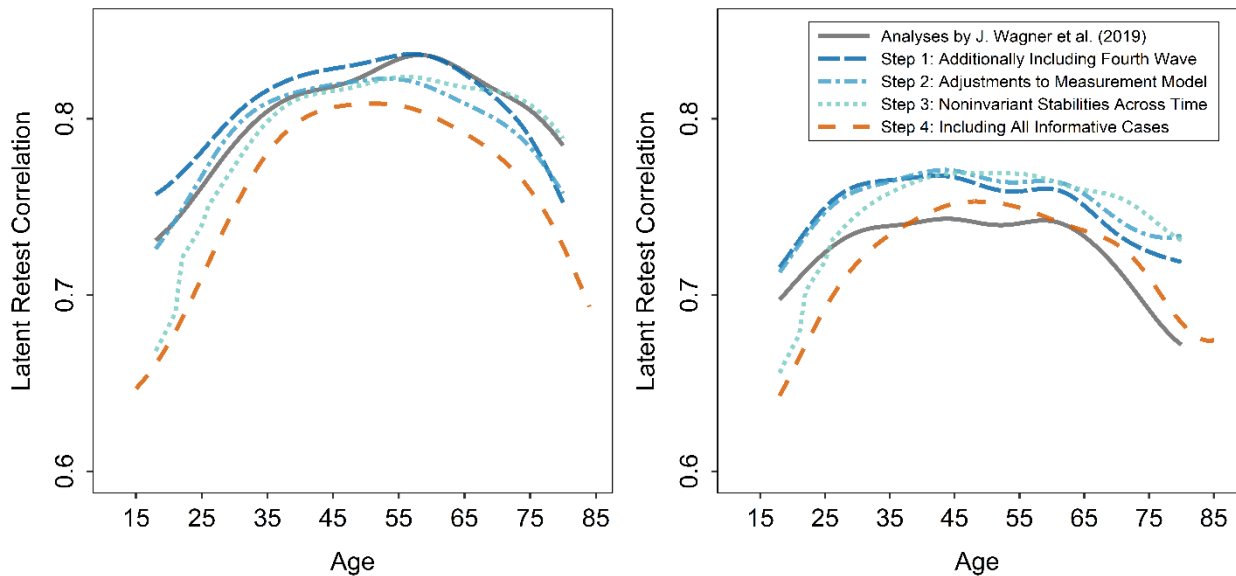
with data on at least two waves. Notably, this change led to a substantial increase in sample sizes: from $n = 6,289$ to $n = 15,465$ in HILDA and from $n = 7,075$ to $n = 21,777$ in SOEP. After these changes, the statistical analyses were the same as the methodology used in the present studies (mirroring the age trajectories in Figures 2 and 4).

Taking a closer look at the resulting trajectories of the mean 4-year stabilities in Figure 6, intriguingly, even the original analyses by J. Wagner et al. (2019) already showed some evidence of an inverted U-shaped age curve (which was not that obvious in the trait-specific curves reported by J. Wagner et al., 2019; for more detailed results including the individual Big Five dimensions, see Figures S14 to S25 in the supplemental material).⁵ Adding another wave of personality data in Step 1 did not change the general pattern but only slightly increased the general age-independent level of stability. The several adjustments made to the measurement model in Step 2 also did not affect the observed age curves. But, allowing the stabilities to be invariant across time in Step 3 led to a sharper increase in personality stability for younger ages. As allowing for such invariance across time is fully compatible with (and indeed implied by) the notion that stability changes with age, we believe that the equality constraint in J. Wagner et al. led to an underestimation of the increase in stability in young adulthood. Lastly, including all informative cases in Step 4 led to a more distinct decline in personality stability in older age. That is, including only cases with complete data in their analyses seems to have led to an underestimation of the decrease in older adulthood in the J. Wagner et al. study. Notably, these changes were quite consistent across HILDA and SOEP.

⁵ J. Wagner et al. (2019) formally examined the shape of the age distribution of stability by testing whether a stability estimate for a certain age deviated from the mean level of stability across all ages. We suggest that this procedure might not be sensitive enough to detect nuanced developmental trends within a relatively flat U-shape.

Figure 6

Impact of Methodological Changes on Mean 4-Year Rank-Order Stabilities in the Household, Income and Labour Dynamics in Australia Survey (Left Side) and the Socio-Economic Panel (Right Side)



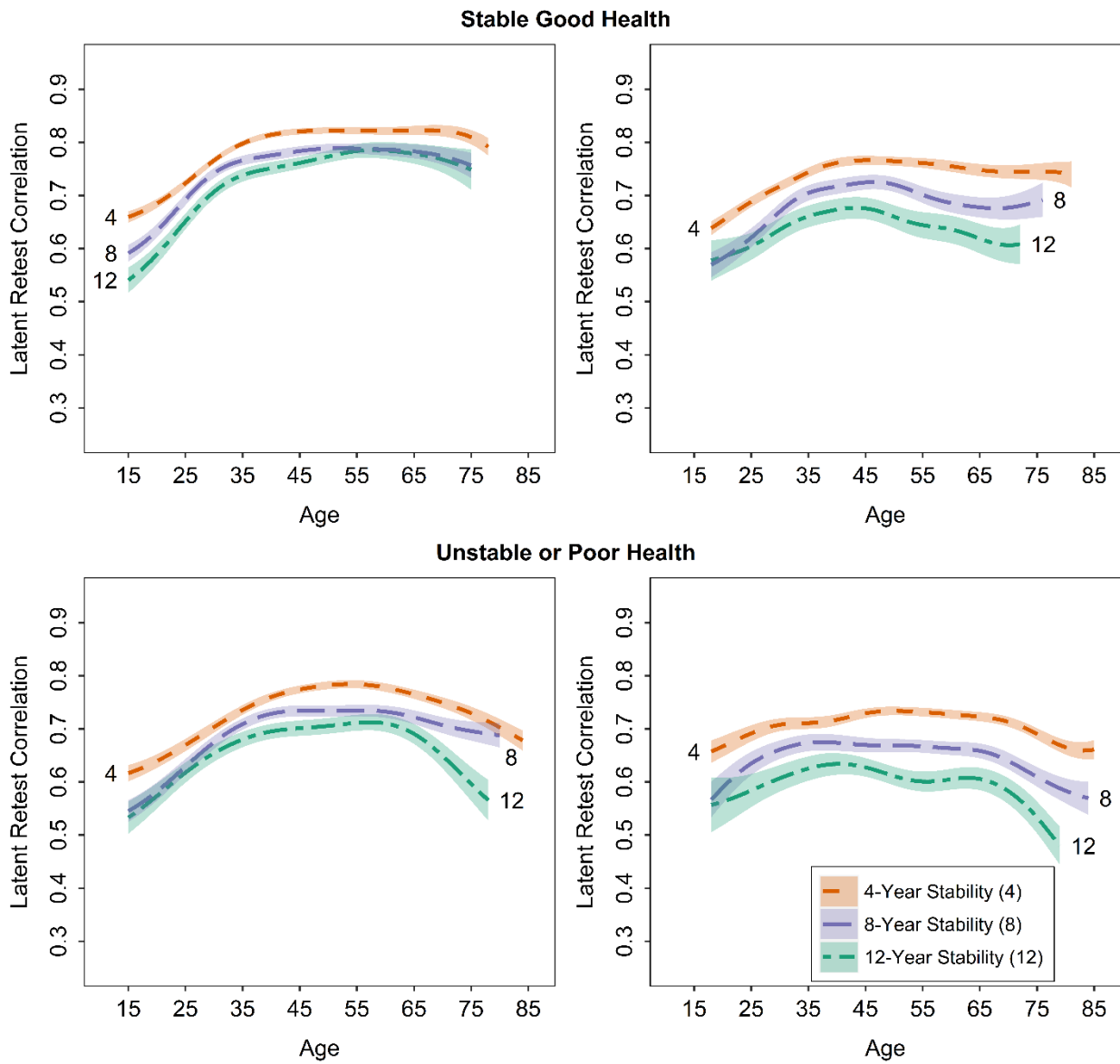
In summary, we identified two main reasons that explain why we found stronger age effects on personality stability in comparison with J. Wagner et al. (2019): idiosyncrasies in the statistical model itself (i.e., assuming that stability is constant within individuals as they age) and the reliance on complete cases, which suppresses a decline in stability in older age. The exclusion of incomplete responses may result in an overly healthy sample, thus resulting in underestimations of age-related changes in stability, especially for older age as changes in health go along with changes in personality (Kornadt et al., 2018; Mueller et al., 2018). Indeed, studies reporting an inverted U-shaped stability pattern in personality development have suggested that health-related processes are a key factor that accounts for the decrease in stability in older age (Ardelt, 2000; Lucas & Donnellan, 2011; Wortman et al., 2012).

Exploring the Role of Health on Personality Stability Development

This explanation that stability trajectories depend on which cases are included leads to the prediction that, if we limit analyses to a sample of respondents who are in good health over the course of the study, the decrease in personality stability will be less pronounced compared with a sample of respondents with unstable or poor health. We tested the hypothesized role of health on personality stability development by splitting the sample according to self-reported health into two groups: stable good health and unstable or poor health. In HILDA, self-reported health was assessed by asking “In general, would you say your health is” with response options *Excellent* (1), *Very Good* (2), *Good* (3), *Fair* (4), and *Poor* (5). Similarly, in SOEP, self-reported health was assessed with the item “How would you describe your current health?” with the answers *Very Good* (1), *Good* (2), *Satisfactory* (3), *Less Good* (4), and *Bad* (5). In both samples, we ascribed stable good health to respondents who respectively reported at least *Good* or *Satisfactory* self-reported health (3 or better), respectively, in all yearly surveys in which they participated between 2005 and 2017 (i.e., the time span of personality data collection). All other individuals were grouped under the label unstable or poor health. According to these criteria, in HILDA, there were $n = 9,191$ participants with stable good health ($M_{\text{age}} = 41.79$; $SD_{\text{age}} = 17.18$; 53% female) and $n = 6,274$ with unstable or poor health ($M_{\text{age}} = 50.81$; $SD_{\text{age}} = 18.47$; 54% female). In SOEP, $n = 11,237$ participants indicated stable good health ($M_{\text{age}} = 46.87$; $SD_{\text{age}} = 16.96$; 50% female) and $n = 10,540$ unstable or poor health ($M_{\text{age}} = 56.38$; $SD_{\text{age}} = 16.65$; 56% female; for trait-specific sample sizes and analysis details, see Table S24 in the supplemental material). We applied the same analyses as in Studies 1 and 2 to the split samples. The resulting trajectories for the mean stabilities are depicted in Figure 7 (for the individual Big Five dimensions, see Figures S26 to S30 in the supplemental material).

Figure 7

Health and the Mean Rank-Order Stabilities for the 4-, 8-, and 12-Year Intervals Across Age in the Household, Income and Labour Dynamics in Australia Survey (Left Column) and the Socio-Economic Panel (Right Column)



Note. Shaded areas represent the bootstrapped standard error.

Both samples showed an increase in personality stability from young to middle adulthood. More importantly, in the samples with stable high self-reported health, personality stability remained on a plateau throughout late adulthood. By contrast, in the samples characterized by unstable or low self-reported health, personality stability more clearly declined in older age. These trends were somewhat more consistent in HILDA than in SOEP (see Figure 7). Taken together, the results were consistent with our expectations regarding the role of health as a key factor in the decrease in personality stability in older age (see, e.g., Lucas & Donnellan, 2011). As a consequence, our additional analyses suggest that the inverted U-shaped pattern in personality stability development may be more difficult to detect in a sample of complete cases because the relevant variability in health has been removed. This pattern may also imply that the trajectories we presented in our FIML analyses based on all informative cases still underestimated the decline in stability in old age: The data were missing not at random (MNAR) in such a manner that individuals with the steepest health declines were likely not sufficiently represented in our data.

General Discussion

How does the rank-order stability of personality develop across the life span? No conclusive answer has been given to this question, and previous findings may lack precision due to methodological limitations. For example, meta-analyses (e.g., Roberts & DelVecchio, 2000) have reported that stability increases until midlife and plateaus thereafter, but older participants have been only sparsely included. More recent studies with age-representative samples (e.g., Specht et al., 2011; Wortman et al., 2012) have found that stability decreases later in life, resulting in an inverted U-shaped life span trajectory; however, these studies have been limited in their flexibility to accurately capture age trajectories in stability by relying on either age

groups or superimposed developmental patterns. Analytical advancements allowed the most recent contribution (J. Wagner et al., 2019) to overcome these limitations by providing continuous and more finely grained developmental trajectories. Surprisingly, J. Wagner et al. found overall comparatively weak age-related changes in stability, and in particular only sporadic evidence for a decline in stability in older age. It remains unclear why an inverted U-shaped pattern in personality stability development cannot be seen consistently.

Our study aimed to resolve the matter. We attempted to overcome the limitations of previous work by analyzing two panel studies that were representative of all ages and by implementing a latent modeling procedure that was similar to the one employed by J. Wagner et al. (2019) but with some crucial improvements, and we tracked how these changes affected the resulting stability trajectories. These analyses enabled us to descriptively display the age curves of stability and to formally test the different developmental assumptions stated in the literature. Furthermore, using Australian and German samples allowed us to assess the robustness of our results.

Sketching the Development of Rank-Order Stability Across the Life Span

Overall, we found clear evidence of an inverted U-shaped pattern for the development of personality stability across the life span. Both descriptive age trajectories and formal tests supported this developmental trend. With some qualifications, the pattern of findings generally held across samples, personality traits, and retest intervals. Despite these pronounced age-related changes in stability, we also confirmed the general trait character of the Big Five across all ages (Anusic & Schimmack, 2016; Costa et al., 2019; Damian et al., 2019): Even the lowest coefficient we found ($r = .34$) could be classified as a medium to large effect (Funder & Ozer, 2019; Gignac & Szodorai, 2016).

In support of the principle of cumulative continuity, we found that personality stability increased from adolescence to middle adulthood (Briley & Tucker-Drob, 2014; Ferguson, 2010; Roberts & DeVecchio, 2000). On average, personality stability peaked at about 50 years of age in our analyses. Although some previous studies suggested a similar age (Ardelt, 2000; Lucas & Donnellan, 2011; Roberts & DeVecchio, 2000; Specht et al., 2011; Wortman et al., 2012), others located the peak earlier at 30 years of age (Briley & Tucker-Drob, 2014; Ferguson, 2010). This difference may be attributed to sampling issues: Briley and Tucker-Drob (2014) and Ferguson (2010) mostly aggregated studies that included younger participants, leading to a loss of precision for age trends later in life. In line with this reasoning, a peak of stability at 50 years of age has especially been found to be present in studies with age-representative samples (Donnellan et al., 2012; Milojev & Sibley, 2014; Specht et al., 2011; Wortman et al., 2012). Further, we consistently showed that personality stability decreased after middle adulthood (Ardelt, 2000; Lucas & Donnellan, 2011; Milojev & Sibley, 2014; Specht et al., 2011; Wortman et al., 2012). This finding is in contrast with previous studies that reported a plateau in older age (Briley & Tucker-Drob, 2014; Ferguson, 2010; Roberts & DeVecchio, 2000), which might again be attributed to a lack of older respondents in the respective samples.

Intriguingly, our results contrast with findings from the most recent major study on personality stability (J. Wagner et al., 2019), which reported comparatively less evidence for age-related changes in personality stability—based on data that considerably overlapped with the data in the present study (i.e., the first three waves of Big Five data in the HILDA and SOEP studies) and a similar methodology within the LSEM framework. In additional analyses, we showed that two methodological decisions in J. Wagner et al. in particular explained the deviating findings: Assuming that stability does not change within participants as they age, and

excluding participants from the analyses with incomplete data flattened the age trajectories. Considering only complete cases does not lead per se to incorrect estimates—but we demonstrated that the resulting age trajectories, which lacked a decline in older age, were only valid for a nonrepresentative overly healthy sample. A sample of participants with changing or poor health indeed showed a more pronounced decline, confirming the inverted U-shaped pattern in development. Thereby, our findings support the assumption that health-related changes play a crucial role in the decline in personality stability in old age (Ardelt, 2000; Lucas & Donnellan, 2011; Wortman et al., 2012).

Comparing the Age Trajectory Across Samples, Dimensions, and Time Intervals

Across the Australian (HILDA; Study 1) and German samples (SOEP; Study 2), the findings were in general comparable, which is reassuring given that different personality questionnaires were used. However, the findings were somewhat less clear in Study 2, with sporadic and rather unsystematic deviations from the generally inverted U-shaped pattern for some personality dimensions and some retest intervals. A possible explanation is that the scope of personality measurement in the SOEP was more limited than in HILDA, with only three items per dimension (in contrast to four to six in HILDA). Whereas the latent modeling approach ruled out the possibility that a lack of reliability was systematically distorting the findings, it could not compensate for differences in the coverage of the targeted construct. This may also be an explanation for the somewhat generally higher stabilities in HILDA.

As there were no consistent deviations in single personality traits from the inverted U-shaped trajectory, our results let us cautiously suggest that the Big Five personality dimensions exhibit no marked differences in their general shape in stability development (cf. Milojev & Sibley, 2014; Specht et al., 2011; J. Wagner et al., 2019). But taking a closer look, the HILDA

analyses suggest that the age at maximum stability might be some years earlier than 50 for Openness and Agreeableness and some years later for the other three dimensions. However, we could not verify the robustness of these results because, as described above, we were unable to determine the dimension-specific age for the highest stability with sufficient accuracy in the SOEP data. In addition, the Big Five varied in their general levels of stability within the HILDA and SOEP studies (the largest difference between traits was $r = .21$ and $.13$, respectively). Across the two studies, Extraversion tended to be the most stable, whereas the remaining traits did not show a systematic sequencing.

Our main finding was replicated across the 4-, 8-, and 12-year retest intervals, indicating an inverted U-shape of stability across the life span that was independent of the interval under investigation. However, as the length of the time interval between personality measurements increased, the general level of stability decreased, which is consistent with findings from previous studies (Anusic & Schimmack, 2016; Ardel, 2000; Conley, 1984; Ferguson, 2010; Fraley & Roberts, 2005; Roberts & DeLucchio, 2000). Expanding previous studies, we increased the retest interval up to 12 years, which, intriguingly, revealed that the decline in stability is more marked for longer time intervals. As an explanation, we suggest that differences in aging (e.g., variability in health trajectories) cumulate with time, and, hence, their differential effects on personality (lowering rank-order stability) are stronger for longer retest intervals.

Limitations

Despite the strengths of our approach, room for improvement in future studies is surely given. First, whereas our study provided evidence that the stability of all five traits followed the same general pattern (inverted U-shape), we could not conclusively answer whether there were systematic differences in the age at which stability peaked. And as we focused on the Big Five

personality traits on a general level, we further cannot preclude the possibility that facets or even nuances of Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness exhibit different stability trajectories (see Möttus et al., 2017). In a similar vein, future assessments could be broadened to incorporate aspects of personality development that go beyond the Big Five. For example, Honesty-Humility as a sixth dimension (Ashton & Lee, 2007) or surface characteristics (e.g., self-esteem, goals, and interests; Kandler et al., 2014) of personality may show different developmental patterns (for age trends in the 2-year stability of Honesty-Humility, see Milojević & Sibley, 2014).

Second, studies on personality development rely heavily on self-reports (but for exceptions, see, e.g., Göllner et al., 2017; Rohrer et al., 2018). Thereby, observed trait stability may be inflated by idiosyncratic effects in individuals' self-reported personality. Idiosyncrasies are certainly also present in personality judgments of an individual by a relevant other (e.g., friend or partner), but it might be illuminating to contrast stability development depending on the source of the information as it might also shed light on the mechanisms that underlie changes in stability (e.g., changes in the stability of reporting tendencies, changes in comparison groups, changes in the underlying true trait variance).

Third, whereas stability estimates always require longitudinal information, our age trajectories (just like the age trajectories in other studies) were still informed by between-subject information. For example, the individuals underlying the stability estimates at age 20 did not overlap with the individuals underlying the stability estimates at age 80. Therefore, we could not exclude the possibility that our age trajectories were also influenced by cohort effects. Our substantive interpretations partly rest on the assumptions that cohorts do not drastically vary with respect to their stability or with respect to age effects on stability.

Lastly, just like the majority of research on personality development, our study focused on samples originating from societies that are predominantly regarded as “western,” educated, industrialized, rich, and democratic (Costa et al., 2019), an issue that is generally present in psychological research (Arnett, 2008; Henrich et al., 2010). Future investigations should broaden the perspective and incorporate data from other geographical regions to address how generalizable these personality development patterns are. In addition, future studies should examine whether other aspects of culture (e.g., socioeconomic status, religion, and values; Cohen & Varnum, 2016; Qu et al., 2021) influence the development of personality stability.

Conclusion

We found that personality stability increases until middle adulthood, peaking at about 50 years of age, and then decreases again in later life. Furthermore, our analyses suggest that health-related processes account for the decline in stability in older age. These findings have implications for the investigation of personality development in general: Individuals who eventually drop out of the study should be included for as long as possible, as they might represent those most prone to frailty and, thus, those who are most likely to experience personality changes. Data collection efforts should invest resources into following up with older participants. Otherwise, we have little chance to fully understand personality development in old age.

References

- Anusic, I., & Schimmack, U. (2016). Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations. *Journal of Personality and Social Psychology, 110*(5), 766–781. <https://doi.org/10.1037/pspp0000066>
- Ardelt, M. (2000). Still stable after all these years? Personality stability theory revisited. *Social Psychology Quarterly, 63*(4), 392–405. <https://doi.org/10.2307/2695848>
- 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>
- Asendorpf, J. B. (2010). Developmental perspectives. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment: Vol. 1. Personality theories and models* (pp. 101–123). SAGE Publications. <https://doi.org/10.4135/9781849200462.n5>
- Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*(2), 150–166. <https://doi.org/10.1177/1088868306294907>
- 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>
- Block, J. (1971). *Lives through time*. Bancroft Books.

- 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>
- Briley, D. A., & Tucker-Drob, E. M. (2014). Genetic and environmental continuity in personality development: A meta-analysis. *Psychological Bulletin, 140*(5), 1303–1331.
<https://doi.org/10.1037/a0037091>
- Brown, M. I., Wai, J., & Chabris, C. F. (2021). Can you ever be too smart for your own good? Comparing linear and nonlinear effects of cognitive ability on life outcomes. *Perspectives on Psychological Science*. Advance online publication.
<https://doi.org/10.1177/1745691620964122>
- Canty, A., & Ripley, B. (2017). *boot: Bootstrap R (S-Plus) functions* (Version 1.3.20) [R package]. <https://CRAN.R-project.org/package=boot>
- Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology, 56*, 453–484.
<https://doi.org/10.1146/annurev.psych.55.090902.141913>
- 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

- Cohen, A. B., & Varnum, M. E. (2016). Beyond east vs. west: Social class, region, and religion as forms of culture. *Current Opinion in Psychology*, 8, 5–9.
<https://doi.org/10.1016/j.copsyc.2015.09.006>
- Cole, D. A. (2015). Latent trait-state models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 585–600). Guilford Press.
- Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological Methods*, 10(1), 3–20. <https://doi.org/10.1037/1082-989X.10.1.3>
- Conley, J. J. (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality and self-opinion. *Personality and Individual Differences*, 5(1), 11–25. [https://doi.org/10.1016/0191-8869\(84\)90133-8](https://doi.org/10.1016/0191-8869(84)90133-8)
- Costa, P. T., McCrae, R. R., & Löckenhoff, C. E. (2019). Personality across the life span. *Annual Review of Psychology*, 70, 423–448. <https://doi.org/10.1146/annurev-psych-010418-103244>
- Damian, R. I., Spengler, M., Sutu, A., & Roberts, B. W. (2019). Sixteen going on sixty-six: A longitudinal study of personality stability and change across 50 years. *Journal of Personality and Social Psychology*, 117(3), 674–695.
<https://doi.org/10.1037/pspp0000210>
- De Fruyt, F., Bartels, M., Van Leeuwen, K. G., De Clercq, B., Decuyper, M., & Mervielde, I. (2006). Five types of personality continuity in childhood and adolescence. *Journal of Personality and Social Psychology*, 91(3), 538–552. <https://doi.org/10.1037/0022-3514.91.3.538>

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>

Donnellan, M. B., Kenny, D. A., Trzesniewski, K. H., Lucas, R. E., & Conger, R. D. (2012).

Using trait-state models to evaluate the longitudinal consistency of global self-esteem from adolescence to adulthood. *Journal of Research in Personality*, *46*(6), 634–645.

<https://doi.org/10.1016/j.jrp.2012.07.005>

Ferguson, C. J. (2010). A meta-analysis of normal and disordered personality across the life

span. *Journal of Personality and Social Psychology*, *98*(4), 659–667.

<https://doi.org/10.1037/a0018770>

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for

conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review*, *112*(1), 60–74. [https://doi.org/10.1037/0033-](https://doi.org/10.1037/0033-295X.112.1.60)

[295X.112.1.60](https://doi.org/10.1037/0033-295X.112.1.60)

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>

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.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78.

<https://doi.org/10.1016/j.paid.2016.06.069>

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>

Göllner, R., Roberts, B. W., Damian, R. I., Lüdtke, O., Jonkmann, K., & Trautwein, U. (2017).

Whose "storm and stress" is it? Parent and child reports of personality development in the transition to early adolescence. *Journal of Personality, 85*(3), 376–387.

<https://doi.org/10.1111/jopy.12246>

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability:

What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930–944. <https://doi.org/10.1177/0013164406288165>

Grothendieck, G. (2013). *nls2: Non-linear regression with brute force* (Version 0.2) [R package]. <https://CRAN.R-project.org/package=nls2>

Harris, M. A., Brett, C. E., Johnson, W., & Deary, I. J. (2016). Personality stability from age 14 to age 77 years. *Psychology and Aging, 31*(8), 862–874.

<https://doi.org/10.1037/pag0000133>

- Hartung, J., Doeblner, P., Schroeders, U., & Wilhelm, O. (2018). Dedifferentiation and differentiation of intelligence in adults across age and years of education. *Intelligence*, *69*, 37–49. <https://doi.org/10.1016/j.intell.2018.04.003>
- Hartung, J., Engelhardt, L. E., Thibodeaux, M. L., Harden, K. P., & Tucker-Drob, E. M. (2020). Developmental transformations in the structure of executive functions. *Journal of Experimental Child Psychology*, *189*, Article 104681. <https://doi.org/10.1016/j.jecp.2019.104681>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29. <https://doi.org/10.1038/466029a>
- 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.
- 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>
- 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.
- Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2018). *semTools: Useful tools for structural equation modeling* (Version 0.5-1) [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>
- Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology*, 63(1), 52–59. <https://doi.org/10.1037/0022-006X.63.1.52>
- Kenny, D. A., & Zautra, A. (2001). Trait-state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 241–263). American Psychological Association.
- Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65(4), 457–474.
<https://doi.org/10.1007/BF02296338>
- 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>

- Liebig, S., Schupp, J., Goebel, J., Richter, D., Schröder, C., Bartels, C., Fedorets, A., Franken, A., Giesselmann, M., Grabka, M., Jacobsen, J., Kara, S., Krause, P., Kröger, H., Kroh, M., Metzging, M., Nebelin, J., Schacht, D., Schmelzer, P., . . . Deutsches Institut für Wirtschaftsforschung. (2019). *Socio-Economic Panel (SOEP): Data for years 1984–2017* (Version 34) [Data set]. <https://doi.org/10.5684/soep.v34>
- Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.
- Losoncz, I. (2009). Personality traits in HILDA. *Australian Social Policy*, 8, 169–198.
- 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>
- Lüdtke, O., Trautwein, U., & Husemann, N. (2009). Goal and personality trait development in a transitional period: Assessing change and stability in personality development. *Personality and Social Psychology Bulletin*, 35(4), 428–441. <https://doi.org/10.1177/0146167208329215>
- 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>
- Maslowsky, J., Jager, J., & Hemken, D. (2015). Estimating and interpreting latent variable interactions: A tutorial for applying the latent moderated structural equations method. *International Journal of Behavioral Development*, 39(1), 87–96. <https://doi.org/10.1177/0165025414552301>

- McAdams, D. P., & Olson, B. D. (2010). Personality development: Continuity and change over the life course. *Annual Review of Psychology, 61*, 517–542.
<https://doi.org/10.1146/annurev.psych.093008.100507>
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. <https://doi.org/10.1037/met0000144>
- Milojev, P., & Sibley, C. G. (2014). The stability of adult personality varies across age: Evidence from a two-year longitudinal sample of adult New Zealanders. *Journal of Research in Personality, 51*, 29–37. <https://doi.org/10.1016/j.jrp.2014.04.005>
- 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>
- 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>
- Newman, D. A. (2014). Missing data. *Organizational Research Methods, 17*(4), 372–411.
<https://doi.org/10.1177/1094428114548590>
- Olaru, G., Robitzsch, A., Hildebrandt, A., & Schroeders, U. (2020). *Examining moderators of vocabulary acquisition from kindergarten throughout elementary school using local structural equation modeling*. PsyArXiv. <https://doi.org/10.31234/osf.io/bcmd8>
- 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>
- 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>
- Qu, Y., Jorgensen, N. A., & Telzer, E. H. (2021). A call for greater attention to culture in the study of brain and development. *Perspectives on Psychological Science*, 16(2), 275–293.
<https://doi.org/10.1177/1745691620931461>
- Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, 126(1), 3–25. <https://doi.org/10.1037/0033-2909.126.1.3>
- 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., & Robins, R. W. (2004). Person-environment fit and its implications for personality development: A longitudinal study. *Journal of Personality*, 72(1), 89–110.
<https://doi.org/10.1111/j.0022-3506.2004.00257.x>
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1), 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.

- Robitzsch, A. (2019). *sirt: Supplementary item response theory models* (Version 3.6-4) [R package]. <https://CRAN.R-project.org/package=sirt>
- Rohrer, J. M., Egloff, B., Kosinski, M., Stillwell, D., & Schmukle, S. C. (2018). In your eyes only? Discrepancies and agreement between self- and other-reports of personality from age 14 to 29. *Journal of Personality and Social Psychology, 115*(2), 304–320. <https://doi.org/10.1037/pspp0000142>
- 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>
- 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
- Simonsohn, U. (2018a). Two lines: A valid alternative to the invalid testing of U-shaped relationships with quadratic regressions. *Advances in Methods and Practices in Psychological Science, 1*(4), 538–555. <https://doi.org/10.1177/2515245918805755>
- Simonsohn, U. (2018b). *Two-lines test* (Version 0.52) [R code]. <http://webstimate.org/twolines/>
- 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>

- Spiess, A.-N., & Neumeier, N. (2010). An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: A Monte Carlo approach. *BMC Pharmacology*, *10*(1), Article 6. <https://doi.org/10.1186/1471-2210-10-6>
- 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, University of Melbourne.
- Wagner, G. G., Frick, J. R., & Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP)—Scope, evolution and enhancements. *Schmollers Jahrbuch*, *127*(1), 139–169. <https://doi.org/10.2139/ssrn.1028709>
- 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>
- 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/llcs.v3i3.208>
- Wettstein, M., Tauber, B., Kuźma, E., & Wahl, H.-W. (2017). The interplay between personality and cognitive ability across 12 years in middle and late adulthood: Evidence for reciprocal associations. *Psychology and Aging*, *32*(3), 259–277. <https://doi.org/10.1037/pag0000166>
- Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety.

Educational and Psychological Measurement, 73(6), 913–934.

<https://doi.org/10.1177/0013164413495237>

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>

Zheng, A., Briley, D. A., Malanchini, M., Tackett, J. L., Harden, K. P., & Tucker-Drob, E. M. (2019). Genetic and environmental influences on achievement goal orientations shift with age. *European Journal of Personality*, 33(3), 317–336. <https://doi.org/10.1002/per.2202>