Sixteen Going on Sixty-Six:

A Longitudinal Study of Personality Stability and Change across 50 Years

Rodica Ioana Damian[1]

Marion Spengler[2]

Andreea Sutu[1]

Brent W. Roberts[3]

[1]University of Houston; [2]University of Tübingen; [3]University of Illinois at Urbana-Champaign

Abstract

How much do people's personalities change or remain stable from high-school to retirement? To address these questions, we used a large US sample ($N = 1,795$) that assessed people's personality traits in adolescence and 50 years later. We also used two independent samples, one cross-sectional and one short-term longitudinal ($N = 3,934$ and $N = 38$, respectively), to validate the personality scales and estimate measurement error. This was the first study to test personality stability/change over a 50-year time span in which the same data source was tapped (i.e., self-report). This allowed us to use four different methods (rank-order stability, mean-level change, individual-level change, and profile stability) answering different developmental questions. We also systematically tested gender differences. We found that the average rank-order stability was .31 (corrected for measurement error) and .23 (uncorrected). The average mean-level change was half of a standard deviation across personality traits, and the pattern of change showed maturation. Individual-level change also supported maturation, with 20-60% of the people showing reliable change within each trait. We tested three aspects of personality profile stability, and found that overall personality profile stability was .37, distinctive profile stability was .17, and profile normativeness was .51 at baseline and .62 at the follow-up. Gender played little role in personality development across the lifespan. Our findings suggest that personality has a stable component across the lifespan, both at the trait level and at the profile level, and that personality is also malleable and people mature as they age.

Keywords: personality traits; rank-order stability; mean-level change; profile stability; lifespan

Sixteen Going on Sixty-Six:

A Longitudinal Study of Personality Stability and Change across 50 Years

In Homer's epic poem, Odysseus, the legendary Greek king, returns home to Ithaca after 20 years of warfare and difficult journeys, only to find his wife, Penelope, faithfully waiting for him despite her numerous suitors. The lovers are happily reunited and Odysseus reclaims his kingdom. Perhaps even more impressive than legend is the love story of Jerzy Bielecki and Cyla Cybulska. The two fell in love in 1943 in the Nazi concentration camp Auschwitz. After managing to break out, they got separated, and through a series of misunderstandings they each came to presume the other dead. Cyla moved to Brooklyn and married, while Jerzy started a family in Poland. In 1983, Cyla told this story to her Polish house cleaner who told her that she had seen a man tell the same story on Polish television. The two were reunited a few weeks later. When Cyla arrived in Krakow, Jerzy gave her 39 roses, one for each year they had been apart. They became very good friends and visited each other regularly until 2005, when Cyla died. In 2010, when Jerzy was last interviewed before passing away, he said he was "still very much in love with Cyla" (Hevesi, 2011).

When hearing such stories, one must wonder what such reunions feel like. When considering the personality traits (i.e., the characteristic patterns of thoughts, feelings, and behaviors) that one exhibited in adolescence, how similar to their old self is that person likely to be 50 years later? Are sociable teens destined to become sociable older adults? And does our relative personality ranking with respect to other people endure over our entire lifespan? For example, if Cyla was more sociable than Jerzy when they were 16, how likely is it that she was still more sociable than Jerzy 50 years later? What about absolute levels of personality: Do people change across their entire lifespan? Were both Cyla and Jerzy perhaps a bit wiser, less

impulsive, when they met as older adults, than when they were teenagers? And do some people change more than others across the lifespan? What about the unique constellations of traits that people have? For example, if Cyla were more neurotic than she was sociable when she was an adolescent, how likely is it that the same idiosyncratic pattern of personality traits would characterize her 50 years later? Finally, are there gender differences in how people's personalities change as they age? Questions regarding the stability and change of personality across the entire lifespan are some of the most interesting, because there are very few longitudinal studies spanning over so many years. The present study seeks to address such questions by using a large US sample that was followed over 50 years.

A major insight from recent research on personality development (e.g., Costa & McCrae, 1988, 1994; Donnellan, Conger, & Burzette, 2007; Fraley & Roberts, 2005; Hampson & Goldberg, 2006; Roberts & Del Vecchio, 2000; Roberts, Walton, & Viechtbauer, 2006; Terracciano, Costa & McCrae, 2006) is that personality traits are both stable and changeable. Personality traits are defined as relatively enduring, automatic patterns of thoughts, feelings, and behaviors that are relatively consistent across a wide variety of situations and contexts (Roberts, 2009). The most commonly used personality trait framework is the Big Five (John, Naumann, & Soto, 2008) or Five-Factor Model (McCrae & Costa, 2008), which includes five broad traits: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. The insight that personality traits are both stable and changeable reflects, in part, the systematic use of different methodological approaches to assess stability and change in individual traits (Block 1971). The two most prominent and most widely used approaches are rank-order stability and mean-level change. Rank-order stability (or differential stability) refers to the relative placement of a person (on a specific trait) within a group over time. A typical research question would be:

Does a 16-year-old who is conscientious relative to her peers develop into a 66-year-old who is also more conscientious compared to her peers? As described below, assessing rank-order stability across the lifespan, and especially over very long periods of time, is essential for understanding personality development and whether personality traits may be caused, in part, by continuous factors, such as (a) certain components of the genetic system (Roberts, in press) and (b) individuals seeking consistent roles and/or environments across time (Roberts & Damian, in press; Roberts & Robins, 2004; Harms, Roberts, & Winter, 2006). Mean-level change refers to how the average level of a trait across all individuals changes over time (see Caspi, Roberts, & Shiner, 2005). Mean-level change can be studied in cross-sectional studies with different cohorts (e.g., 16 and 66-year-olds assessed at the same time) or longitudinally, with the same cohort (e.g., assess people when they are 16 and then again 50 years later). Typical research questions would be: Are 16-year-olds less responsible than 66-year-olds? Or are people less responsible when they are 16 than 50 years later?

Although rank-order stability and mean-level change are essential in understanding personality development at the population level, they limit the understanding of development at the individual level (e.g., Roberts, Caspi, & Moffitt, 2001; Robins et al., 2001). Thus, it is important to investigate individual differences in change, that is, the magnitude of increase or decrease within each person over the course of the study. Typical research questions would be: Across the entire sample, what percentage of people showed reliable increases or decreases in responsibility going from 16- to 66-years-old?

Rank-order stability, mean-level change, and individual-level change are all "variable-centered" approaches (Block, 1971), meaning that they focus on stability and change in single personality traits. As a result, these approaches cannot account for the fact that personality is a

''peculiar patterning of attributes within the single person'' (Allport, 1954, p. 9), as opposed to a set of disconnected traits. Thus, a fourth way to conceptualize change and stability in personality is to take a "person-centered approach" and focus on the stability of the *pattern* of personality traits within a person across time, that is, personality profile stability (or ipsative stability). Assessing personality profile stability requires measurements of multiple personality traits that are ranked with respect to each other and that are collected across a minimum of two time-points (see Bleidorn et al., 2012; Furr, 2008; Klimstra et al., 2010). A typical research question would be: Does a 16-year-old, whose neuroticism is higher than her conscientiousness, develop into a 66-year-old whose neuroticism is still higher than her conscientiousness?

Rank-order Stability in Personality over Extended Periods of Time

Longitudinal studies that investigated the rank-order stability of personality traits across extremely long time-spans are rare and often plagued by methodological drawbacks (e.g., having obtained informant-reports at baseline and self-reports at the follow-up). Thus, it is unclear what level of rank-order stability we should expect across long periods of time (e.g., 50 years).

For example, several previous studies (Shiner et al., 2017; Edmonds, Goldberg, Hampson, & Barckley, 2013; Hampson & Goldberg, 2006) that assessed rank-order stability across long time spans, that is, across 20 years ($N = 205$; ages 10-30) and 40 years ($N = 799$, starting the assessment in elementary school with teacher reports and following-up with self-reports), found that the average rank-order stability of personality traits was about .20. Note that, although still substantial, this estimate is much lower than the rank-order stability evident over shorter time-spans (see Roberts & DelVecchio, 2000). In other words, the stability of traits over time should decrease as the time-span increases, but it should not asymptote to zero (Fraley & Roberts, 2005).

Contradicting previous research, a recent study suggested there may be little to no rank-order stability in personality traits when very long time-intervals are considered. Specifically, Harris and colleagues (2016) used a sample of 174 people to test the rank-order stability of personality over 63 years. At baseline (age 14), teacher reports were collected using only one item for each of six personality characteristics (Self-Confidence, Perseverance, Stability of Moods, Conscientiousness, Originality, and Desire to Excel). At the follow-up (age 77), self- and informant-reports (by close others) were collected on the same six personality items. The authors found no statistically significant rank-order stability. The test-retest coefficients found across 63 years when correlating teacher ratings at age 14 with self-ratings at age 77 ranged from -.05 (Perseverance) to .12 (Stability of Moods). When correlating teacher ratings at age 14 with ratings by close others at age 77, the test-retest coefficients ranged from -.14 (Desire to Excel) to .12 (Conscientiousness). The authors' conclusion that personality traits show little to no stability over extended periods of time is at odds with the previous research cited above, with two meta-analyses (Roberts & DelVecchio, 2000; Anusic & Schimack, 2016), and with genetic models which would predict some personality stability, even over long time-spans (e.g., Bleidorn, Kandler, Riemann, Angleitner, & Spinath, 2009). One possible explanation, which was put forward by the authors, is that the unusually long time-span (63 years) is responsible for the lack of observed stability. However, there are two alternative explanations, due to methodological issues: (a) it is possible the six items did not constitute a comprehensive measure of personality and (b) it is possible that the source of the lack of observed stability was, not the long time-interval, but the use of measurements from different informants at the two time-points (teacher vs. self; teacher vs. close other). Thus, the status of personality stability over extended time periods (i.e., over 50 years) remains unclear.

What level of stability in personality should we expect over a time span of five decades? Given the lack of long-term longitudinal data, it would be prudent to rely on estimates drawn from the corpus of prior longitudinal research, as an aggregate estimate would be an ideal benchmark upon which to base expectations. Accordingly, Fraley and Roberts (2005), used an aggregate of stability coefficients assessed at varying time intervals (including very long intervals) and with different starting ages to estimate the expected levels of stability over time and age. Using a variety of models to better incorporate developmental processes, such as random life events, person-environment transactions, and developmental constancies (e.g., genetic factors), Fraley and Roberts (2005) estimated that personality stability coefficients over long time spans (including over 50 years), when the first measurement was in adolescence, should asymptote at a value of about .20 on a correlational metric (not corrected for error), though estimates varied across different personality traits, with asymptotes ranging from .18 for Openness to .36 for Conscientiousness. However, their data did not include studies that covered a stability time span longer than 30 years. Thus, the current study provides the first opportunity to empirically test the stability of personality traits over a 50-year time span in which the same data source was tapped (i.e., self-report).

Mean-level Changes in Personality over Extended Periods of Time

Understanding mean-level change across the lifespan is important because it can further inform our knowledge of developmental processes. How do people change as they age? Like with studies on stability, longitudinal studies assessing mean-level change over extended periods of time are lacking. Such studies are important because without them we cannot discern whether the changes seen over shorter time-spans dissipate or cumulate with time.

Despite the lack of longitudinal studies investigating mean-level change over very long time-spans, research has made extensive progress in understanding principles of personality change. For example, a principle of personality change that has received extensive attention is the maturity principle, which can explain mean-level changes in personality over time. According to this principle, people become more psychologically mature with age, if maturity is defined as becoming more socially adapted, and, specifically, if being socially adapted is reflected in changes that increase a person's ability to negotiate social relationships and challenges more effectively (Roberts & Damian, in press). A meta-analysis of the mean-level changes in personality traits over time solidified empirical support for the maturity principle (Roberts, Walton, & Viechtbauer, 2006). This study found that most people become more agreeable, conscientious, and emotionally stable over their lifespan. Interestingly, the meta-analysis also showed robust increases in a facet of extraversion, described as social dominance, which reflects higher levels of assertiveness, self-confidence, and dominance.

The interesting question concerning long-term personality trait change is whether the changes seen over shorter time spans cumulate or dissipate. To the extent that personality traits are governed by "set points" that anchor the range of potential change that a person can realize in their life, then the longer the time span, the more likely a person will return to their set point (Fraley, 2002). On the other hand, if changes in personality cumulate over the life course, the longer the time period examined, the greater the amount of change that should occur. Using meta-analytic data, Roberts and colleagues (2006) found that changes in personality traits within different decades of life (e.g., from 20 to 30 or from 30 to 40), were each about a quarter to a third of standard deviation in the direction of maturation and that change was consistently positive across different age cohorts (note that, in this meta-analysis, longitudinal data spanned

across an average of 7 years, but used different age cohorts). Given these meta-analytic data,

they estimated that changes in specific personality traits, like conscientiousness and emotional

stability, should be around a full standard deviation between ages 20 and 70 *if* changes seen over

shorter time spans (10 years) cumulated. While consistent with cross-sectional estimates of age

differences in personality traits (Costa & McCrae, 1988), these estimates were extrapolations

from the data set rather than reflections of actual tests of how personality traits should behave

across the life course (because most longitudinal studies included in the meta-analysis did not

span for longer than 10 years). If the magnitude of personality trait change were lower than the

estimates from the meta-analysis when examined for the first time across 50 years (i.e., lower

than one standard deviation for traits like conscientiousness and emotional stability), this would

be consistent with a set point model that would argue for a braking system on change. On the

other hand, if personality trait change continued to accumulate, we would expect estimates closer

to half to one standard deviation (i.e., estimates that are higher than the quarter to a third of a

standard deviation change estimated over shorter time spans of 10 years). The latter potential

finding would be more consistent with a plasticity model of personality traits, and it would

contradict a strong set point model, indicating that once positive gains are made, they are likely

to continue in a form of a virtuous cycle.

Individual-Level Change in Personality

Although mean-level change can help us understand personality development at the

population level, it overlooks potential individual differences in change. The existence of

individual differences is pertinent to personality development in two ways. First, if they did not

exist, this would bolster the argument that normative changes in personality traits are universal

and uniform. If people demonstrated normative increases in traits without individual differences

in change, then one could argue that a universal genetic factor might be the cause of personality development (McCrae & Costa, 2008). On the other hand, if individual differences in personality trait change did exist, this would bolster the argument that personality trait change is contingent on each person's particular experiences. A recent review of prospective research (Bleidorn, Hopwood, & Lucas, 2018) showed that life experiences are associated with change in personality traits, and that different life experiences are differentially related to personality trait domains. Specifically, the most robust findings across the review were that transitioning to the first romantic relationship increased Extraversion and decreased Neuroticism (e.g., Neyer & Lehnart, 2007; Wagner, Becker, Lüdtke, & Trautwein, 2015), and that transitioning from high school to college/work increased Agreeableness, Conscientiousness, Openness, and decreased Neuroticism (e.g., Bleidorn, 2012; Lüdtke, Roberts, Trautwein, & Nagy, 2011). Furthermore, studies have found that life experiences are associated with personality change in middle- (van Aken, Denissen, Branje, Dubas, & Goossens, 2006) and old-adulthood (Mottus, Johnson, & Deary, 2012). Nevertheless, these effects of life experiences may be relatively modest and the evidence is still preliminary (Costa et al., 2000; Bleidorn et al., 2018). Thus, the existence of individual differences in personality trait change is key for understanding which model of personality development we should hold as an assumption—one that does not propose the influence of environmental experiences (McCrae & Costa, 2008) or one that does (Roberts, Wood, & Caspi, 2008).

One way to measure individual-level change in each personality trait, when only two waves of data are available, is to calculate the Reliable Change Index (RCI; Christensen & Mendoza, 1986; Jacobson & Truax, 1991), and classify people into three groups: decreased, increased, or stayed the same on each trait level. The Reliable Change Index is a widely used and

very conservative index of change that was primarily developed to assess whether the changes

resulted from therapeutic interventions were larger than chance (Jacobson & Truax, 1991).

We are not aware of any previous studies that used the RCI to assess individual-level

change in personality traits across very long periods of time (e.g., 50 years), but two previous

longitudinal studies of personality traits used the RCI to assess individual-level change across

four years (Robins et al., 2001) and across eight years (Roberts et al., 2001), in samples that had

starting ages similar to ours. Across both of these studies, the observed distributions of changers

and non-changers was significantly different from chance across all personality traits examined.

However, in both studies, the percentage of people found in the group who "stayed the same"

was quite large. Specifically, across four years, 73% (Neuroticism) to 91% (Openness) of the

people in the sample stayed the same (Robins et al., 2001). Across eight years, 72.2% (Negative

emotionality) to 84.4% (Constraint) of the people in the sample stayed the same (Roberts et al.,

2001). Notably, the percentages of people increasing or decreasing on various traits were

consistent with previously found mean-level change patterns, that is, they showed a maturation

trajectory, where a higher percentage of people increased (vs. decreased) in conscientiousness,

emotional stability, agreeableness, and dominance aspects of extraversion (Roberts et al., 2001).

Personality Profile Stability

Personality profile stability, also known as ipsative stability (Caspi & Herbener, 1990) or

within-person coherence (Biesanz & West, 2000), represents the degree to which a person's

unique pattern of traits remains stable across time (Ozer & Gjerde, 1989) and it is usually

assessed with q-correlations, that is correlations, within-individuals, between ranked sets of traits

across time.[1]

Previous studies that assessed personality with self-reports found average profile stability coefficients ranging from .37 (over a 15-year time span) to .85 (over a 3-year time span), with the distributions of these profile stability coefficients often ranging from -.95 to 1.00 (Block, 1971; Caspi & Herbener, 1990; De Fruyt et al., 2006; Donnellan et al., 2007; Klimstra et al., 2009; Lönnqvist et al., 2008; Ozer & Gjerde, 1989; Roberts et al., 2001; Robins et al., 2001). Notably, based on our review, the study with the longest time-span covered (i.e., 15 years, ages 20-35, $N = 74$) revealed an average personality profile stability of .37 ($SD = .32$), with individuals ranging from -.54 to .90 (Lönnqvist et al., 2008).

Extensive previous research has measured personality profile stability as a unitary construct. However, Furr (2008) argued that personality profile stability is not a unitary construct because of the so-called "normativeness problem." Specifically, Furr (2008) proposed that (a) most people's personality profiles will be similar to the normative profile (i.e., the average profile of the sample), as the degree of normativeness might reflect psychological adjustment and adaptation; and (b) normative profiles are likely stable across time, to the extent to which similar norms are relevant across different developmental periods. Thus, overall profile stability (i.e., "classic" profile stability, which was previously measured as a unitary construct) might reflect two separate processes: (a) the tendency to retain idiosyncratic (non-normative) personality profiles or (b) the tendency to be consistently normative across time.

To address the normativeness issue, Furr (2008) suggested decomposing overall profile stability into two additional components: distinctive profile stability (i.e., the degree to which a person's personality profile consistently diverges from the normative profile within the sample) and within-time normativeness (a.k.a., profile normativeness, or the degree to which a person's personality profile is similar to the average personality profile in a sample at the respective

measurement point). Notably, overall profile stability and distinctive profile stability require two measurement time points, whereas within-time profile stability can be computed at each measurement point.

Following Furr's (2008) recommendations, Klimstra and colleagues (2010) found, in a longitudinal study of 565 college students, that the average profile stability of self-reported personality traits was .74 (*SD* = .35) and the average distinctive profile stability was .61 (*SD* = .43), across four consecutive 1-year intervals. The average profile normativeness across the four time points was .58 (*SD* = .48), and seemed to increase with age (going from .54 at time 1 to .68 at time 4). In another study, Bleidorn and colleagues (2012) found, in a sample of 805 twin pairs (ages 34-46) who self-reported on their personalities, that, across two 5-year intervals, the average overall profile stability was .86 (*SD* = .23), whereas the average distinctive profile stability was .76 (*SD* = .29). Average profile normativeness ranged from .53 at Time 1 to .71 at Time 3, again showing an increasing trend with age. Across all stability coefficients, the range of estimates across participants was very wide, as previously reported going from -.96 to 1.00.

Across past research, we observed several trends. First, overall personality profile stability tended to be higher when the time span between assessments was shorter and to increase from childhood to late adolescence (see Klimstra et al., 2009; Ozer & Gjerde, 1989). Second, distinctive personality profile stability tended to be lower than overall profile stability (which should happen by definition, given Furr's theorizing on the normativeness problem; Furr, 2008). Third, profile normativeness was about .50 to .70 within time points, and it increased with age, where older people had more normative profiles (e.g., Bleidorn et al., 2012). Fourth, very few studies investigated profile stability over a timespan longer than eight years (e.g., Bleidorn et al., 2012; Caspi & Herbener, 1990; Roberts et al., 2001), and the longest time span covered was 15

years (Lönnqvist et al., 2008). And fifth, there were only a handful of studies that assessed

personality profile stability following Furr's (2008) recommendations, and thus, addressing the

normativeness issue.

Gender Differences in Personality Stability and Change

Previous studies have shown that men and women differ in their personality trait levels at

any given point in time, such that, women are higher in conscientiousness, agreeableness, the

sociability facet of extraversion, and lower in emotional stability and agency (e.g., Schmitt et al.,

2008; Costa et al., 2001; Roberts et al., 2001). However, cross-sectional gender differences may

not necessarily translate into gender differences in stability and change across the lifespan.

We are not aware of any previous studies on gender differences in personality stability

and change across very long time spans (e.g., 50 years), however, previous studies over shorter

time spans (e.g., 8-10 years) provide some clues. For example, a meta-analysis of 152

longitudinal studies that examined rank-order stability across the lifespan found no statistically

significant gender differences, that is, personality traits were similarly trait-like for both men and

women (Roberts & DelVecchio, 2000). Furthermore, a meta-analysis of 92 longitudinal studies

that examined mean-level change across the lifespan also found no statistically significant gender

differences, that is, men and women changed at similar rates across the lifespan (Roberts,

Walton, & Viechtbauer, 2006). Regarding gender differences in individual-level change and in

profile stability, one study, across 8 years (Roberts et al., 2001), found that men showed slightly

more reliable change, and that women showed slightly more personality profile stability,

however, it is worth noting that only overall profile stability was computed in this study (without

addressing the normativeness issue).

In sum, extensive previous research has shown that (a) personality traits show both rank-order stability and mean-level change over time, (b) there is reliable change at the individual-level, and (c) personality profiles show stability across time. Furthermore, scientists have started uncovering the developmental antecedents and processes underlying stability and change. Despite these advances, most previous research, including the meta-analyses described above, has focused on relatively short time-spans, up to 10 years, in different age cohorts. Thus, very little research has examined personality stability and change over the entire life-span (i.e., long periods of time, over 50 years). The reason for this oversight is a lack of longitudinal samples that included personality traits over the entire life-span. Nevertheless, the question of how personality develops over the entire lifespan is an important one. The critical research question becomes "If we are presented with an adolescent, can we reliably predict their personality when they are in their 60s?" And, do personality traits continue to change in a positive direction, such that personality trait change is cumulative across the life course? In the present investigation, we had access to self-reported personality traits recorded in adolescence, as well as self-reported personality traits recorded 50 years later, which enabled us to assess long-term stability and change both at the trait level and at the person level.

The Present Investigation

In the "Main Study" of the present investigation, we used a subsample of the Project Talent data set to test personality stability and change over a 50-year time span. Specifically, we used a large US sample of high-school students who had their personality assessed in 1960 and 50 years later ($N = 1,795$), using the Project Talent Personality Inventory (PTPI). Using these data, we tested (a) to what extent people maintained their relative standing on personality trait dimensions relative to others over time (i.e., rank-order stability), (b) to what extent people's

personality traits changed across time (i.e., mean-level changes), (c) what percentage of people

had reliable change in their personality traits (i.e., individual-level change), (d) to what extent

people's personality profiles remained stable across time (to address this question, we evaluated

three aspects of profile stability, that is, overall personality profile stability, distinctive profile

stability, and profile normativeness), and (e) gender differences in each of the four types of

continuity and change.

      Based on previous research and theory (Fraley & Roberts, 2005; Roberts & DelVecchio,

2000; Roberts et al., 2006; Caspi et al., 2005), we had several predictions. First, due to the long

time-span and the relatively young age of the participants (~16) at baseline, personality should

show lower levels of rank-order stability than it did in samples tested over shorter time spans and

where testing started at older ages (see Roberts & DelVecchio, 2000). Specifically, rank-order

stability over 50 years (with the first assessment in adolescence) should be about .20 on average,

and it should not differ markedly across different traits. Second, assuming a plasticity (vs. a set

point) model of personality change across the lifespan, mean-level changes in personality traits

across 50 years should be similar to the estimated cumulative mean-level changes estimated from

the meta-analysis by Roberts and colleagues (2006), with an average effect of half to one

standard deviation, and changes should be in the direction of maturation, such that people should

become more agreeable, conscientious, and emotionally stable over their lifespan, as well as

higher in dominance-related facets of extraversion. Regarding individual-level changes in

personality traits and personality profile stability, we did not have clear point-estimate

predictions, because, to our knowledge, no previous studies have examined individual-level

changes using the Reliable Change Index or profile stability across a time span longer than eight

and fifteen years, respectively. Nevertheless, regarding individual-level changes in personality

traits, and consistent with the idea that change across the lifespan is cumulative and depends on life experiences as opposed to being universal and uniform (Roberts et al., 2006; Roberts, Wood, & Caspi, 2008), we expected more than 20% of the people to show reliable changes on each personality trait, because 20% was the average percentage of "changers" found across shorter time spans within each trait (see Robins et al., 2001; Roberts et al., 2001). We also expected the percentages of people increasing or decreasing on various traits to correspond to the expected mean-level change patterns, that is, to show a maturation trajectory, where a higher percentage of people should increase (as opposed to decrease) in conscientiousness, emotional stability, agreeableness, and dominance aspects of extraversion (see Roberts et al., 2001). Regarding personality profile stability, we expected our estimate to be below .37, which was the overall profile stability observed across a 15-year time span (Lönnqvist et al., 2008). Furthermore, due to the normativeness issue (see Furr, 2008), we expected distinctive profile stability to show a much smaller effect than overall profile stability. We also expected profile normativeness within each time point to range between .50 and .70, and to increase slightly with age, which is what other studies have shown (e.g., Bleidorn et al., 2012). Finally, regarding gender differences, we expected to replicate previously found cross-sectional differences, whereby women should be higher in conscientiousness, agreeableness, the sociability facet of extraversion, and lower in emotional stability and agency at any given point in time (e.g., Schmitt et al., 2008; Costa et al., 2001; Roberts et al., 2001), but, given previous longitudinal research on shorter time spans (Roberts & DelVecchio, 2000; Roberts et al., 2006), which found no statistically significant gender differences, we did not expect gender differences in the patterns of personality stability and change across 50 years.

The inventory used at 50th year follow-up was an abbreviated form of the original scale, which necessitated preliminary work to be validated.[2] Therefore, we conducted additional analyses on two independent data sets ($N = 3,934$ and $N = 38$), which were previously collected by Pozzebon and colleagues (2013). Using the first sample ($N = 3,934$), which was cross-sectional, we obtained descriptive statistics and correlations between long- and short-forms of the PTPI. We also tested mean-level personality differences between two different age cohorts (20s vs. 60s) that were included in this sample and that were similar in age to the participants from the two waves of the main study. This allowed us to test whether short- vs. long-forms of the PTPI showed the same patterns of cross-sectional mean-level change across different cohorts. Furthermore, this allowed us to compare cross-sectional mean-level change across different cohorts (computed from the validation study) with longitudinal mean-level change within the same cohort (computed from the main study), using the same personality measures. Using this validation sample, we also computed mean-level differences between the long- and short-forms of the PTPI, which allowed us to conduct additional robustness checks of the longitudinal mean-level changes observed in the main study, because we could correct for potential mean differences due to the type of scale used. Finally, using the cross-sectional validation sample, we tested measurement invariance across relevant age cohorts (20- vs. 60-year-olds, which we could not do in the main study due to the lack of item-level data). Using the second sample ($N = 38$), which was longitudinal, we obtained the short-term test-retest reliabilities of the long- and short-form PTPI scales over a two-week time span. We used the test-retest estimates obtained from this independent validation study to help us better estimate the rank-order stability of personality across the lifespan in the Project Talent sample (i.e., in the main study). Specifically, in addition to raw correlation coefficients (unadjusted for error) found in the Project Talent sample, we also

obtained correlation coefficients in a latent framework, which are likely to be more accurate, as we accounted for measurement error (Little, 2013). To that effect, we used single-indicator latent constructs with the marker variable convention, and used the observed variance and estimated test-retest reliability of the long- vs. short-form PTPI scales obtained from the validation study.

<div align="center">**Validation Study**</div>

This validation study included two samples, one cross-sectional and one short-term longitudinal. Pozzebon and colleagues (2013) previously used these data to publish results on the long-forms of the Project Talent Personality Inventory (PTPI). We extend those results to encompass the short-forms of the PTPI and validate them against the long-forms.

Method

Because the data were publicly available and de-identified, this study was deemed exempt by the University of Houston, Division of Research Institutional Review Boards (IRB ID: STUDY00000793).

Participants (cross-sectional sample). As reported by Pozzebon and colleagues (2013), 3,934 participants (65% females; 88% European American) were available for data analysis. The mean age was 50 (SD = 19.32) and the sampling was largely focused on young (20s) and older (60s) adults. The goal was to test the validity of the PTPI in samples close to the age of the Project Talent sample at baseline (when the average age was 16) and 50 years later (when the average age was 67).

Participants (short-term longitudinal sample). As reported by Pozzebon and colleagues (2013), a sample of 38 participants (47% females, 87% European American, M_{age} = 31.87, SDage = 13.58) was available for data analysis. These participants completed two assessments collected two weeks apart.

Measures (across both samples). Participants responded to questions about their age, gender, and ethnicity. To measure personality traits, we used the same measure that was administered in Project Talent, namely, the Project Talent Personality Inventory (PTPI). Participants completed all 108 PTPI items, from which the 10 PTPI scales were scored. Item responses were assessed using the same measurement scale used in the Project Talent surveys and scale scores were computed using the same procedures described below in the "Main Study" section. Specifically, item scores were dichotomized and then scale scores were computed taking the average of the relevant items. Notably, the scoring procedures for the PTPI cannot be changed in the Project Talent data because item-level data are not available at baseline; thus, to validate the scales used in the main study, we followed the same scale scoring procedures in this validation study. We computed scale scores for both long- and short-forms of the PTPI.

All the measures, data analysis scripts, and de-identified data used in this validation study, and which are necessary to reproduce the results, are publicly available at the following address: https://osf.io/vxba7/.

Results (cross-sectional sample)

Descriptive statistics and internal consistencies for the long- and short-forms of the PTPI scales can be found in Tables 1 and 2, respectively. When comparing the internal consistencies of the long- vs. short- forms, we can observe that the average decrement in reliability was only .02 for the short-forms. Furthermore, regardless of form type (long vs. short) reliabilities did not change markedly when examined in younger (20s) and older (60s) subgroups. Moreover, we observe that both the long- and the short-forms showed a similar pattern of mean age differences between the 20s and 60s groups, with the 60s age group (compared to the 20s age group) being higher in calmness, mature personality, self-confidence, social sensitivity and tidiness. These

findings were consistent with the maturation hypothesis and previously published age differences in the Big Five personality traits (see Roberts, Walton, & Viechtbauer, 2006).

Table 3 shows the correlations among the 10 PTPI scales in both their long- and short-forms, as well as across the long- and short-forms. Importantly, all correlations between long- and short-form versions of the same scale were higher than .83.

When comparing the means of long- vs. short- forms (across the entire sample), paired samples t-tests showed some statistically significant differences (see Table 4), where in some cases the short-form score was slightly lower (e.g., vigor, impulsiveness, self-confidence, culture, and tidiness), whereas in other cases the short-form score was slightly higher (e.g., mature personality, sociability, and social sensitivity). To address this issue in the Project Talent data, where the long-form was administered at baseline and the short-form was administered at the 50th year follow-up, we conducted a robustness check in the context of the longitudinal mean-level change analysis, where we adjusted the means from the 50th year follow-up by adding or subtracting a constant equal to the mean-difference observed between long- and short-forms in this validation study.

In longitudinal studies, it is important to ascertain that the changes observed in manifest indicators (e.g., personality traits in the present case) are due to real changes in the phenomena being studied and not due to changes in measurement properties across time or across different age groups, that is, one needs to establish measurement invariance (Schmitt & Kuljanin, 2008; Widaman et al., 2010). To test for measurement invariance, however, item-level data must be available at all time points. Unfortunately, in the main longitudinal study presented in this paper (Project Talent sample), we did not have item-level data available at baseline. To alleviate this limitation, we used the cross-sectional data from this validation study to test for measurement

invariance in the PTPI scales across the two relevant age groups: 20- vs. 60-year-olds. We performed the analyses for each PTPI scale separately, both for long- and short-form versions of each scale. Across all scales, we found evidence for configural (similar factor structure in two-group confirmatory factor analysis), metric (equal factor loadings), scalar (equal factor loadings and intercepts), and strict (equal factor loadings, intercepts, and residuals) measurement invariance across the two age groups (the only exception was for the short-form of the self-confidence scale, which showed evidence for metric, but not scalar invariance). The results, along with further details on how to interpret them, can be found in the Supplemental Materials in Tables 4S and 5S. These findings suggested that people across different age groups (age groups that were very similar to the age groups we had in the main longitudinal study) used the PTPI measures in similar ways. This brought some empirical evidence that measurement invariance might also hold in the longitudinal data form the main study, which allowed us to interpret the coefficients resulting from the longitudinal analyses described below as we did (with the caveat that mean-level differences for self-confidence should be treated with caution since the evidence for scalar invariance was limited in this case).

Together, these findings suggest that the short-form versions represent good measures of the original constructs, with the caveat that the scale means are slightly different (see Table 4), but we addressed this issue in the main study by conducting a robustness check using data from this validation study. Furthermore, given the findings for measurement invariance across the two critical age groups in the cross-sectional validation study data, we believe that the results presented in the main longitudinal study (Project Talent sample) can be taken as not being the result of measurement artifacts.

Results (short-term longitudinal sample)

Descriptive statistics can be found in Table 2S of the Supplementary Material. The test-retest reliabilities for the long-form PTPI scales were as follows: Vigor (.81), Calmness (.77), Mature Personality (.79), Impulsiveness (.61), Self-Confidence (.89), Culture (.82), Sociability (.80), Leadership (.76), Social Sensitivity (.85), Tidiness (.88).

The test-retest reliabilities for the short-form PTPI scales were as follows: Vigor (.80), Calmness (.73), Mature Personality (.66), Impulsiveness (.68), Self-Confidence (.84), Culture (.79), Sociability (.78), Leadership (.76), Social Sensitivity (.80), Tidiness (.80).

Thus, the average decrement in test-retest reliability resulting from using short- versus long-form scales was .04, indicating that the short-form versions represent good measures of the original constructs. Furthermore, these results provided us with critical information needed to address measurement error in the rank-order stability analyses from the main study (see Table 2S).[3]

Conclusion

In sum, this validation study provided us with enough information to be confident that the short-forms represent good approximations of the long-forms and likely measure the same constructs. Furthermore, this study provided us with the necessary information to (a) account for measurement error in a latent framework (using the observed test-retest reliabilities and variance of the scales from the validation study to inform measurement error in the main study), and (b) conduct a robustness check of the mean-level change analysis in the main study, by accounting for mean differences observed in the validation study between long- and short-form versions of the PTPI.

Main Study

Method

  Because the data were publicly available and de-identified, this study was deemed exempt by the University of Houston, Division of Research Institutional Review Boards (IRB ID: STUDY00000793).

  Participants. The data came from Project Talent (see Wise et al., 1979), a longitudinal study that started in 1960 with a 5% representative sample of US high-school students.[4] Over 440,000 students in grades 9 through 12, coming from 1,300 schools, participated at the baseline assessment in 1960. Personality measures were available at baseline and at the 50th year follow-up. Thus, we used these two waves of data to test the rank order stability, mean-level change, and profile stability of personality, over the lifespan. Participants in the 50th year follow-up were selected using the following procedures. First, a representative subsample of 4,879 participants was randomly selected from a 10% random subsample of the schools that were originally surveyed in 1960. Next, using a wide variety of tracking methods (see Stone et al., 2014), the project team managed to locate 84.8% of the random subsample: 15.5% were deceased, 50.3% were located with an address and verified, 19% were located with an address and not verified. Survey materials were mailed to the presumably surviving subjects whose address had been identified (i.e., 3,462 people). Of these, about 56% responded to the survey and were included in the final dataset of the 50th year follow-up ($N = 1,952$, out of which 1,858 were coded as "credible," see "Data Analysis" below, and therefore used in our analyses), however, due to missing data on the personality variables, our final longitudinal sample using listwise deletion was $N = 1,795$.

The participant demographics across the two waves used were as follows: (a) the gender distribution was stable across the two time-points, with 52.5% females at baseline and 52% females at the follow-up; (b) the race/ethnicity distribution was also fairly similar across waves with 93% Whites/Caucasians at baseline and 95.3% at the follow-up; and (c) the ages were on average 16 years old at baseline (with participants ranging from 9th to 12th grades) and 67 years old at the follow-up.

Measures.

Project Talent Personality Inventory (PTPI; baseline). The PTPI included 108 items from which ten different scale composites were scored and recorded. All the scale items are available in Table 1S in the Supplementary Materials.

The Vigor scale (7 items) measures the physical activity level of a person. The Calmness scale (9 items) measures the ability to react to emotional situations in an appropriate manner without extreme emotions. The Mature Personality scale (24 items) measures the ability to get work done efficiently and to accept assigned responsibility. The Impulsiveness scale (9 items) measures the tendency to make quick decisions without full consideration of the outcomes. The Self-Confidence scale (12 items) measures one's feelings of social acceptability and the willingness to act and think independently. The Culture scale (10 items) measures the tendency to recognize the value of aesthetic things, and to display refinement and good taste. The Sociability scale (12 items) measures the tendency to enjoy being with people. The Leadership scale (5 items) measures activities such as taking charge and seeking out responsibilities. The Social Sensitivity scale (9 items) measures the propensity to put oneself in another's place. Finally, the Tidiness scale (11 items) measures the desire for order and neatness in one's environment. For each item, participants rated how well the item described them on a 5-point scale ("extremely well" to "not very well"). Item-level data are unfortunately not available to

researchers today for the entire sample (only for 4% of the sample), which is why we relied on the scale scores computed by the Project Talent staff. Furthermore, when computing the PTPI scale scores that are currently available to researchers, the Project Talent staff did not use the original Likert scale coding. Instead, they dichotomized the individual item scores as follows: answers A (*extremely well*) and B (*quite well*) were coded as 1, whereas answers C (*fairly well*), D (*slightly*), and E (*not very well*) were coded as 0; in the case of reverse scored items, answers D and E were coded as 1, whereas answers A, B, and C were coded as 0. After the item-level answers were dichotomized, the Project Talent staff summed them up, to form the 10 PTPI scale scores that are currently available. To make the long-form scale scores used at baseline comparable with the short-form scale scores used at the follow-up, we computed item averages, instead of sums, for each scale.[5]

As mentioned earlier, in previous work on independent but comparable samples (Pozzebon et al., 2013), researchers established the validity and reliability of the 10 (long-form) PTPI scales (see also Table 1 in the present paper), and they identified how the 10 PTPI scales relate to modern Big Five inventories (e.g., John, Donahue, & Kentle, 1991). Thus, Self-Confidence and Calmness were most reflective of Emotional Stability; Sociability, Vigor, and Leadership were most reflective of Extraversion; Culture was best reflective of Openness, Social Sensitivity reflected Agreeableness; and Mature Personality, Impulsiveness (reverse scored), and Tidiness reflected Conscientiousness.

Project Talent Personality Inventory (PTPI; 50th year follow-up). At the 50th year follow-up, the Project Talent staff administered a short-form version of the PTPI. Specifically, the 10 PTPI scales were measured using a subset of 5 of the original items for each scale. All the scale items are available in Table 1S in the Supplementary Materials, and detailed codebooks

and scale construction syntax (note that the scale construction syntax used was the same as the one used in the validation study, as we recoded the item names to match) are publicly available at the following address: https://osf.io/vxba7/. As it was the case at baseline, participants rated how well each item described them on a 5-point scale ("extremely well" to "not very well"). Although item-level data are available at the 50th year follow-up, to make the short-form versions of the PTPI scales as comparable as possible to the long-form versions used at baseline, we followed the same scale computation procedure that was used by the Project Talent staff in 1960. Specifically, we dichotomized and averaged the items. Therefore, all the analyses presented in this paper are based on the dichotomized coding of the items.

Given the similar internal consistencies, test-retest reliabilities, and high inter-correlations between the long- and short-form versions of the PTPI, as presented in the validation study, we concluded that the two measures were highly comparable and therefore testing the rank-order stability, mean-level change, individual-level change of personality traits, and personality profile stability over time using these measures was appropriate.

Data Analysis

Participants were excluded prior to all analyses based on response credibility. Specifically, we only analyzed cases that were coded as "credible" on the original response credibility index (see Wise et al., 1979). This credibility index was computed based on a Screening scale, which included questions such as "How many days are in a week?" that should have been answered easily by anyone who did not suffer from a reading problem, a clerical problem in recording answers, general slowness, or a lack of cooperation. Because the 50th year follow-up was conducted on a representative sub-sample of the original baseline data (i.e., as described earlier, at the 50th year follow-up, the researchers did not attempt to contact everyone who had participated at baseline, but only a representative 10% sub-sample), for the purposes of

this study, the baseline sample consisted of 4,879 participants. Of these, 4,513 cases were coded

as credible. Furthermore, cases were excluded based on having missing data on personality

measures at baseline. Thus, we were left with a total sample of 4,510 participants at baseline.

Due to the longitudinal design of this study and the long time-span covered, we had missing data

at the 50th year follow-up. To better understand how participants who stayed in the study at the

50th year follow-up differed from the participants who dropped out, we conducted an attrition

analysis, which we present in Table 6S of our Supplementary Material. We present the results of

these analyses in the next section. Because participants who stayed in the study differed

systematically from participants who dropped out (see Table 6S), we dealt with missing data in

two different ways. First, in all our subsequent analyses, we analyzed the data using listwise

deletion ($N = 1,795$). Second, we used the full information maximum likelihood (FIML)

approach (maximum likelihood estimation with robust standard errors) to obtain parameter

estimates and standard errors that accounted for the missing data (Enders, 2010; Enders &

Bandalos, 2001). In this approach, all the model covariates were used to predict the missing data,

and the estimation was based on $N = 4,510$, which was the baseline sample relevant for this

study. We present FIML results in Tables 7S and 8S of our Supplementary Material. Notably,

there were no meaningful differences between the results using listwise deletion and those using

FIML estimation.

The main analysis consisted of five parts. First, we assessed the rank-order stability of the

PTPI scales across 50 years. To do so, we obtained a correlation matrix across all 10 PTPI scales

and across both time points. The correlations of interest for testing the rank-order stability of

personality over the lifespan were those between the same scales across time points (e.g., the

correlation between Vigor at baseline and Vigor at the 50th year follow-up, represents the rank-

order stability over 50 years of the trait Vigor). All the PTPI scales scores were based on dichotomized item coding and averaging the relevant items. In addition to the *raw* rank-order stability coefficients, we also obtained rank-order stability coefficients in a structural equation modeling framework, which are likely to be more accurate, as we accounted for measurement error (Little, 2013). To that effect, we used single-indicator latent constructs with the marker variable convention, and used the observed variance and estimated test-retest reliability of the long- and short-form PTPI scales, respectively, obtained from the validation study to account for measurement error (Watson, 2004).

Second, we tested the mean-level change of the 10 PTPI scores across 50 years using paired-samples t-tests. Because the validation study showed slight mean-level differences between long- and short-form versions of the PTPI scales, and because the Project Talent data included long-forms at baseline and short-forms at the 50th year follow-up, we addressed this issue by conducting a robustness check, where we adjusted the means from the 50th year follow-up by adding or subtracting a constant equal to the mean-difference observed between long- and short-forms in the validation study. After adjusting the means, we re-computed longitudinal mean-level change.

Third, we tested individual-level change using the Reliable Change Index (RCI; Christensen & Mendoza, 1986; Jacobson & Truax, 1991). The RCI is calculated for each trait, for each person, separately. $RCI = (X_2 - X_1)/S_{diff}$, where X_2 is a person's score at Time 2, X_1 is a person's score at Time 1, and $S_{diff} = (2(S_E)^2)^{1/2}$, that is, the standard error of difference between the two test scores. The standard error of measurement, $S_E = S_{dev}(1-\alpha_{test-retest})^{1/2}$, where S_{dev} is the standard deviation of the measure and $\alpha_{test-retest}$ is its test-retest reliability. Because the Project Talent data included long-forms at baseline and short-forms at the 50th year follow-up, we

computed two different standard errors of measurement, one for each form, using data from the short-term longitudinal validation study.[6] RCI scores larger than 1.96 or smaller than -1.96 are considered indicative of reliable change. Based on RCI scores, people are split into increasers, decreasers, and nonchangers and then the distribution of people across these three groups is compared, via a chi-square test, with a distribution expected by chance (i.e., 2.5% increasers, 2.5% decreasers, and 95% nonchangers).

Fourth, following Furr (2008) and Klimstra and colleagues (2010), we used q-correlations to assess personality profile stability based on the PTPI self-reports. To assess overall profile stability, we correlated the rank-ordered set of PTPI traits assessed at baseline with the same set assessed at the 50th year follow-up, within each person. To assess distinctive profile stability, we first subtracted average scale scores from the corresponding raw scores for each person at each time point. Next, we computed distinctive stability scores, for each person, by correlating the rank-ordered set of PTPI difference scores at baseline with the rank-ordered set of the same difference scores at the 50th year follow-up. To assess within-time normativeness (a.k.a. profile normativeness) we correlated each person's rank-ordered set of PTPI traits at each time point with the rank-ordered set of sample means on PTPI traits at that same time point, thus rendering two sets of profile normativeness (one for baseline and one for the 50th year follow-up). To assess whether the magnitude of the within-person correlations observed across time was meaningful, we had to evaluate them against the distribution of within-person correlations that could be found in a sample with the same mean and standard deviation, but in which profiles had been randomly paired across time. Thus, we conducted a simulation study following previous recommendations (Robins et al., 2001; De Fruyt et al., 2006). We used the same original data set, but instead of using the correct pairs across time, we randomly assigned follow-up measurement

occasions to baseline measurement occasions. Note that the simulation benchmarks are only relevant for overall profile stability and distinctive stability (i.e., the cross-time analyses).

Fifth, we systematically tested gender differences. To test for gender differences in rank-order stability, we obtained stability coefficients for each trait and for each gender and then conducted Fisher tests to compare the magnitudes of the stability coefficients. To test for gender differences in mean-levels, we conducted independent-samples t-tests to test cross-sectional gender mean-differences at baseline and at the 50th year follow-up, and we used repeated-measures ANOVA to test the interaction between change in personality across time and gender. To test for gender differences in individual-level change, we dummy-coded RCI scores, such that people who showed reliable change in either direction got a 1 and people who did not show a reliable change got a 0. Then we cross-tabulated that with gender (0=male, 1= female) and obtained phi correlations, which indicated whether women or men were more likely to show reliable change on each personality trait across 50 years. To test for gender differences in personality profile stability, we correlated gender with each of the three aspects of personality profile stability (overall profile stability, distinctive profile stability, and profile normativeness).

All the data analysis scripts necessary to reproduce these results are publicly available at the following address: https://osf.io/vxba7/. Furthermore, all the output files can be found at the same address in case the reader is interested in exact p-values for each of our effects, in addition to the effect sizes and 95% confidence intervals that we report in this paper.

Results

Attrition Analysis

In the attrition analyses we tested mean-level differences in the PTPI scales measured at baseline between participants who dropped out and those who stayed in the study at the 50th year follow-up. Results can be found in Table 6S of the Supplementary Material.

The attrition analyses showed that participants who stayed in the study at the follow-up differed slightly on their PTPI scores. Specifically, participants who stayed (vs. those who dropped) were higher in vigor, calmness, and mature personality (Cohen's d = .18). All other attrition effects were not statistically significant and smaller than a Cohen's d of .10.

Rank-order Stability

The raw correlation coefficients for the rank-order stability of the 10 PTPI scales can be found in Table 5. The average rank-order stability over 50 years was .23, ranging from .09 for Impulsiveness to .34 for Culture, with most scales showing rank-order stability above .20. The latent framework rank-order stabilities (corrected for error using test-retest reliabilities and variances of the PTPI scales obtained from the longitudinal validation sample) were as follows: Vigor (.22), Calmness (.32), Mature Personality (.43), Impulsiveness (.13), Self-Confidence (.28), Culture (.41), Sociability (.29), Leadership (.32), Social Sensitivity (.35), Tidiness (.36). Thus, when using the latent framework that accounts for measurement error (as estimated from the validation study and as detailed in Table 2S in the Supplementary Material), the average rank-order stability across PTPI scales over 50 years was .31, with most scales showing rank-order stability above .25.[7] The raw and error-adjusted rank-order stabilities can be found in Figure 1 along with their 95% confidence intervals.

Mean-level Change

Table 6 presents descriptive statistics of each of the 10 PTPI scales at baseline and at the 50ᵗʰ year follow-up, as well as standardized mean-level changes across the 50-year period. For the sake of comparison with previous research, to calculate standardized mean-level changes, we used the same procedure used by Roberts and colleagues (2006), namely the single-group, pretest-posttest raw score effect size (Morris & DeShon, 2002). Specifically, $d = (Mfollow\text{-}up - Mbaseline)/SDbaseline$.

The means for all the scales were reasonably close to the theoretical midpoints (i.e., .50) and the standard deviations were reasonably wide. We can observe that across the board, people increased over time in all the traits, except for Impulsiveness on which they decreased. The average change (in absolute value) was slightly above one half of a standard deviation ($d = .63$), the standardized mean-level change ranging between -.17 for Impulsiveness and 1.64 for Mature Personality. The average mean-level change as well as the direction of change was consistent with previous research on the Big Five personality traits (Roberts et al., 2005). Thus, when people were in their 60s (as opposed to when they were in their teens), they were higher in Calmness and Self-confidence (indicative of higher Emotional Stability), higher in Mature Personality and Tidiness (indicative of higher Conscientiousness), higher in Leadership (indicative of the dominance facet of Extraversion), and higher in Social Sensitivity (indicative of Agreeableness). These standardized mean-level changes found in the Project Talent longitudinal sample are also consistent (albeit larger) with the standardized mean-level changes we observed in our cross-sectional validation study (see Tables 1 and 2), when comparing different age cohorts. The effect directions were very similar, but the average effect size was smaller in the cross-sectional analyses, only about a quarter of a standard deviation ($d = .25$ when

using the long-forms of the PTPI and $d = .21$ when using the short-forms), as opposed to the half

standard deviation observed in the longitudinal sample.

One possibility is that the observed longitudinal mean-level change observed in Project

Talent was inflated by using different scale versions (long-form at baseline and short-form at the

follow-up). To address this issue, we conducted a robustness check. Specifically, we adjusted the

means from the 50th year follow-up by adding or subtracting a constant equal to the mean-

difference observed between long- and short-forms in the validation study (for mean-differences

observed in the validation study, see Table 4). After adjusting the means, we re-computed

longitudinal mean-level change, using the same formula used before, namely the single-group,

pretest-posttest raw score effect size (Morris & DeShon, 2002). As we can see in Table 6, the

mean-level changes that included the robustness check were very similar to those without the

robustness check. Some effects increased slightly while others decreased slightly, but the average

change observed (in absolute value) was exactly the same, that is, about half a standard deviation

($d = .63$). The pattern of change observed was consistent, with two exceptions. Specifically, the

Impulsiveness effect changed from a small negative to a non-statistically significant effect, and

the Sociability effect decreased from about a third of a standard deviation to close to zero. The

new Sociability effect is more consistent with past research, which does not predict that people

should increase in Sociability as they grow older. In sum, the robustness-check suggests our

results hold up and are consistent with past research, that is, personality changes across the

lifespan in the maturational direction. Furthermore, this is the first study to test mean level

changes in personality in a longitudinal setting across 50 years.

Individual-level Change

Table 7 shows the percentages of people, for each trait, who either increased, decreased, or stayed the same according to the Reliable Change Index across the 50-years between the two assessments. Of the ten personality traits assessed, Mature Personality exhibited the highest level of reliable change, with 60.7% of the people in the sample showing change, mostly increases (58.7%). The lowest level of reliable change was found for Leadership, where 21.4% of the people in the sample showed reliable change, mostly increases again (17.3%). On average, on any given trait, about 40% of the people in the sample showed reliable change, whereas 60% showed no reliable change.

As expected, the percentages of people increasing or decreasing on any given trait corresponded to the mean-level changes presented in Table 6 and followed a maturation pattern, with highest percentages of "increasers" (vs. "decreasers") being found for traits indicative of Conscientiousness (Mature Personality, Tidiness), Emotional Stability (Calmness, Self-confidence), Agreeableness (Social Sensitivity), and dominance facets of Extraversion (Leadership). Furthermore, as seen in Table 7, chi-square tests showed that, for each of the ten traits, the pattern of increasers, nonchangers, and decreasers differed significantly from an expected random-change pattern where 95% of people would show no change, 2.5% would show increases, and 2.5% would show decreases. Thus, there appears to be reliable change in each of the ten personality traits. Looking across traits, according to the RCI, 97.9% of the people showed reliable change on at least one of the ten personality traits assessed across the 50-year period, 58.9% of the people showed reliable change on four or more traits, and only .2% of people showed reliable change on all ten traits.

In sum, across 50 years, there was evidence of reliable change for every single trait assessed, patterns of individual-level change followed mean-level maturational patterns, almost everyone showed reliable change on at least one trait, more than half of the people showed reliable change on four or more traits, but very few people changed reliable on all ten traits.

Personality Profile Stability

Overall profile stability, which represents the similarity between a person's trait profile at one time and their trait profile at a later time (e.g., Ozer & Gjerde, 1989), over a 50-year time span, ranged from -.69 to .98, with a mean of .37 (SD = .31) and a median of .40. Interestingly, this is the same overall profile stability estimate that Lönnqvist and colleagues (2008) found across 15 years (going from age 20 to age 35). Furthermore, our overall profile stability estimate was well above the corresponding estimate produced in our simulation study, where profiles were matched randomly across time (M = .25, SD = .33, Mdn =.28), suggesting that overall profile stability across 50-years was higher than chance (Mdiff = .12, t(3588) = 11.23, p < .001, 95% CI Mdiff [.10; .14], Cohen's d = .38). Notably, the overall profile stability estimates produced in our simulation were very similar to estimates produced in previous simulations; for example, Robins and colleagues (2001) found an average value of .20 in their simulation; the somewhat large overall profile stability estimates in simulations where data were randomly matched across time are likely due to what Furr (2008) called the normativeness problem, which is why we assessed, in addition to overall profile stability, two additional estimates of profile stability.

Distinctive profile stability, which represents the similarity between a person's distinctive trait profile at one time and their distinctive trait profile at a later time (Furr, 2008), over a 50-year time span, ranged from -.80 to .97, with a mean of .17 (SD = .35) and a median of .20.

These estimates were well above the corresponding estimates produced in the simulation ($M =$ .01, $SD = .35$, $Mdn =$.01), suggesting that distinctive profile stability across 50-years was higher than chance (Mdiff = .16, $t(3588) = 13.7$, $p < .001$, 95% CI Mdiff [.14; .18], Cohen's $d = .46$).

Regarding within-time normativeness or profile normativeness, which represents the degree to which a person's profile is similar to the average profile within each developmental period (Furr, 2008), we had two sets of estimates. During adolescence (at baseline), within-time normativeness ranged from -.76 to .98, with a mean of .51 ($SD = .28$) and a median of .56. Within-time normativeness during older adulthood (at the 50th year follow-up) ranged from -.71 to .97, with a mean of .62 ($SD = .24$) and a median of .69. These estimates are in line with previous findings (e.g., Bleidorn et al., 2012; Klimstra et al., 2010) and suggest that most people have personality profiles that are normative, regardless of age. To better understand what profile normativeness might mean across time, we also investigated generalized normative stability, which reflects the degree to which two age periods (adolescence and older adulthood in our case) have similar normative trait profiles. Figure 2 shows the normative trait profiles that we found in 1960 adolescents (average age 16) versus older adults (50 years later). Based on these profiles, generalized normative stability was .77 (note that this is a single estimate, not the mean of a distribution).

Gender Differences in Personality Stability and Change

We tested gender differences across all four kinds of stability and change. We started with cross-sectional gender differences, for each time point (baseline and 50th year follow-up). As seen in Table 8, at baseline (average age 16), women scored higher than men in Mature Personality and Tidiness (indicative of Conscientiousness), Sociability (Extraversion), Culture (Openness), and Social Sensitivity (Agreeableness). These same effects replicated at the follow-

up, and additionally, women scored lower than men in Self-confidence (indicative of lower

Emotional Stability).

Regarding the rank-order stability of personality traits across 50 years, we found that the

average stability was .21 for women and .20 for men, and none of the stability coefficients

differed by gender at $p < .001$.

Regarding mean-level change, men and women showed similar patterns of change, and

only two gender-by-time interactions were statistically significant at $p < .001$, specifically, men

(vs. women) increased more in self-confidence across 50 years ($d = .86$ vs. $d = .50$, respectively),

whereas women (vs. men) increased more in social sensitivity as they aged ($d = 1.24$ vs. $d =$

1.04, respectively).

Regarding individual-level change, we found no statistically significant gender

differences in reliable change.

The correlation between gender and personality profile stability was .16 (p < .001),

indicating that women showed higher levels of overall profile stability. However, this was likely

because women showed higher levels of profile normativeness within each time point (the

correlation between gender and profile normativeness was .12, $p < .001$, at baseline; and .19, $p <$

.001 at the 50th year follow-up). Indeed, there were no gender differences in distinctive profile

stability across the 50 years ($r = .04$, $p = .087$).

Across the board, these results are largely consistent with past research, suggesting that,

although there are cross-sectional gender differences in personality, men and women do not

differ much in their patterns of personality stability and change across the lifespan.

Discussion

In the present paper, we used a large US sample that assessed people's personality during adolescence, as well as 50 years later. We tested rank-order stability, mean-level change, individual-level change, personality profile stability, gender differences in stability and change, and we also used two independent samples to validate the short-forms of the PTPI against the long-forms.

We found that the average rank-order stability across the 10 personality traits and across 50 years was .31 (when accounting for measurement error in a latent framework) and .23 (when estimated without correcting for unreliability). The only previous study inconsistent with our finding is the recent study by Harris and colleagues (2016), but this study may have underestimated the stability of personality traits over the lifespan due to their use of one item-measures and different data sources at the different time points (teacher at time 1 vs. self or close other at time 2). Furthermore, our present results, where the average raw stability coefficient was .23, with an average 95% confidence interval of [.19; .27] (see Fig. 1) are consistent with the full developmental model put forward by Fraley and Roberts (2005), where all three developmental processes (stochastic-contextual, person–environment transactions, and developmental constancies) are present.

Why would personality traits be consistent from age 16 through age 66? Broadly speaking authors have argued for environmental factors, genetics, or both when attempting to explain why individuals would remain consistent over long periods of time (Fraley & Roberts, 2005). What environments or environmental factors could promote consistency over a 50-year period stretching from adolescence to old age? It is difficult to imagine that many people in their 60s would find themselves in similar environments to those occupied in their teens. Thus, except

for people who remained in their parents' home or nearby in the same community, strictly

environmental consistency would be an unlikely explanation. Alternatively, one could imagine

that people could play similar roles in adolescence and old age and that the continuity in the role

may help explain the continuity in personality (Roberts, 2007). If a person played the clown in

high school, or was the leader or nurturer, it would not be out of the question that they could play

the same role in their 60s. The other factor thought to contribute to personality continuity is

genetics. One argument would be that some temperamental factor that resulted from genetic

differences at conception would play out as a permanent signal in one's personality over time.

The extreme version of this argument, that people are "hard wired from birth" to possess a

specific personality, is difficult to support given the incredibly small relation between early

childhood temperament and personality traits in adulthood (Caspi & Silva, 1995). It would be

impossible for a signal of such small magnitude (correlations at or below .10) to result in the .20

correlation that we find on average from adolescence to old age. However, newer perspectives on

genetics that focus on developmental genetics may provide a potential answer. Specifically, a

sociogenomic perspective on personality trait development (Roberts, in press) argues that

experience acts on gene systems during development and that this process results in fixed

phenotypes that emerge through development (referred to as "pliable" systems). Thus,

experience modifies the genome through epigenetic mechanisms to help create a phenotype that

reflects both fixed and dynamic genetic processes that may then create a consistent signal in

personality from adolescence to adulthood. Of course, both environmental and genetic

explanations for continuity in personality across the life course are necessarily speculative

because no longitudinal study has tracked yet, genes, epigenetic systems, nor experience over the

time-line of this study.

We found that the average mean-level change was about half of a standard deviation across the 10 personality traits, which is also consistent with past research. The pattern of change was consistent with the maturity principle, where most traits are assumed to increase in an adaptive manner (e.g., higher conscientiousness, agreeableness, emotional stability, and the dominance facet of extraversion) (Roberts et al., 2006). The standardized mean-level changes found in our Project Talent longitudinal sample were also consistent with the standardized mean-level changes we observed in our cross-sectional validation study, though the effects were twice the size. In sum, mean-level change in personality traits over 50 years was, on average, slightly over a half of a standard deviation. For several scales, such as the maturity scale, the amount of change was over one standard deviation, which is substantial. In fact, the changes in the scales that reflect the largest changes estimated from prior work (e.g., agreeableness, conscientiousness, and emotional stability) all showed changes around 1 to 1.5 standard deviations which are quite large by psychological standards. The present findings that personality trait change across 50 years was (a) larger than change across shorter periods of time (e.g., 10 years, where effects were about a quarter to a third of a standard deviation), and (b) consistent with estimates of cumulative change across the lifespan extrapolated from meta-analytic findings (Roberts et al., 2006), suggest that the plasticity model, where change continues and cumulates across the lifespan, might be more fitting than a "set point" model that would argue for a braking system on change.

The mean-level changes also beg the question of why individuals would show such dramatic shifts with age. Like explanations for rank-order stability, the explanations for mean-level change have focused on genetic and environmental factors. While longitudinal twin studies point to the typical finding that both are important for development (Bleidorn et al., 2009), as of yet, there are no insights into how genetics would contribute to these changes. It is a possibility

that personality trait development works like puberty in that it is pre-programed and universal,

but to date no molecular genetic evidence supports that claim, and the present findings on

individual-level change bring further evidence against this claim (that is, change is unlikely to be

universal if there are individual-differences in change). Some of the environmental factors

thought to contribute to personality development include normative transitions to adulthood

(Roberts & Damian, in press). These experiences include applying oneself to achievement

situations, such as school and work (Bleidorn et al., 2013; Hudson, Roberts, & Lodi-Smith,

2012) and some aspects of relationships, such as relationship duration and satisfaction (Lehnart,

Neyer, & Eccles, 2010). Unfortunately, because there was such a long-time span between the

assessments of the Project Talent sample, we do not have the experiential data to test these ideas.

The maturational patterns we found are also consistent with the classic account of Erik Erikson

(1963), who postulated that people mature as they age, and change continuously throughout the

lifespan, pressed to adapt by the ever-increasing social demands and developmental tasks

required by each life stage.

Regarding individual-level change, we found that, on average, on any given trait, about

40% of the people in the sample showed reliable change, whereas 60% did not. This is higher

than the 30-10% estimates of people who showed reliable change across shorter time spans (e.g.,

Robins et al., 2001; Roberts et al., 2001), which suggests that greater change accumulates with

time. Looking across traits, we found that 97.9% of the people showed reliable change on one or

more of the 10 personality traits assessed across the 50-year period (compared to 84% over 8

years in the study by Roberts and colleagues, 2001), again suggesting that more personality

change may accumulate with time, thus causing more individuals to show reliable change on

more traits. Furthermore, like mean-level changes, individual-level change patterns supported the maturation hypothesis.

The fact that most people showed reliable change in one or more personality traits supports the perspective that individual differences in change are an important developmental phenomenon. Not everyone changes in the same way despite normative trends. Some people change less than their peers, while others change more than the norm. Our study is consistent with past research identifying the existence of individual differences in change in shorter longitudinal studies (e.g., De Fruyt et al., 2006; Mõttus, Johnson, & Deary, 2012; Robins, et al, 2001; Roberts, Caspi, & Moffitt, 2001; Schwaba & Bleidorn, 2017). It makes sense that with a longer time span, more people showed unique patterns of change. These non-normative patterns of change beg the question of why? In many other studies, this question has been answered by showing that life experiences correlate with individual differences in change (Bleidorn, 2012; Göllner et al., 2017; Lehnart, Neyer, & Eccles, 2010; Takahashi, et al., 2013). Unfortunately, the unique feature of the present study—the 50-year time lag, also prevented the tracking of life experiences that could have shown relations to these individual differences in change in the present case. Future research would, optimally, track personality and life experiences over a 50-year period, with multiple assessments, to further establish evidence for the links between life experiences and personality trait development in adulthood.

Regarding personality profile stability, we found that overall profile stability across 50 years was .37, which was consistent with previous findings of profile stability over 15 years (e.g., Lönnqvist et al., 2008). As expected, we found distinctive profile stability to be lower, with a mean of .17. Importantly, both these estimates were well above estimates of stability from the simulation study that we ran, indicating that profile stability across 50 years (overall and

distinctive) was higher than chance. Profile normativeness was high at each age period and it
increased as people aged (.51 and .62, during adolescence and older age, respectively), which
was also consistent with previous findings that looked at profile normativeness across much
shorter time spans (e.g., Bleidorn et al., 2012; Klimstra et al., 2010).[8] Interestingly, normative
personality profiles were very similar between 1960 adolescents and the same sample assessed
50 years later correlating .77 (see Fig. 2). To better understand how normative personality
profiles might have changed (or not changed) across historical cohorts, we conducted additional
exploratory analyses using the cross-sectional data used from the validation study, for 20- and
60-year-olds separately (see these additional normative trait profiles in Figure 2). Notably, the
correlation between the normative trait profile of 16-year-olds in 1960 and 20-year-olds in 2013
was .67. This estimate is comparable to previously reported generalized normative stability
estimates (.64-.78), where 12- and 20-year-olds' normative profiles were compared (Klimstra et
al., 2012; note that in this latter study, youth cohorts were assessed within a 5-year period of each
other, unlike across our two studies, where the gap was 53 years). All in all, the relatively high
degree of consistency in personality norms, both across US history and across time, suggests that
certain personality profiles may be normative across time because they might facilitate
adjustment (e.g., it might always be advantageous to be more sociable than you are impulsive,
regardless of age and historical time).

Why would personality profiles be stable across 50 years and why would profile
normativeness increase with age? As with rank-order stability, the answer to the first question
probably lies with a combination of genetic and environmental factors, whereby profile stability
reflects fixed phenotypes that emerge through development (see Bleidorn et al., 2012; Roberts, in
press). Regarding the second question, researchers have suggested that profile normativeness

might increase across time because it reflects psychological adjustment (Bleidorn et al., 2012; Klimstra et al., 2010). Presumably, a higher level of psychological adjustment, might allow people to select environments that fit them better, thus exposing themselves to less environmental pressure to change (see also Donnellan et al., 2007), which would in turn lead to increasing stability and normativeness across time.

Although our study has many advantages, including its large sample and the inclusion of personality measures across a 50-year time span, our study also has several limitations. First, item-level data were not available at baseline, which means that we were not able to conduct more complex statistical analyses, such as structural equation modeling, which would have allowed us to test for invariance over time in the measurement model and use latent constructs. To alleviate this limitation, we used cross-sectional data from the validation study to test for measurement invariance in the PTPI scales across the two relevant age groups: 20- vs. 60-year-olds. Across all scales (except for the short-form of self-confidence), we found evidence for configural, metric, scalar, and strict invariance, which gave us more confidence that we may interpret the longitudinal findings as we did (though the mean-level changes in self-confidence should be interpreted with caution since we only had evidence for metric but not scalar invariance for its short-form version). Second, as the attrition analyses presented in Tables 6S of the supplementary materials showed, the sample available at the follow-up differed slightly from the sample that dropped out. Specifically, the people who stayed in the study at the 50th year follow-up (as opposed to those who dropped out) were higher in vigor, calmness, and mature personality at baseline (i.e., during adolescence), which means our sample was no longer representative of the US high-school population in 1960. It is possible that this impacted our estimates of personality stability and change, though it is difficult to assess in what way. Another

limitation is that the Project Talent staff used long forms of the personality measures at baseline, but short forms at the follow-up. We tried to alleviate this shortcoming by conducting a validation study on two independent but comparable samples to obtain internal and test-retest reliabilities for both the long and the short forms (see validation study), as well as investigate their mean-level differences (which we adjusted for in the robustness check) and cross-sectional correlations (which were in the order of .90 or higher, indicating that the short- and long-forms likely measure the same constructs). Although we did everything we could to address the issue of long- versus short-forms at the two time-points, it is still possible that our estimates of change are biased. Furthermore, it is possible that the robustness check employed did not correct for this issue because our robustness check made one crucial assumption, namely that differences between long- and short-forms of the PTPI would be constant across samples and across historical time. This is a rather strict assumption that is likely to be violated, so our mean-level change estimates should be treated with caution. However, the fact that the effects found are consistent with past research and with the cross-sectional results gives us some confidence. Finally, another limitation is that the present study only assessed personality at two time-points, which prevented us from delving more deeply into developmental patterns and processes. For instance, by having such a long time-span in between the two assessments with no other assessments in between, we were not able to capture when changes occurred or whether some changes occurred that were later reversed. We were also limited in the kinds of models we could fit to the data. With three or more assessments we might have been able to fit growth curve models to better understand developmental trajectories.

Despite these limitations, the present study advances our understanding of personality stability and change, because it examines rank-order stability, mean-level change, individual-

level change, and profile stability in a large sample and over a very long timespan (50 years). We found evidence for continuity in the way individuals ranked relative to each other, but there was also a great deal of change in mean-levels of specific traits, a large percentage of individuals showed reliable change, and their individual patterns of change varied, though overall there was evidence for maturation. Furthermore, patterns of personality also showed some degree of stability, within individuals across the lifespan, and this stability was explained by both distinctive stability and within-time normativeness. Together, these results suggest that, although individuals maintain some of their core personality across the lifespan, they also change. Moreover, that change in personality across the lifespan is likely cumulative, following a maturational adaptive pattern, and that change is not uniform or universal, meaning that environmental factors, such as life experiences, are likely to contribute to change, thus, resulting in individual differences in change. These results are in line with previous research, supporting developmental models that include a developmental constancy (e.g., genetic component) and personality stability over the lifespan (Fraley & Roberts, 2005; Roberts, in press). These results are also in line with the maturity principle, whereby people adapt and mature as they grow older so they can successfully engage in life's responsibilities, and they support a plasticity (vs. a "set point" model) of personality change (Roberts et al., 2006; Fraley, 2002). The profile stability results are also in line with genetic models, with the cumulative continuity principle, and with the social investment principle, whereby people's personality profiles may increase in normativeness across the lifespan as they invest in normative social roles (see Bleidorn et al., 2012; Klimstra et al., 2010).

Future research should include more long-term studies, over long periods of time, that assess personality at multiple waves, to enable scientists to investigate more deeply developmental process as well as look at inter-individual variation in stability and change.

In sum, there is evidence for stability in personality traits across the lifespan, while at the same time there is evidence of change presumably resulting from life's trials and tribulations. Going back to the love story we presented at the beginning, an interesting question raised by these findings is how likely are people to stay within the same personality "zones" (e.g., above or below the median) after a 50-year long hiatus? Looking at the variable-centered approach, where raw rank-order stability was .23 on average, and using the Binomial Effect Size Display (BESD; Rosenthal & Rubin, 1982) to transform it, we could say that Jerzy and Cyla would each have about 60:40 odds of having stayed in the same personality "zone," on each trait, during their long separation. But people are not characterized by single traits. Thus, looking at the whole person, where overall profile stability was .37 on average, and using BESD to transform it, we could say that Jerzy and Cyla would each have about 70:30 odds of having stayed in the same personality "zone" after their long separation. Given our findings in mean-level changes, individual-level changes, and profile normativeness, we can also assume that both Jerzy and Cyla would have been likely to be more mature and more normative in their personality profiles when they met later in life. These being said, perhaps it makes more sense now how Jerzy and Cyla could fall in love all over again after 40 years of being apart. Although life's experiences had undoubtedly changed them (likely for the better), somewhere deep down they each managed to find in the other a glimmer of the person they had fallen in love with many years before.

References

Allemand, M., Zimprich, D., & Hertzog, C. (2007). Cross-sectional age differences and

longitudinal age changes of personality in middle adulthood and old age. *Journal of

Personality, 75*, 323-358.

Allport, G. W. (1954). *Personality: A psychological interpretation*. New York: Henry Holt.

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*, 766-81.

Biesanz, J. C., & West, S. G. (2000). Personality coherence: Moderating self-other profile

agreement and profile consensus. *Journal of Personality and Social Psychology*, *79*, 425–

437.

Bleidorn, W. (2012). Hitting the road to adulthood: Short-term personality development during a

major life transition. *Personality and social psychology bulletin*, *38*(12), 1594-1608.

Bleidorn, W., Hopwood, C. J., & Lucas, R. E. (2018). Life events and personality trait change.

Journal of Personality, 86, 83-96.

Bleidorn, W., Kandler, C., Riemann, R., Angleitner, A., & Spinath, F. M. (2009). Patterns and

sources of adult personality development: Growth curve analyses of the NEO PI-R scales

in a longitudinal twin study. *Journal of personality and social psychology*, *97*(1), 142.

Bleidorn, W., Kandler, C., Riemann, R., Angleitner, A., & Spinath, F. M. (2012). Genetic and

environmental influences on personality profile stability: Unraveling the normativeness

problem. *Journal of personality*, *80*(4), 1029-1060.

Bleidorn, W., Klimstra, T. A., Denissen, J. J., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2013).

Personality maturation around the world: A cross-cultural examination of social-

investment theory. *Psychological Science*, *24*(12), 2530-2540.

Block, J. (1971). *Lives through time.* Berkeley, CA: Bancroft Books.

Caspi, A., & Herbener, E. S. (1990). Continuity and change: Assortative marriage and the

consistency of personality in adulthood. *Journal of Personality and Social Psychology*,

58, 250–258.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and

change. *Annu. Rev. Psychol.*, *56*, 453-484.

Caspi, A., & Silva, P. A. (1995). Temperamental qualities at age three predict personality traits

in young adulthood: Longitudinal evidence from a birth cohort. *Child

development*, *66*(2), 486-498.

Cattell, R.B. (1981). *Sixteen Personality Factor Questionnaire (16PF)*. Psykologien Kustannus

OY, Helsinki.

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

measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Costa Jr, P. T., Herbst, J. H., McCrae, R. R., & Siegler, I. C. (2000). Personality at midlife:

Stability, intrinsic maturation, and response to life events. *Assessment*, *7*(4), 365-378.

Costa, P. T., & McCrae, R. R. (1988). Personality in adulthood: a six-year longitudinal study of

self-reports and spouse ratings on the NEO Personality Inventory. *Journal of personality

and social psychology*, *54*(5), 853.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and

NEO Five-Factor Inventory: Professional manual*. Odessa, FL: Psychological

Assessment Resources.

Costa, P. T., & McCrae, R. R. (2008). The revised neo personality inventory (neo-pi-r). *The

SAGE handbook of personality theory and assessment, *2*, 179-198.

Damian, R. I., & Roberts, B. W. (2015). The associations of birth order with personality and intelligence in a representative sample of U.S. high school students. *Journal of Research in Personality, 58*, 96-105.

Damian, R. I., Spengler, M., & Roberts, B. W. (2017). Whose job will be taken over by a computer? The role of personality in predicting job computerizability over the lifespan. *European Journal of Personality, 31*, 291–310.

Damian, R. I., Su, R., Shanahan, M., Trautwein, U., Roberts, B. (2015). Can personality traits and intelligence compensate for background disadvantage? Predicting status attainment in adulthood. *Journal of Personality and Social Psychology, 109*, 473-89.

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*, 538–552.

Donnellan, M. B., Conger, R. D., & Burzette, R. G. (2007). Personality development from late adolescence to young adulthood: Differential stability, normative maturity, and evidence for the maturity-stability hypothesis. *Journal of personality*, *75*(2), 237-264.

Edmonds, G. W., Goldberg, L. R., Hampson, S. E., & Barckley, M. (2013). Personality stability from childhood to midlife: relating teachers' assessments in elementary school to observer-and self-ratings 40years later. *Journal of research in personality*, *47*(5), 505-513.

Ferguson, C. J. (2010). A meta-analysis of normal and disordered personality across the life span. *Journal of Personality and Social Psychology, 98*, 659-67.

Fraley, R. C. (2002). Attachment stability from infancy to adulthood: Meta-analysis and dynamic

modeling of developmental mechanisms. *Personality and Social Psychology Review, 6*,

123-151.

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.

Furr, M. R. (2008). A framework for profile similarity: Integrating similarity, normativeness, and

distinctiveness. *Journal of Personality*, *76*, 1267–1316.

Göllner, R., Damian, R. I., Rose, N., Spengler, M., Trautwein, U., Nagengast, B., & Roberts, B.

W. (2017). Is doing your homework associated with becoming more conscientious?

Journal of Research in Personality, 71, 1-12.

Hampson, S. E., & Goldberg, L. R. (2006). A first large cohort study of personality trait stability

over the 40 years between elementary school and midlife. *Journal of personality and

social psychology*, *91*(4), 763.

Harms, P.D., Roberts, B.W., & Winter, D. (2006). Becoming the Harvard man: Person-

environment fit, personality development, and academic success. *Personality and Social

Psychology Bulletin, 32*, 851-865.

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.

Hevesi, D. (2011). Jerzy Bielecki dies at 90; Fell in love in a Nazi camp. *The New York Times.*

Retrieved from http://www.nytimes.com/2011/10/24/world/europe/jerzy-bielecki-dies-at-

90-fell-in-love-in-a-nazi-camp.html?_r=0

Hoekstra, H. A., Ormel, J., & De Fruyt, F. (1996). *NEO persoonlijkheids vragenlijsten: NEO-PI-

R, NEO-FFI. Handleiding [NEO personality questionnaires: NEO-PI-R and NEO-FFI-*

manual]. Lisse, The Netherlands: Swets & Zeitlinger.

Hudson, N.W., Roberts, B.W., & Lodi-Smith, J. (2012). Personality trait development and
     social investment at work. *Journal of Research in Personality, 46,* 334-344.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory—versions 4a and
     54.*

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait
     taxonomy. *Handbook of personality: Theory and research*, *3*, 114-158

Klimstra, T. A., Hale, W. W., III, Raaijmakers, Q. A. W., Branje, S. J. T., & Meeus, W. H. J.
     (2009). Maturation of personality in adolescence. *Journal of Personality and Social
     Psychology*, *96*, 898–912.

Klimstra, T. A., Luyckx, K., Hale, W. W., III, Goossens, L., & Meeus, W. H. J. (2010).
     Longitudinal associations between personality profile stability and adjustment in college
     students: Distinguishing among overall stability, distinctive stability, and within-time
     normativeness. *Journal of Personality*, *78*, 1163–1184.

Lehnart, J., Neyer, F. J., & Eccles, J. (2010). Long-term effects of social investment: The case of
     partnering in young adulthood. *Journal of personality*, *78*(2), 639-670.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Lodi-Smith, J., & Roberts, B. W. (2007). Social investment and personality: A meta-analysis of
     the relationship of personality traits to investment in work, family, religion, and
     volunteerism. *Personality and Social Psychology Review*, *11*(1), 68-86

Lönnqvist, J.-E., Mäkinen, S., Paunonen, S. V., Henriksson, M., & Verkasalo, M. (2008).
     Psychosocial functioning in young men predicts their personality stability over 15 years.
     Journal of Research in Personality, *42*, 599–621.

Lüdtke, O., Roberts, B. W., Trautwein, U., & Nagy, G. (2011). A random walk down university

avenue: life paths, life events, and personality trait change at the transition to university

life. *Journal of personality and social psychology*, *101*(3), 620.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and

determination of sample size for covariance structure modeling. *Psychological Methods,

1*, 130.

Major, J. T., Johnson, W., & Deary, I. J. (2014). Linear and nonlinear associations between

general intelligence and personality in Project Talent. Journal of personality and social

psychology, 106, 638.

McCrae, R. R. (2008). A note on some measures of profile agreement. *Journal of Personality

Assessment*, *90*, 105–109.

McCrae, R. R., & Costa, P. T. (1994). The stability of personality: Observations and

evaluations. *Current directions in psychological science*, *3*(6), 173-175.

McCrae, R. R., & Costa, P. T. (2008). Empirical and theoretical status of the five-factor model of

personality traits. *The SAGE handbook of personality theory and assessment, 1*, 273-294.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with

repeated measures and independent-groups designs. *Psychological methods*, *7*(1), 105.

Mõttus, R., Johnson, W., & Deary, I. J. (2012). Personality traits in old age: measurement and

rank-order stability and some mean-level change. *Psychology and aging, 27(1)*, 243.

Neyer, F. J., & Lehnart, J. (2007). Relationships matter in personality development: Evidence

from an 8-year longitudinal study across young adulthood. *Journal of personality*, *75*(3),

535-568.

Ostendorf, F., & Angleitner, A. (2004). *NEO Persönlichkeitsinventar nach Costa und McCrae:*

Revidierte Fassung [NEO Personality Inventory by Costa and McCrae: Revised edition]. Göttingen, Germany: Hogrefe.

Ozer, D. J., & Gjerde, P. F. (1989). Patterns of personality consistency and change from childhood through adolescence. *Journal of Personality*, *57*, 483–507.

Pozzebon, J., Damian, R. I., Hill, P. L., Lin, Y., Lapham, S., & Roberts, B. W. (2013). Establishing the validity and reliability of the Project Talent Personality Inventory. *Frontiers in Psychology—Personality Science and Individual Differences, 4*, 968.

Roberts, B. W. (2007). Contextualizing personality psychology. *Journal of Personality, 75*, 1071-1081.

Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of research in personality*, *43*(2), 137-145.

Roberts, B.W. (in press). A revised sociogenomic model of personality traits. *Journal of Personality*.

Roberts, B. W., Caspi, A., & Moffitt, T. E. (2001). The kids are alright: Growth and stability in personality development from adolescence to adulthood. *Journal of Personality and Social Psychology*, *81*, 670–683.

Roberts, B. W., & Damian, R. I. (in press). The principles of personality trait development and their relation to psychopathology. In D. Lynam & D. Samuel (Eds.), *Using basic personality research to inform the personality disorders.* Oxford, UK: Oxford University Press.

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.

Roberts, B.W., & Robins, R. W. (2004). A longitudinal study of person-environment fit and

    personality development. *Journal of Personality, 72*, 89-110.

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.

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 edition, Ch 14, pp. 375-398). New York, NY: Guilford.

Roberts, B. W., Wood, D., & Smith, J. L. (2005). Evaluating five factor theory and social

    investment perspectives on personality trait development. *Journal of Research in*

    Personality, *39*(1), 166-184.

Robins, R. W., Fraley, R. C., Roberts, B. W., & Trzesniewski, K. H. (2001). A longitudinal study

    of personality change in young adulthood. *Journal of Personality*, *69*, 617–640.

Shiner, R. L., Allen, T. A., & Masten, A. S. (2017). Adversity in adolescence predicts

    personality trait change from childhood to adulthood. *Journal of Research in*

    Personality, *67*, 171-182.

Stone, C., Scott, L., Battle, D., & Maher, P. (2014). Locating longitudinal respondents after a

    50-Year hiatus. *Journal of Official Statistics, 30*, 311-334.

Takahashi, Y., Edmonds, G.E., Jackson, J.J., & Roberts, B.W. (2013). Longitudinal correlated

    changes in conscientiousness, preventative health-related behaviors, and self-perceived

    physical health. *Journal of Personality, 81*, 417-427.

Terracciano, A., Costa Jr, P. T., & McCrae, R. R. (2006). Personality plasticity after age

    30. *Personality and Social Psychology Bulletin*, *32*(8), 999-1009.

Van Aken, M. A., Denissen, J. J., Branje, S. J., Dubas, J. S., & Goossens, L. (2006). Midlife concerns and short-term personality change in middle adulthood.*European Journal of Personality*, *20*(6), 497-513.

Wise, L., McLaughlin, D., & Steel, L. (1979). *The Project TALENT Data Bank Handbook, Revised*. Palo Alto, CA: American Institutes for Research.

Wagner, J., Becker, M., Lüdtke, O., & Trautwein, U. (2015). The first partnership experience and personality development: A propensity score matching study in *young adulthood. Social Psychological and Personality Science, 6*, 455-463.

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality, 38*, 319–350.

Footnotes

[1]Notably, *q*-correlations reflect only the degree of similarity in the *shape* of two personality profiles across time, leaving out information regarding the elevation (i.e., the mean of the different profile elements) and scatter (i.e., the variability around the profile's elevation) of the profiles (see Biesanz & West, 2000). Some studies (e.g., Donnelan et al., 2007; Robins et al., 2001) assessed profile stability with a related index (D^2), in addition to *q*-correlations, but these studies suggested that both indices led to the same conclusions. Furthermore, McCrae (2008) showed that *q*-correlations are reliable measures of profile stability.

[2]The items of the original scales at baseline were not recorded so we were unable to create duplicate scales at both time points.

[3]Using the short-term longitudinal validation study data, another way to compute the residual errors used to account for measurement error in the latent models from the "Main Study," is to use correlations between the long-forms at Time 1 and the short-forms at Time 2 instead of test-retest reliabilities for each form, in order to simultaneously take into account unreliability due to (a) re-testing and (b) switching from long- to short-forms. More details on how these residuals were computed can be found in Table 3S in the Supplementary Materials. Corrected stability coefficients using these alternative residual errors can be found in Footnote 7 and all syntax and output can be found on our OSF page in the file titled "Main Study Syntax."

[4]Although several previous papers have been published using the personality data available at baseline in Project Talent (Damian & Roberts, 2015; Damian et al., 2015; 2017; Major et al., 2015; Spengler et al., 2018), to our knowledge, no previous papers have used the personality data available at the 50th year follow-up, and thus, no previous papers have analyzed stability and change in personality across 50 years, using this data set, which is the topic of our

current submission. A comprehensive list of papers published using *other* variables from the

Project Talent dataset can be found at the following link:

http://www.projecttalent.org/about/biblio.

[5]To test for mean-level changes in personality across time, we could not use the summed

scale scores because the PTPI scales had different number of items at baseline vs. the follow-up.

Therefore, we used individual mean scale scores. At the follow-up, where we had item-level data

available we simply averaged the item scores for each scale. At baseline, where we did not have

item-level data available (we only had the summed scale scores available), we simply divided the

available summed score by the known number of items present in each scale (reported in the

methods section). This method is somewhat problematic because we have no way of knowing

whether there were missing items for any of the participants in the scales measured at baseline.

Specifically, because all the items were dichotomized and summed, a score of 0 going into the

sum could either mean that an item had a score of 0 or it could be a missing item. Thus, without

this knowledge, dividing the total sum score available by the total number of known items might

underestimate the mean scores at baseline. Because this problem is not present at the follow-up

(where we had item-level data and could therefore distinguish between items with a zero score

and missing items), it follows that the standardized mean-level change reported in Table 6 could

be overestimated.

[6]An alternative way to compute RCI scores was to use only one standard error of

measurement (as opposed to a different one for each kind of form) computed using the

correlations between long-forms of the PTPI at Time 1 and short-forms of the PTPI at Time 2

(from the short-term longitudinal validation study), to simultaneously correct for measurement

error resulted from two different sources (a) re-testing and (b) switching from long- to short-

forms. The results using this alternative RCI computation can be found in Table 9S (in the

Supplementary Materials) and they did not differ in any meaningful way from the results

presented in Table 7 (which used the RCI computation presented in the paper). Specifically, the

average percentage of people who stayed the same was 59.67% in Table 7 vs. 60% in Table 9S.

[7]The latent framework rank-order stabilities (corrected for error as estimated in Table 3S

in the Supplementary Material, see also Footnote 3 for more details) were as follows: Vigor

(.22), Calmness (.32), Mature Personality (.31), Impulsiveness (.13), Self-Confidence (.31),

Culture (.41), Sociability (.29), Leadership (.32), Social Sensitivity (.35), Tidiness (.37). Thus,

using this alternative way to account for measurement error (as opposed to the way proposed in

Table 2S in the Supplementary Material), the average rank-order stability across PTPI scales

over 50 years was .30 (compared to .31 when we used errors from Table 2S). Thus, the particular

way of computing the errors used in the latent models (Table 2S vs. 3S) did not make a

meaningful difference.

[8]We also conducted one further analysis, where we looked at within-time normativeness

in our own cross-sectional data from the validation study and we looked at 20-year-olds and 60-

year-olds separately. We found that, at age 20, the average profile normativeness was .57 ($SD =$

.34), and that, at age 60, the average profile normativeness was .62 ($SD = .25$). This replicated

our longitudinal findings, thus, bringing further evidence that normativeness might increase with

age.

Table 1. Descriptive statistics of the *long forms* of Project Talent Personality Inventory Scales and mean-level age differences (cross-sectional validation study).

Scale	Total			20s			60s			
	α	*M*	*SD*	α	*M*	*SD*	α	*M*	*SD*	*d*
Vigor	.82	.33	.32	.79	.34	.31	.84	.32	.33	-.06
Calmness	.85	.68	.30	.84	.60	.32	.84	.73	.29	**.41**
Mature Personality	.91	.69	.25	.92	.62	.27	.89	.73	.22	**.41**
Impulsiveness	.56	.19	.17	.62	.19	.19	.53	.20	.16	.05
Self Confidence	.78	.52	.25	.75	.40	.24	.74	.58	.23	**.75**
Culture	.79	.53	.27	.78	.52	.27	.80	.55	.27	.11
Sociability	.79	.42	.25	.80	.39	.25	.78	.44	.24	**.20**
Leadership	.76	.27	.30	.74	.26	.30	.77	.28	.30	.07
Social Sensitivity	.82	.68	.29	.83	.65	.31	.80	.70	.28	**.16**
Tidiness	.85	.55	.30	.84	.51	.30	.85	.58	.29	**.23**

Note. Ns = 3,907-3,926 for total sample, *N* for the 20s = 1,243-1,246, *N* for the 60s = 2,405-2,420. The remaining participants were not in their 20s or 60s. Minor differences in *Ns* across scales were due to different numbers of missing items on each scale, which sometimes prevented scale computation. d = standardized mean-level change between 60 and 20 year olds. For the sake of comparison with previous research, to calculate d, we used the same procedure used by Roberts et al., 2006, namely the single-group, pretest-posttest raw score effect size (Morris & DeShon, 2002). Specifically, $d = (M_{age60} - M_{age20})/SD_{age20}$. Bold font indicates effect was statistically significant at $p < .001$.

Table 2. Descriptive statistics of the *short forms* of Project Talent Personality Inventory Scales and mean-level age differences (cross-sectional validation study).

Scale	Total			20s			60s			
	α	*M*	*SD*	α	*M*	*SD*	α	*M*	*SD*	*d*
Vigor	.83	.32	.35	.80	.30	.34	.84	.33	.36	.09
Calmness	.83	.68	.35	.82	.61	.37	.82	.73	.33	**.32**
Mature Personality	.79	.75	.31	.82	.69	.34	.75	.79	.28	**.29**
Impulsiveness	.69	.15	.24	.74	.18	.26	.66	.14	.22	**-.15**
Self Confidence	.68	.45	.31	.69	.33	.30	.64	.52	.29	**.63**
Culture	.75	.45	.33	.73	.45	.33	.76	.45	.34	.00
Sociability	.77	.51	.34	.76	.48	.34	.77	.53	.34	**.15**
Leadership	.76	.27	.30	.74	.26	.30	.77	.28	.30	.07
Social Sensitivity	.84	.73	.34	.83	.69	.36	.84	.77	.33	**.22**
Tidiness	.83	.50	.38	.80	.47	.37	.85	.52	.39	**.14**

Note. Ns = 3,920-3,927 for total sample, *N* for the 20s = 1,244-1,246, *N* for the 60s = 2,414-2,420. The remaining participants were not in their 20s or 60s. Minor differences in *Ns* across scales were due to different numbers of missing items on each scale, which sometimes prevented scale computation. *d* = standardized mean-level change between 60 and 20 year olds. For the sake of comparison with previous research, to calculate *d*, we used the same procedure used by Roberts et al., 2006, namely the single-group, pretest-posttest raw score effect size (Morris & DeShon, 2002). Specifically, $d = (M\text{age}60 – M\text{age}20)/SD\text{age}20$. Bold font indicates effect was statistically significant at *p* < .001

Table 3. Inter-correlations among the Project Talent Personality Inventory scales (cross-sectional validation study).

Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. Vigor_L	1	.33	.47	.20	.33	.40	.45	.45	.30	.39	.95	.36	.42	.20	.11	.41	.46	.45	.27	.38
2. Calmness_L		1	.65	.00	.52	.53	.42	.30	.60	.49	.36	.93	.59	-.05	.31	.43	.47	.30	.56	.43
3. Mature Personality_L			1	.10	.52	.57	.47	.45	.62	.63	.49	.67	.89	.04	.24	.49	.51	.45	.57	.53
4. Impulsiveness_L				1	.17	.07	.21	.28	.04	.03	.22	.02	.09	.83	.09	.11	.19	.28	.05	.04
5. Self-Confidence_L					1	.30	.43	.40	.28	.29	.37	.44	.42	-.05	.83	.26	.38	.41	.22	.23
6. Culture_L						1	.50	.38	.62	.57	.43	.54	.54	.11	.04	.92	.57	.38	.56	.53
7. Sociability_L							1	.40	.51	.37	.50	.42	.45	.19	.14	.47	.91	.40	.47	.35
8. Leadership_L								1	.29	.29	.48	.33	.38	.23	.14	.41	.40	1.00	.25	.25
9. Social Sensitivity_L									1	.46	.33	.59	.59	.07	.02	.50	.58	.29	.93	.41
10. Tidiness_L										1	.41	.52	.58	.04	.09	.53	.41	.29	.41	.93
11. Vigor_S											1	.38	.44	.20	.14	.43	.52	.48	.29	.40
12. Calmness_S												1	.63	.00	.21	.46	.48	.33	.56	.47
13. Mature Personality_S													1	.07	.15	.46	.50	.38	.56	.51
14. Impulsiveness_S														1	-.16	.16	.19	.23	.08	.06
15. Self-Confidence_S															1	.00	.06	.14	-.02	.04
16. Culture_S																1	.53	.41	.44	.50
17. Sociability_S																	1	.40	.53	.39
18. Leadership_S																		1	.25	.25
19. Social Sensitivity_S																			1	.38
20. Tidiness_S																				1

Note. Ns = 3,891-3,927. All values in bold font were statistically significant at $p < .001$. Values show correlations for the total sample. _L = long-form PTPI scales, _S = short-form PTPI scales.

Table 4. Mean-level differences between long- and short-forms of each PTPI scale in cross-sectional validation study (from paired samples t-tests).

PTPI Scale	Long Form		Short form		Mean Difference (Paired samples t-test)		
	M	*SD*	*M*	*SD*	*M-Diff*	*SD*	95% CI
Vigor	.33	.32	.32	.35	**.01**	.11	[.01;.01]
Calmness	.68	.3	.68	.35	.00	.13	[-.01;.00]
Mature Personality	.69	.25	.75	.31	**-.06**	.14	[-.07;-.06]
Impulsiveness	.19	.17	.15	.24	**.04**	.13	[.04;.05]
Self Confidence	.52	.25	.45	.31	**.06**	.17	[.06;.07]
Culture	.53	.27	.45	.33	**.09**	.14	[.09;.09]
Sociability	.42	.25	.51	.34	**-.09**	.15	[-.10;-.09]
Leadership	.27	.3	.27	.3	.00	.13	*N/A* *
Social Sensitivity	.68	.29	.73	.34	**-.05**	.15	[-.06;-.05]
Tidiness	.55	.3	.5	.38	**.05**	.11	[.05;.06]

Note: All mean differences in bold font were statistically significant at p < .001. CI= confidence interval. *95% CI was not available for the leadership scale because the test could not be performed as the long- and short- forms of the leadership scale were exactly the same.

Table 5. Rank-order stabilities and inter-correlations among the Project Talent Personality Inventory scales at baseline (T1) and at the 50th year follow-up (T2).

Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. VigorT1	.81																			
2. CalmnessT1	.40	.77																		
3. Mature PersonalityT1	.48	.57	.79																	
4. ImpulsivenessT1	.23	.12	.20	.61																
5. Self-ConfidenceT1	.30	.44	.39	.11	.89															
6. CultureT1	.39	.50	.55	.17	.28	.82														
7. SociabilityT1	.48	.38	.33	.24	.35	.39	.80													
8. LeadershipT1	.41	.38	.47	.23	.32	.40	.37	.76												
9. Social SensitivityT1	.39	.53	.52	.20	.25	.56	.48	.36	.85											
10. TidinessT1	.37	.50	.59	.09	.24	.57	.35	.33	.46	.88										
11. VigorT2	.20	.09	.17	.08	.08	.13	.15	.16	.11	.14	.80									
12. CalmnessT2	.13	.22	.19	.01	.11	.19	.17	.13	.19	.20	.41	.73								
13. Mature PersonalityT2	.11	.06	.18	.06	.08	.13	.15	.14	.15	.15	.54	.55	.66							
14. ImpulsivenessT2	.04	-.04	-.01	.09	-.03	.01	.07	.04	.02	.04	.24	.09	.17	.68						
15. Self-ConfidenceT2	.08	.12	.10	.01	.23	.04	.04	.04	.00	-.01	.07	.14	.05	-.24	.84					
16. CultureT2	.09	.18	.22	.06	.10	.34	.17	.18	.27	.25	.40	.45	.43	.18	-.03	.79				
17. SociabilityT2	.13	.10	.14	.06	.07	.19	.24	.12	.17	.16	.43	.43	.40	.25	-.03	.51	.78			
18. LeadershipT2	.14	.13	.19	.07	.16	.16	.14	.25	.13	.14	.44	.39	.47	.25	.11	.44	.41	.76		
19. Social SensitivityT2	.07	.13	.17	.03	.04	.24	.17	.10	.26	.22	.29	.50	.40	.15	-.12	.50	.56	.30	.80	
20. TidinessT2	.08	.12	.16	.00	.03	.18	.17	.09	.18	.30	.40	.48	.48	.13	-.08	.47	.39	.28	.42	.80

Note. N = 1,795 (we used listwise deletion to deal with missing data). On the main diagonal are the test-retest reliabilities from the validation study. On the grey diagonal are the rank-order stabilities of each personality trait across 50 years. All values in Bold font were statistically significant at $p < .001$.

Table 6. Descriptive statistics and mean-level changes in Project Talent Personality Inventory scales at baseline (T1) and at the 50th year follow-up (T2)

PTPI Scale (Big Five Corresponding Scale)	Time 1 (Baseline)		Time 2 (50th year follow-up)		Standardized Mean-level change	Standardized Mean-level change with robustness check
	M	*SD*	*M*	*SD*	*d*	*d-adjusted*
Vigor (Extraversion)	.56	.31	.56	.40	.00	.03
Calmness (Emotional Stability)	.50	.28	.80	.28	**1.07**	**1.07**
Mature Personality (Conscientiousness)	.49	.22	.85	.23	**1.64**	**1.34**
Impulsiveness (Low Conscientiousness)	.22	.18	.19	.24	**-.17**	.07
Self-Confidence (Emotional Stability)	.44	.21	.58	.26	**.67**	**.97**
Culture (Openness/Intellect)	.53	.24	.55	.33	**.08**	**.46**
Sociability (Extraversion)	.57	.24	.65	.33	**.33**	-.04
Leadership (Extraversion)	.27	.27	.40	.33	**.48**	**.48**
Social Sensitivity (Agreeableness)	.52	.26	.81	.30	**1.12**	**.91**
Tidiness (Conscientiousness)	.52	.26	.71	.36	**.73**	**.93**

Note. $N = 1,795$. d = standardized mean-level change between baseline and 50th year follow-up. For the sake of comparison with previous research, to calculate d, we used the same procedure used by Roberts et al., 2006, namely the single-group, pretest-posttest raw score effect size (Morris & DeShon, 2002). Specifically, $d = (Mfollow\text{-}up – Mbaseline)/SDbaseline$. Bold font indicates effect was statistically significant at $p < .001$.

Table 7. Individual-Level Change in Personality Traits from baseline (T1) to the 50th year follow-up (T2)

PTPI Scale (Big Five Corresponding Scale)	Decreased (%)	Stayed the same (%)	Increased (%)	χ^2 (2, N = 1,795)
Vigor (Extraversion)	17.4	67.2	15.4	**2,915.5**
Calmness (Emotional Stability)	3.4	54.7	41.9	**11,452**
Mature Personality (Conscientiousness)	2	39.3	58.7	**23,212**
Impulsiveness (Low Conscientiousness)	16.4	72.7	10.9	**1,989.9**
Self-Confidence (Emotional Stability)	10.2	52.5	37.3	**9,416.6**
Culture (Openness/Intellect)	12.6	69.9	17.5	**2,475.9**
Sociability (Extraversion)	12.3	66.3	21.4	**3,405**
Leadership (Extraversion)	4.1	78.6	17.3	**1,642.1**
Social Sensitivity (Agreeableness)	5.6	43.7	50.7	**17,191**
Tidiness (Conscientiousness)	9.9	51.8	38.3	**9,927.2**

Note. N = 1,795. Percentages for decrease, staying the same, and increase were based on the Reliable Change Index (RCI), where change less than -1.96 or greater than 1.96 is considered reliable change. The RCIs used for this table were computed using two different standard errors of measurement derived from two separate sets of test-retest reliabilities (from long and short scales, respectively). The chi-square tested whether the observed distribution of non-changers and changers differed from the expected distribution if change were random (i.e., 95% stayed the same, 2.5% each increased and decreased). Bold font indicates χ^2 test was statistically significant at *p* < .001

Table 8. Gender Differences in Personality Stability and Change from baseline (T1) to the 50th year follow-up (T2).

PTPI Scale (Big Five Corresponding Scale)	Cross-sectional differences						Longitudinal differences in change			
	Baseline (T1)		50th year (T2)		Gender d score		Change over time	Gender X Time	Gender correlation	Rank-order stability
	M	SD	M	SD	Baseline (T1)	50th year (T2)	d	Partial eta squared	with reliable change	
Vigor (Extraversion)					-.06	-.03		.00	-.01	
Women	.55	.31	.56	.39			.03			**.19**
Men	.57	.31	.57	.40			.00			**.20**
Calmness (Emotional Stability)					.11	.14		.00	-.01	
Women	.51	.29	.82	.26			**1.07**			**.20**
Men	.48	.28	.78	.30			**1.07**			**.22**
Mature Personality (Conscientiousness)					**.14**	.13		.00	-.05	
Women	.50	.22	.86	.21			**1.64**			**.18**
Men	.47	.22	.83	.24			**1.64**			.17
Impulsiveness (Low Conscientiousness)					.05	.08		.00	.00	
Women	.23	.19	.20	.24			-.16			.09
Men	.22	.18	.18	.24			**-.22**			.08
Self-Confidence (Emotional Stability)					.05	**-.24**		**.01**	-.06	
Women	.44	.22	.55	.26			**.50**			**.28**
Men	.43	.21	.61	.24			**.86**			**.18**
Culture (Openness/Intellect)					**.53**	**.49**		.00	-.06	
Women	.58	.22	.63	.31			**.23**			**.32**
Men	.46	.23	.47	.34			.04			**.27**
Sociability (Extraversion)					**.34**	**.36**		.00	-.05	
Women	.61	.23	.71	.32			**.43**			**.22**
Men	.53	.24	.59	.34			**.25**			**.21**
Leadership (Extraversion)					.00	-.12		.00	-.03	
Women	.27	.28	.38	.33			**.39**			**.25**
Men	.27	.27	.42	.33			**.56**			**.25**
Social Sensitivity (Agreeableness)					**.52**	**.64**		**.01**	-.07	
Women	.58	.25	.89	.23			**1.24**			**.22**
Men	.45	.25	.71	.33			**1.04**			**.19**
Tidiness (Conscientiousness)					**.40**	**.43**		.01	-.05	
Women	.57	.25	.78	.32			**.84**			**.26**
Men	.47	.25	.63	.38			**.64**			**.29**

Note. Nmales = 860; *Nfemales* = 935; Gender *d* score = differences between women and men divided by the pooled standard deviation (for the sake of comparison with Roberts et al., 2001). Change over time *d* = standardized mean-level change between baseline and 50th year follow-up; for the sake of comparison with previous research, to calculate the change over time *d*, we used the same procedure used by Roberts et al., 2006, namely the single-group, pretest-posttest raw score effect size (Morris & DeShon, 2002), that is, *d* = (*Mfollow-up – Mbaseline*)/*SDbaseline*. The Gender X Time Partial Eta Squared represents the magnitude of the effect size of the interaction between gender and time on personality change. Reliable change was recoded so that 0 = no change and 1 = reliable change; gender was coded 0=male and 1=female; the numbers represent phi correlations. Bold font indicates effect was statistically significant at *p* < .001.

Figure 1. Rank-order stability in Project Talent over 50 years (raw coefficients and coefficients adjusted for measurement error).
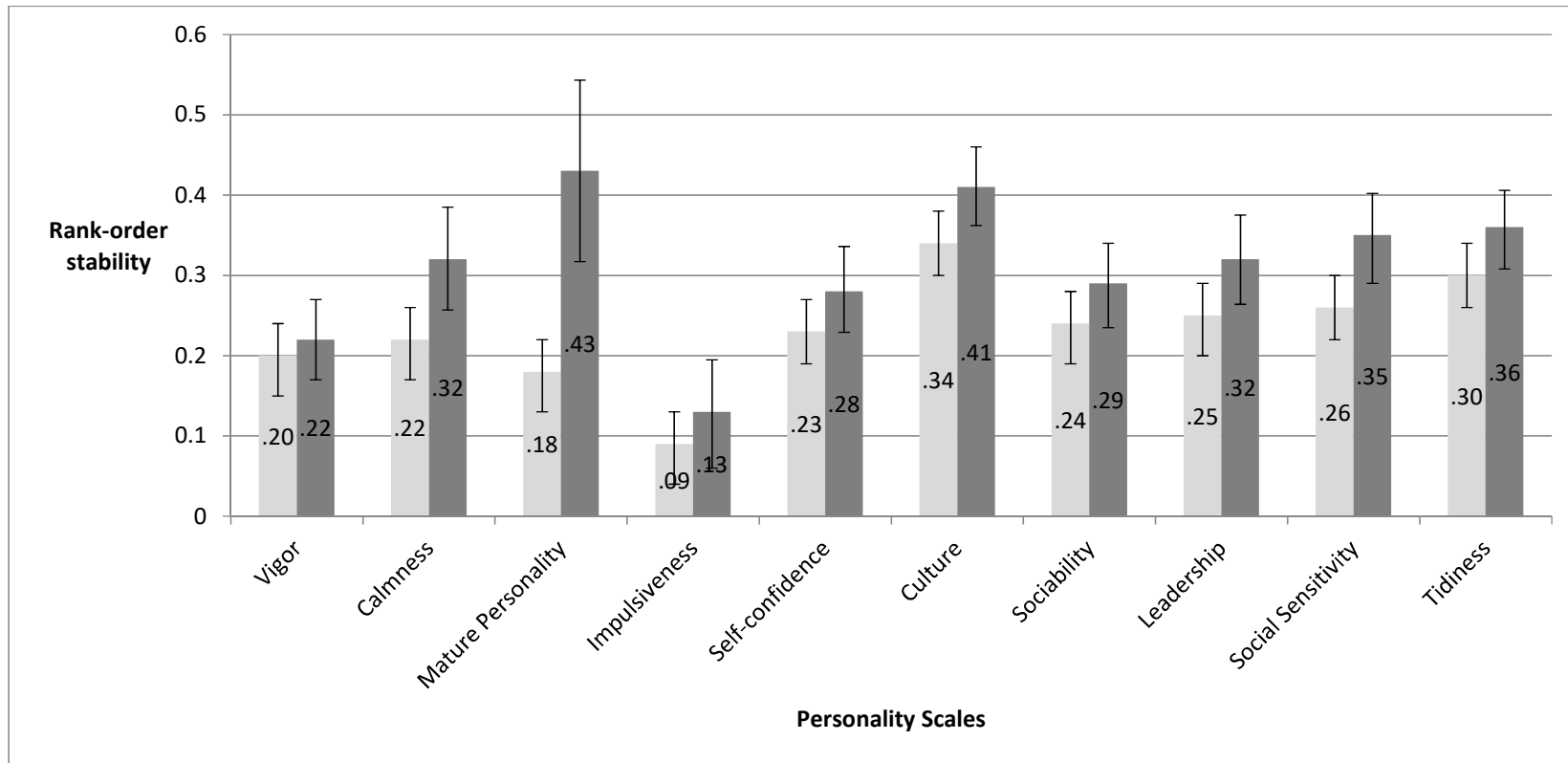
Note. Light gray bars represent raw stability coefficients, whereas dark gray bars represent stability coefficients obtained from using single-item latent models, corrected for measurement error. Bars represent 95% confidence intervals.

Figure 2. Normative Personality Trait Profiles from the main longitudinal study (baseline = adolescent norms in 1960; 50th year follow-up = older adult norms 2010) and from the cross-sectional validation study (20-year-olds vs. 60-year-olds norms in 2013).
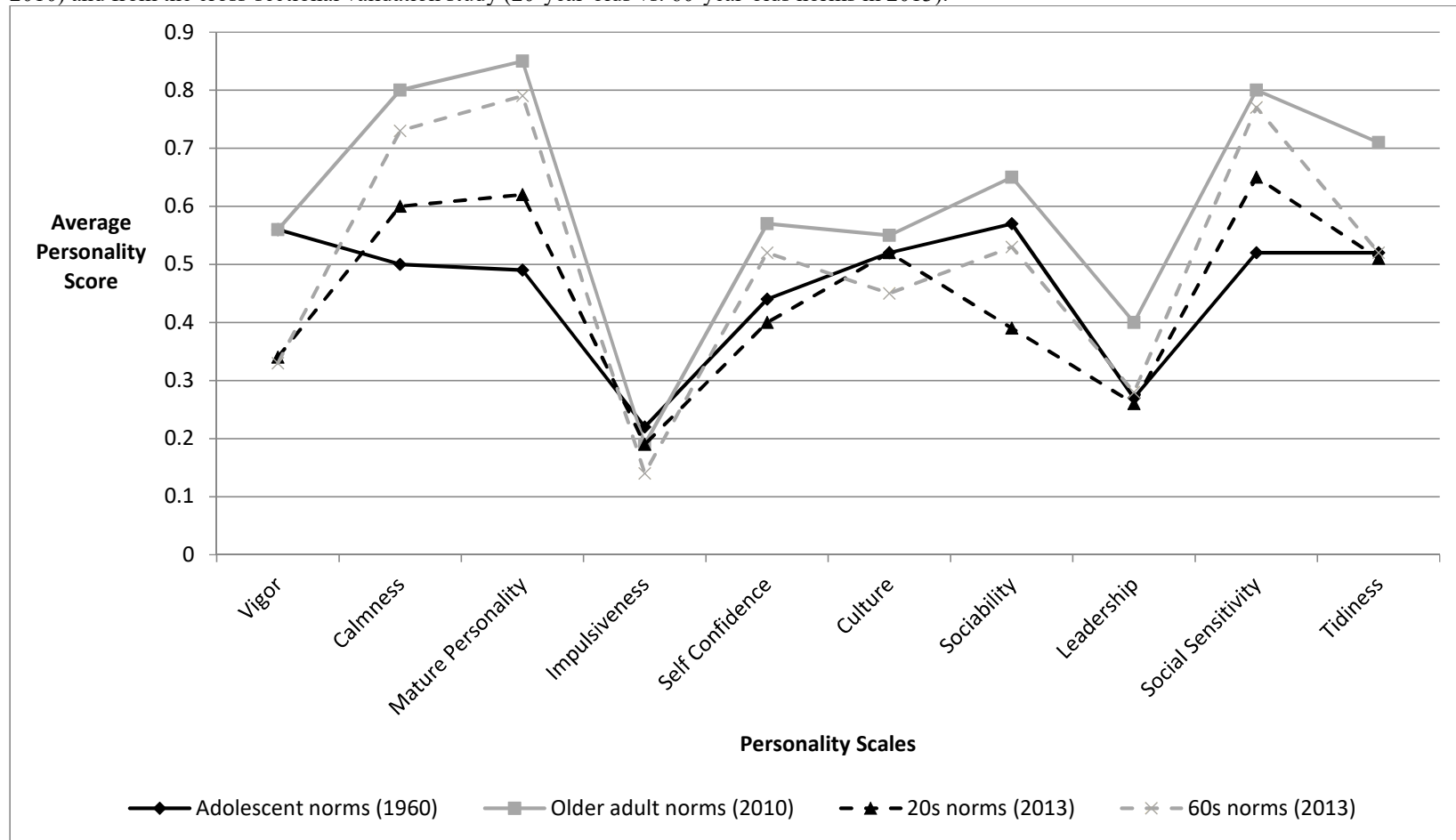
Online Supplementary Material for:

Sixteen Going on Sixty-Six:

A Longitudinal Study of Personality Stability and Change Across 50 Years

Table 1S. Items included in Long and Short Forms of the Project Talent Personality Inventory (PTPI). Long Forms Include ALL the items below, whereas short-forms include only items followed by an asterisk.

PTPI Items Organized by Scale	Coding
Vigor	
I can work or play outdoors for hours without getting tired.	
I am a fast walker.	
I am full of pep and energy.*	
People seem to think I lead a vigorous life.*	
I am active.*	
I am vigorous.*	
I am energetic.*	
Calmness	
I often lose my temper.	R
I can usually keep my wits about me even in difficult situations.	
People seem to think I get angry easily.	R
People seem to think I have good self-control.*	
People consider me level-headed.*	
I am even-tempered.	
I am calm.*	
I am stable.*	
I am usually self-controlled.*	
Mature Personality	
I make good use of all my time.	
I never seem to get things done on time.	R
I work fast and get a lot done.	

When I say I'll do something I get it done.	
It bothers me to leave a task half done.	
I can turn out a lot more work than average.	
I am hard-working.*	
People consider me an efficient worker.	
I do my job, even when I don't like it.	
I find it hard to keep working toward long-range goals.	**R**
I am productive.*	
As soon as I finish one project or assignment, I always have something else I want to begin.	
I never volunteer for a tough job.	**R**
I think that if something is worth starting its worth finishing.	
I do things the best I know how, even if no one checks up on me.	
I lose interest in most projects before I get them done.	**R**
People seem to think they can count on me.	
People consider me persistent.	
I am dependable.*	
People have criticized me for leaving things undone.	**R**
I am conscientious.	
I am persistent.	
I am reliable.*	
People consider me determined.*	
Impulsiveness	
I like to do things on the spur of the moment.*	
I usually act on the first plan that comes to mind.*	
I feel that I'm impulsive.*	
People seem to think I sometimes make decisions too quickly.*	
I am impulsive.*	

I don't believe in rushing into things.	R
I am cautious.	R
When I have a problem, I make up my mind and don't worry about it.	
It takes me quite a while to come to a decision.	R
Self-Confidence	
I am confident.	
I'd enjoy speaking to a club group on a subject I know	
Being around strangers makes me ill-at-ease.	R
I'm troubled by people making fun of me.	R
People seem to think my feelings are hurt too easily.*	R
I am usually at ease.	
People seem to think I am easily discouraged when criticized.*	R
I am often self-conscious.*	R
People consider me shy.	R
I am sensitive.*	R
I am often worried.*	R
People seem to think I usually do a good job on whatever I'm doing.	
Culture	
I enjoy beautiful things.	
I feel that good manners are very necessary for everyone.	
I think culture is more important than wealth.	
I enjoy cultural things.*	
I am a cultured person.*	
People seem to think I have good taste.*	
I take part in the cultural activities in my community.*	
I tend to have good taste.*	

I am refined.	
I am sometimes crude.	**R**
Sociability	
I like to spend a good deal of time by myself.	**R**
I'd rather be with a group of friends than at home by myself.	
People consider me the quiet type.	**R**
People seem to think I make new friends more quickly than most people do.	
I couldn't get along without having people around me most of the time.	
I enjoy getting to know people.*	
I like to be with people most of the time.*	
I go out of my way to be with friends.*	
I prefer reading a good book to going out with friends.	**R**
People consider me good-natured.	
People consider me sociable.*	
I am friendly.*	
Leadership	
I am the leader in my group.*	
I am influential*	
I have held a lot of elected offices.*	
People naturally follow my lead.*	
I like to make decisions.*	
Social Sensitivity	
I like to tease people.	**R**
I never hurt another person's feelings if I can avoid it.*	
I seem to know how other people will feel about things.	

I sympathize with my friends and encourage them when they have problems.*	
People consider me a sympathetic listener.*	
People consider me very helpful in dealing with other people.	
I am sympathetic*	
I am considerate.*	
People consider me understanding.*	
Tidiness	
I am never sloppy in my personal appearance.*	
I have a definite place for all of my things.	
Before I start a task, I spend some time getting it organized.	
It bothers me to be with someone who dresses carelessly.	
I like to do things systematically.	
My work suffers from lack of neatness.	R
People consider me very careful about my personal appearance.*	
I am tidy.*	
I am neat.*	
I am orderly.*	
I tend to be untidy.	R

Note: PTPI Instructions read as follows: "For each statement below mark which the one of the five choices which best describes how the statement applies to you. Regarding the things I do and the way I do them, this statement describes me: 1 (not very well), 2 (slightly), 3 (fairly well), 4 (quite well), 5 (extremely well)." "R" means the item was reverse-coded. *item included in the short-form version of the scale.

Validation Study

Table 2S. Descriptive statistics for short-term longitudinal validation study and residual (error) used in the latent-model (see "Main Study").

Scale Type	Variable Name	Mean Time 1	Variance Time1	Mean Time 2	Variance Time 2	Sample Size T1 and T2	Pooled Variance	Test-retest	Error
Long	Vigor	0.33	0.06	0.33	0.08	38.00	0.07	0.81	0.01
Long	Calmness	0.70	0.08	0.65	0.10	38.00	0.09	0.77	0.02
Long	Mature Personality	0.64	0.07	0.59	0.07	38.00	0.07	0.79	0.01
Long	Impulsiveness	0.19	0.03	0.19	0.03	38.00	0.03	0.61	0.01
Long	Self Confidence	0.48	0.05	0.48	0.07	38.00	0.06	0.89	0.01
Long	Culture	0.49	0.07	0.47	0.06	38.00	0.07	0.82	0.01
Long	Sociability	0.31	0.05	0.31	0.05	38.00	0.05	0.80	0.01
Long	Leadership	0.27	0.08	0.23	0.08	38.00	0.08	0.76	0.02
Long	Social Sensitivity	0.57	0.10	0.60	0.10	38.00	0.10	0.85	0.01
Long	Tidiness	0.46	0.08	0.44	0.07	38.00	0.08	0.88	0.01
Short	Vigor	0.27	0.07	0.26	0.10	38.00	0.09	0.80	0.02
Short	Calmness	0.67	0.11	0.63	0.12	38.00	0.12	0.73	0.03
Short	Mature Personality	0.74	0.11	0.65	0.15	38.00	0.13	0.66	0.04
Short	Impulsiveness	0.16	0.06	0.13	0.05	38.00	0.05	0.68	0.02
Short	Self Confidence	0.50	0.07	0.52	0.10	38.00	0.09	0.84	0.01
Short	Culture	0.39	0.11	0.39	0.10	38.00	0.11	0.79	0.02
Short	Sociability	0.38	0.10	0.36	0.08	38.00	0.09	0.78	0.02
Short	Leadership	0.27	0.08	0.23	0.08	38.00	0.08	0.76	0.02
Short	Social Sensitivity	0.66	0.15	0.66	0.15	38.00	0.15	0.80	0.03
Short	Tidiness	0.41	0.11	0.37	0.13	38.00	0.12	0.80	0.02

Note: Formula used for pooled variance: [Var_t1*(N_t1-1)+Var_t2(N_t2 -1)]/[(N_t1-1)+(N_t2 -1)]. Formula used for residual (error) that we fixed in the latent framework: Pooled_var*(1-Test-retest).

Table 3S. Descriptive statistics for short-term longitudinal validation study and residual (error) used in the second set of latent-models (see Footnotes 2 and 7).

Scale Type	Variable Name	Mean Time 1	Variance Time1	Mean Time 2	Variance Time 2	Sample Size T1 and T2	Pooled Variance	Test-retest	Error
Long	Vigor	0.33	0.06	0.33	0.08	38.00	0.07	0.81	0.01
Long	Calmness	0.70	0.08	0.65	0.10	38.00	0.09	0.77	0.02
Long	Mature Personality	0.64	0.07	0.59	0.07	38.00	0.07	0.79	0.01
Long	Impulsiveness	0.19	0.03	0.19	0.03	38.00	0.03	0.61	0.01
Long	Self Confidence	0.48	0.05	0.48	0.07	38.00	0.06	0.89	0.01
Long	Culture	0.49	0.07	0.47	0.06	38.00	0.07	0.82	0.01
Long	Sociability	0.31	0.05	0.31	0.05	38.00	0.05	0.80	0.01
Long	Leadership	0.27	0.08	0.23	0.08	38.00	0.08	0.76	0.02
Long	Social Sensitivity	0.57	0.10	0.60	0.10	38.00	0.10	0.85	0.01
Long	Tidiness	0.46	0.08	0.44	0.07	38.00	0.08	0.88	0.01
Long T1 to Short T2	Vigor	0.33	0.06	0.26	0.10	38.00	0.08	0.76	0.02
Long T1 to Short T2	Calmness	0.70	0.08	0.63	0.12	38.00	0.10	0.68	0.03
Long T1 to Short T2	Mature Personality	0.64	0.07	0.65	0.15	38.00	0.11	0.72	0.03
Long T1 to Short T2	Impulsiveness	0.19	0.03	0.13	0.05	38.00	0.04	0.42	0.02
Long T1 to Short T2	Self Confidence	0.48	0.05	0.52	0.10	38.00	0.08	0.75	0.02
Long T1 to Short T2	Culture	0.49	0.07	0.39	0.10	38.00	0.08	0.73	0.02
Long T1 to Short T2	Sociability	0.31	0.05	0.36	0.08	38.00	0.07	0.74	0.02
Long T1 to Short T2	Leadership	0.27	0.08	0.23	0.08	38.00	0.08	0.76	0.02
Long T1 to Short T2	Social Sensitivity	0.57	0.10	0.66	0.15	38.00	0.12	0.77	0.03
Long T1 to Short T2	Tidiness	0.46	0.08	0.37	0.13	38.00	0.10	0.74	0.03

Note: Formula used for pooled variance: [Var_t1*(N_t1-1)+Var_t2(N_t2 -1)]/[(N_t1-1)+(N_t2 -1)]. Formula used for residual (error) that we fixed in the latent framework: Pooled_var*(1-Test-retest). In the "Main Study," when using this alternative set of residuals to correct for measurement error, for Time 1 (baseline), we used the same set of residuals as before (like in Table 2S, which is why the Long-form section of Table 3S is identical to the Long-form section of Table 2S), however, for Time 2 (50th year follow-up), instead of using means, standard deviations, and test-retest reliabilities obtained from the short-term longitudinal validation study for the short-forms of the PTPI, we used data from both long- and short-forms of the PTPI and correlations between long-forms of the PTPI at Time 1 and short-forms of the PTPI at Time 2 (from the short-term longitudinal validation study), to simultaneously correct for unreliability resulted from two different sources (a) re-testing and (b) switching from long- to short-forms.

Notes regarding measurement invariance testing (see Tables 4S ad 5S below):

We ran 20 sets of invariance models in a multi-group framework to test for measurement invariance across the 20-year-olds and the 60-year-olds in the cross-sectional validation study. For individual models, model fit is considered adequate if CFI is > 0.95 and RMSEA is < .08. Most individual models, across scales, showed reasonably good fit, though short-forms showed better fit than long-forms, and some scales had larger RMSEA values. We did not consider this to be a major problem, because we are not using these latent factors in our main analyses (we only used manifest variables in the main analyses, and this was mentioned as a limitation in the discussion section on p. 46 of the manuscript). Furthermore, the focus when testing measurement invariance is on relative fit.

Based on the measurement invariance literature, we determined the level of invariance with ΔCFI (Cheung & Rensvold, 2002) of two consecutive models. For a model to be accepted, it had to yield acceptable relative fit compared with the preceding invariance model (ΔCFI > -.01, that is, if the model had a worse fit, it was not supposed to deviate more than -.01, but to be clear, if the CFI increased in a subsequent, more constrained, model then the rule did not need to be applied, because in those cases the model fit better, so there was no reason to reject it). As can be seen in Table 4S (long-forms), following the ΔCFI > -.01 rule, all scales showed evidence for measurement invariance (configural, metric, scalar, and strict) (the Sociability scale's CFI deviated by -.011 when going from metric to scalar, but in that case, the RMSEA 90% confidence intervals overlapped, so we still had some evidence for invariance). As can be seen in Table 5S (short-forms), following the ΔCFI > -.01 rule, 9 out of 10 scales showed evidence for measurement invariance (configural, metric, scalar, and strict), the only exception being self-confidence, which showed evidence for metric, but not scalar invariance. As alluded to earlier, a second guideline for testing relative fit and assessing measurement invariance is to check whether the 90% confidence intervals for the RMSEA are overlapping across consecutive models (Allemand, Zimprich, & Hertzog, 2007; MacCallum, Browne, & Sugawara, 1996). Following, this guideline, not all scales showed measurement invariance, but according to Little (2013), these are just general guidelines, cutoffs being somewhat arbitrary, and a ΔCFI > -.01 might bring sufficient evidence for measurement invariance, especially in cases where the RMSEA fit itself does not change or even becomes better with subsequent models (which was the case for several of our scales).

To further facilitate a detailed interpretation of the results presented in Tables 4S and 5S, we give one specific example. The model with the unconstrained factor loadings and intercepts can be found in the first row labeled "configural." A configural invariance model was initially specified in which a

single factor was estimated simultaneously in both age groups. All item factor loadings (one per item) and thresholds (one per item given two response options) were estimated. The residual variances are not uniquely identified in the configural invariance model (when applying it to dichotomous/categorical items) and as such were all constrained to 1 in both groups. As shown in Table 5S, the configural invariance model for the Social Sensitivity scale had a good fit. The analysis proceeded by applying parameter constraints in successive models to examine potential decreases in fit resulting from measurement non-invariance between 20-year-olds and 60-year-olds. In the metric invariance model, we constrained the respective factor loadings to be equal across age-groups. The metric invariance model did not fit significantly worse than the configural invariance model. In a next step, the scalar invariance model was fitted with the item thresholds constrained to be equal across age-groups. The scalar invariance model did not fit significantly worse than the metric invariance model. In a last step we tested for strict measurement invariance. The model comparison at this step proceeded backwards, such that a model with all residual variances freely estimated in the 60-year-olds was fitted first, and then compared with a model in which all residual variances were fixed to 1 in both groups. Model "strict b" did not fit significantly worse than model "strict a". Therefore, strict measurement invariance in Social Sensitivity can be assumed across the 20- and 60-year-olds.

Table 4S. Measurement Invariance (20- vs. 60-year-olds) Long Forms of the PTPI

	χ^2	df	p	CFI	RMSEA	90% CI
Calmness						
Configural	2,354.26	54	.00	.930	.152	[.147; .158]
Metric	1,931.04	62	.00	.943	.128	[.123; .133]
Scalar	2,175.54	70	.00	.936	.128	[.123; .133]
Strict a	2,418.93	61	.00	.929	.145	[.140 ; .150]
Strict b	2,175.54	70	.00	.936	.128	[123; .133]
Culture						
Configural	1,492.36	70	.00	.936	.105	[.101; .110]
Metric	1,223.88	79	.00	.948	.089	[.084; .093]
Scalar	1,412.24	88	.00	.940	.091	[.086; .095]
Strict a	1,422.21	78	.00	.939	.097	[.093; .101]
Strict b	1,412.24	88	.00	.940	.091	[.086; .095]
Impulsiveness						
Configural	493.27	54	.00	.990	.067	[.061; .072]
Metric	430.29	62	.00	.992	.057	[.052; .062]
Scalar	684.76	70	.00	.986	.069	[.065; .074]
Strict a	644.35	64	.00	.990	.070	[.065; .075]
Strict b	684.76	70	.00	.986	.069	[.065; .074]
Leadership						
Configural	36.50	10	.00	.999	.038	[.025; .052]
Metric	35.62	14	.00	.999	.029	[.017; .041]
Scalar	115.96	18	.00	.995	.054	[.045; .064]
Strict a	74.88	13	.00	.997	.051	[.040; .062]
Strict b	115.96	18	.00	.995	.054	[.045; .064]

Mature Personality						
Configural	8,332.79	504	.00	.903	.092	[.090; .094]
Metric	6,533.02	527	.00	.925	.079	[.077; .081]
Scalar	7,064.26	550	.00	.919	.080	[.079; .082]
Strict a	8,443.72	526	.00	.902	.091	[.089; .092]
Strict b	7,064.26	550	.00	.919	.080	[.079; .082]
Self Confidence						
Configural	2,461.33	108	.00	.796	.109	[.105; .113]
Metric	2,261.96	119	.00	.815	.099	[.096; .103]
Scalar	2,389.24	130	.00	.805	.097	[.094; .101]
Strict a	2,508.78	118	.00	.793	.105	[.102; .109]
Strict b	2,389.24	130	.00	.805	.097	[.094; .101]
Sociability						
Configural	2,103.48	108	.00	.902	.100	[.097; .104]
Metric	1,723.052	119	.00	.921	.086	[.082; .089]
Scalar	1,945.30	130	.00	.910	.087	[.084; .091]
Strict a	2,116.30	118	.00	.901	.096	[.092; .100]
Strict b	1,945.30	130	.00	.910	.087	[.084; .091]
Social Sensitivity						
Configural	528.64	54	.00	.982	.069	[.064; .075]
Metric	452.81	62	.00	.985	.059	[.054; .064]
Scalar	624.28	70	.00	.979	.066	[.061; .070]
Strict a	517.23	61	.00	.983	.064	[.059; .069]
Strict b	624.28	70	.00	.979	.066	[.061; .070]
Tidiness						

Configural		2,007.03	88	.00	.956	.109	[.105; .113]
Metric		1,490.30	98	.00	.968	.088	[.084; .092]
Scalar		1,773.30	108	.00	.962	.092	[.088; .095]
Strict a		1,982.09	97	.00	.957	.103	[.099; .107]
Strict b		1,773.30	108	.00	.962	.092	[.088; .095]
Vigor						
Configural		283.45	28	.00	.991	.071	[.063; .078]
Metric		235.94	34	.00	.993	.057	[.050; .064]
Scalar		487.96	40	.00	.984	.078	[.072; .084]
Strict a		504.86	33	.00	.983	.088	[.082; .095]
Strict b		487.96	40	.00	.984	.078	[.072; .084]

Table 5S. Measurement Invariance (20- vs. 60-year-olds) Short Forms of the PTPI

		χ^2	df	p	CFI	RMSEA	90% CI
Calmness							
Configural		53.51	10	.00	.998	.049	[.036; .062]
Metric		58.99	14	.00	.997	.042	[.031; .053]
Scalar		197.59	18	.00	.990	.074	[.065; .083]
Strict a		97.76	13	.00	.995	.060	[.049; .071]
Strict b		197.59	18	.00	.990	.074	[.065; .083]
Culture							
Configural		733.41	10	.00	.945	.199	[.187; .211]
Metric		681.83	14	.00	.949	.161	[.151; .172]
Scalar		719,109	18	.00	.946	.146	[.137; .155]
Strict a		769,347	13	.00	.942	.178	[.167; .189]
Strict b		719,109	18	.00	.946	.146	[.137; .155]
Impulsiveness							
Configural		68.45	10	.00	.999	.056	[.044; .069]
Metric		76,671	14	.00	.999	.049	[.039; .060]
Scalar		95,918	18	.00	.998	.049	[.039; .058]
Strict a		69,263	13	.00	.999	.049	[.038; .060]
Strict b		95,918	18	.00	.998	.049	[.039; .058]
Leadership							
Configural		36.50	10	.00	.999	.038	[.025; .052]
Metric		35.62	14	.00	.999	.029	[.017; .041]
Scalar		115.97	18	.00	.995	.054	[.045; .064]
Strict a		74.88	13	.00	.997	.051	[.040; .062]
Strict b		115.97	18	.00	.995	.054	[.045; .064]

Mature Personality						
Configural	327.49	10	.00	.991	.132	[.120; .144]
Metric	348.75	14	.00	.991	.114	[.104; .125]
Scalar	423.44	18	.00	.989	.111	[.102; .120]
Strict a	371.63	13	.00	.990	.123	[.112; .134]
Strict b	423.44	18	.00	.989	.111	[.102; .120]
Self Confidence						
Configural	46.96	10	.00	.990	.045	[.032; .058]
Metric	59.71	14	.00	.987	.042	[.031; .053]
Scalar	158.24	18	.00	.961	.065	[.056; .075]
Strict a	126.54	13	.00	.969	.069	[.058; .080]
Strict b	158.24	18	.00	.961	.065	[.056; .075]
Sociability						
Configural	144.26	10	.00	.988	.086	[.073; .098]
Metric	140.30	14	.00	.988	.070	[.060; .081]
Scalar	258.22	18	.00	.978	.085	[.076; .095]
Strict a	231.50	13	.00	.980	.096	[.085; .107]
Strict b	258.22	18	.00	.978	.085	[.076; .095]
Social Sensitivity						
Configural	174.39	10	.00	.992	.095	[.083; .107]
Metric	160.37	14	.00	.992	.075	[.065; .086]
Scalar	192,597	18	.00	.991	.073	[.064; .082]
Strict a	181.75	13	.00	.991	.084	[.074; .095]
Strict b	192,597	18	.00	.991	.073	[.064; .082]
Tidiness						

Configural		566.46	10	.00	.983	.174	[.162; .187]
Metric		471.89	14	.00	.986	.134	[.123; .144]
Scalar		508.70	18	.00	.985	.122	[.113; .131]
Strict a		621.54	13	.00	.982	.160	[.149; .171]
Strict b		508.70	18	.00	.985	.122	[.113; .131]
Vigor						
Configural		76.74	10	.00	.997	.060	[.048; .073]
Metric		89.58	14	.00	.997	.054	[.044; .065]
Scalar		141.51	18	.00	.995	.061	[.052; .071]
Strict a		103.21	13	.00	.996	.061	[.051; .073]
Strict b		141.51	18	.00	.995	.061	[.052; .071]

Main Study

Table 6S. Attrition analyses based on participation in personality assessments at baseline and at the 50th year follow-up.

| Variable | Baseline only vs. Year 50 | | | Frequency distributions | |
	Mean Difference	95% CI	d	Baseline only	Year 50
Gender (Female)	N/A	N/A	N/A	52.8%	52%
Race (White)	N/A	N/A	N/A	90%	95.3%
Vigor	**-.06**	[-.07; -.04]	.18	N/A	N/A
Calmness	**-.05**	[-.07; -.03]	.18	N/A	N/A
Mature Personality	**-.04**	[-.05; -.02]	.18	N/A	N/A
Impulsiveness	-.01	[-.02; .00]	.04	N/A	N/A
Self-Confidence	-.02	[-.03; .00]	.08	N/A	N/A
Culture	-.02	[-.03; .00]	.07	N/A	N/A
Sociability	-.02	[-.04; -.01]	.09	N/A	N/A
Leadership	-.02	[-.04; .00]	.08	N/A	N/A
Social Sensitivity	-.02	[-.04; -.01]	.09	N/A	N/A
Tidiness	-.02	[-.03; .00]	.06	N/A	N/A

Notes. For the continuous variables, we provide mean differences, 95% confidence intervals, and absolute effect sizes (Cohen's *d*), where negative mean differences indicate higher scores for the people who stayed in the study at year 50 versus those who dropped out. For the dichotomous variables, we provide frequency distributions for the people who dropped (baseline only) and the people who stayed in the study at year 50. *N/A* indicates that the respective statistics are not available because the variable is either dichotomous or continuous, respectively. Total sample size of people who dropped out was 2,655. Total sample size of people who participated at year 50 was 1,858.

Table 7S (equivalent of Table 5, but with FIML). Rank-order stabilities and inter-correlations among the Project Talent Personality Inventory scales at baseline (T1) and at the 50th year follow-up (T2).

Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21. VigorT1	.81																			
22. CalmnessT1	.42	.77																		
23. Mature PersonalityT1	.51	.60	.79																	
24. ImpulsivenessT1	.24	.13	.19	.61																
25. Self-ConfidenceT1	.29	.44	.39	.09	.89															
26. CultureT1	.42	.53	.58	.20	.29	.82														
27. SociabilityT1	.50	.42	.40	.25	.34	.44	.80													
28. LeadershipT1	.40	.39	.48	.26	.29	.43	.36	.76												
29. Social SensitivityT1	.43	.57	.57	.23	.27	.61	.51	.41	.85											
30. TidinessT1	.40	.53	.62	.12	.26	.59	.42	.35	.53	.88										
31. VigorT2	.20	.10	.17	.08	.08	.13	.16	.16	.12	.14	.80									
32. CalmnessT2	.14	.22	.21	.02	.11	.20	.18	.14	.20	.21	.41	.73								
33. Mature PersonalityT2	.12	.08	.19	.06	.08	.13	.15	.15	.16	.16	.54	.56	.66							
34. ImpulsivenessT2	.05	-.02	-.00	.09	-.03	.02	.07	.05	.03	.05	.24	.10	.17	.68						
35. Self-ConfidenceT2	.07	.10	.09	.00	.22	.03	.03	.04	.00	-.01	.07	.14	.05	-.24	.84					
36. CultureT2	.11	.19	.24	.07	.11	.34	.18	.20	.28	.26	.40	.46	.44	.19	-.03	.79				
37. SociabilityT2	.14	.12	.16	.07	.08	.20	.24	.12	.18	.18	.43	.45	.40	.25	-.03	.52	.78			
38. LeadershipT2	.14	.14	.20	.08	.15	.16	.14	.24	.15	.15	.45	.40	.47	.25	.10	.44	.41	.76		
39. Social SensitivityT2	.10	.15	.19	.04	.05	.25	.19	.12	.27	.25	.30	.51	.41	.15	-.11	.50	.57	.31	.80	
40. TidinessT2	.08	.14	.18	.01	.04	.19	.19	.10	.19	.31	.40	.49	.49	.13	-.08	.47	.40	.28	.43	.80

Note. These analyses used a FIML estimation based on $N = 4{,}510$, which was the baseline sample. On the main diagonal are the test-retest reliabilities from the validation study. On the grey diagonal are the rank-order stabilities of each personality trait across 50 years. All values in Bold font are statistically significant at $p < .001$.

Table 8S (equivalent of Table 6, but with FIML). Descriptive statistics and mean-level changes.

PTPI Scale (Big Five Corresponding Scale)	Time 1 (Baseline)		Time 2 (50th year follow-up)		Standardized Mean-level change	Standardized Mean-level change with robustness check
	M	*SD*	*M*	*SD*	*d*	*d-adjusted*
Vigor (Extraversion)	.53	.30	.56	.40	**.12**	**.15**
Calmness (Emotional Stability)	.46	.28	.80	.28	**1.20**	**1.20**
Mature Personality (Conscientiousness)	.46	.21	.85	.23	**1.79**	**1.49**
Impulsiveness (Low Conscientiousness)	.22	.18	.19	.24	**-.15**	.08
Self-Confidence (Emotional Stability)	.43	.20	.57	.25	**.72**	**1.03**
Culture (Openness/Intellect)	.51	.23	.56	.33	**.18**	**.56**
Sociability (Extraversion)	.56	.24	.65	.33	**.40**	.03
Leadership (Extraversion)	.26	.27	.40	.33	**.50**	**.50**
Social Sensitivity (Agreeableness)	.51	.26	.80	.30	**1.14**	**.94**
Tidiness (Conscientiousness)	.52	.25	.71	.36	**.76**	**.97**

Note. These analyses used a FIML estimation based on *N* = 4,510, which was the baseline sample. All mean differences in bold font were statistically significant at *p* < .001. *d* = standardized mean-level change between baseline and the 50th year follow-up. For the sake of comparison with previous research, to calculate *d*, we used the same procedure used by Roberts et al., 2006, namely the single-group, pretest-posttest raw score effect size (Morris & DeShon, 2002). Specifically, *d* = (*Mfollow-up* – *Mbaseline*)/*SDbaseline*. Bold font indicates effect was statistically significant at *p* < .001

Table 9S. Individual-Level Change in Personality Traits from baseline (T1) to the 50th year follow-up (T2)

PTPI Scale (Big Five Corresponding Scale)	Decreased (%)	Stayed the same (%)	Increased (%)	χ^2 (2, N = 1,795)
Vigor (Extraversion)	19.2	61.7	19.1	**4,183.2**
Calmness (Emotional Stability)	3.2	55.7	41.1	**10,968**
Mature Personality (Conscientiousness)	2	39.3	58.7	**23,212**
Impulsiveness (Low Conscientiousness)	8.2	81.3	10.5	**727.49**
Self-Confidence (Emotional Stability)	7.6	60.1	32.3	**6,774.4**
Culture (Openness/Intellect)	12.6	70	17.5	**2,475.9**
Sociability (Extraversion)	14.9	58	27.1	**5,695.2**
Leadership (Extraversion)	11.3	58.6	30.1	**6,264.9**
Social Sensitivity (Agreeableness)	4.8	53.5	41.7	**11,377**
Tidiness (Conscientiousness)	8	61.8	30.2	**5,910.2**

Note. $N = 1,795$. Percentages for decrease, staying the same, and increase were based on the Reliable Change Index (RCI), where change less than -1.96 or greater than 1.96 is considered reliable change. The RCIs used for this table were computed using only one standard error of measurement derived from the correlation between the long-form at Time 1 and the short-form at Time 2 in the validation short-term longitudinal study. The chi-square tested whether the observed distribution of non-changers and changers differed from the expected distribution if change were random (i.e., 95% stayed the same, 2.5% each increased and decreased). Bold font indicates χ^2 test was statistically significant at $p < .001$