Journal of Personality and Social Psychology
2011, Vol. 100, No. 6, 1124–1138

0022-3514/11/$12.00 DOI: 10.1037/a0022874

Cohort Differences in Big Five Personality Factors Over a Period of 25 Years

Iris A. M. Smits
University of Groningen

Conor V. Dolan, Harrie C. M. Vorst, and Jelte M. Wicherts
University of Amsterdam

Marieke E. Timmerman
University of Groningen

The notion of personality traits implies a certain degree of stability in the life span of an individual. But what about generational effects? Are there generational changes in the distribution or structure of personality traits? This article examines cohort changes on the Big Five personality factors Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience, among first-year psychology students in the Netherlands, ages 18 to 25 years, between 1982 and 2007. Because measurement invariance of a personality test is essential for a sound interpretation of cohort differences in personality, we first assessed measurement invariance with respect to cohort for males and females separately on the Big Five personality factors, as measured by the Dutch instrument Five Personality Factors Test. Results identified 11 (females) and 2 (males) biased items with respect to cohort, out of a total of 70 items. Analyzing the unbiased items, results indicated small linear increases over time in Extraversion, Agreeableness, and Conscientiousness and small linear decreases over time in Neuroticism. No clear patterns were found on the Openness to Experience factor. Secondary analyses on students from 1971 to 2007 of females and males of different ages together revealed linear trends comparable to those in the main analyses among young adults between 1982 onward. The results imply that the broad sociocultural context may affect personality factors.

Keywords: Big Five personality factors, cohort, measurement invariance

In many ways, the present society is incomparable to the society of 10 years ago, let alone 20 or 30 years ago. Yet are people, in psychological respect, likewise subject to change? While generational changes in intelligence (IQ) test scores are well established and much debated (Flynn, 2007; Neisser, 1998; Wicherts et al., 2004), changes in personality test scores have enjoyed less attention. The stability implied by the notion of personality traits pertains to the life span of an individual and does not preclude generational changes in the distribution or structure of personality traits. However, previous research has suggested that, besides well-established genetic and individual environmental influences (Eaves, Eysenck, & Martin, 1989), the broad sociocultural context may also affect human personality (e.g., Twenge, 2000, 2001a).

The sociocultural environment changes over the years. Influenced by historical events such as the destruction of the Twin Towers (9-11 for short), the end of the Cold War (1989), and technological developments like the emergence of the Internet, the sociocultural environment of a society changes. As a consequence, people from different cohorts experience different sociocultural events, and they experience the same sociocultural events at different ages. Previous research (e.g., Elder, 1998; Ryder, 1965; Schaie & Elder, 2005; Stewart & Healy, 1989) has suggested that social-historical events may have different consequences for people of different ages and may therefore contribute to differences between birth cohorts. A historical event like 9-11 may well affect young children, adolescents, and adults differently. As such, people of the same generation may share more than their chronological age. They share the sociocultural contexts of their societies. For this reason, birth cohort is sometimes referred to as being a representative of the broad sociocultural environment (Twenge, 2001a). Twenge (2000, 2001a) argued that this broad sociocultural environment should be considered as an additional influence on personality, in addition to genetic and personal environmental influences, suggesting that personality may well be more changeable in nature than previously assumed.

Empirical Evidence for Change Across Birth Cohorts

In the United States, a number of studies have produced evidence for the hypothesis that there are generational changes in personality traits. Studies have found birth cohort effects on personality-related variables, such as depression (e.g., Kovacs & Gatsonis, 1994; Lewinsohn, Rohde, Seeley, & Fischer, 1993;

This article was published Online First May 2, 2011.

Iris A. M. Smits and Marieke E. Timmerman, Heymans Institute of Psychology, University of Groningen, Groningen, the Netherlands; Conor V. Dolan, Harrie C. M. Vorst, and Jelte M. Wicherts, Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands.

Correspondence concerning this article should be addressed to Iris A. M. Smits, Heymans Institute of Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS, Groningen, the Netherlands. E-mail: I.A.M.Smits@rug.nl

Ryan, Williamson, Iyengar, & Orvaschel, 1992; Twenge & Nolen-Hoeksema, 2002), assertiveness (e.g., Duncan & Agronick, 1995; Twenge, 2001b), personal and social adjustment (e.g., Woodruff & Birren, 1972), masculine and feminine traits (e.g., Sherman & Spence, 1997; Twenge, 1997), self-esteem (e.g., Duncan & Agronick, 1995; Gentile, Twenge, & Campbell, 2010; Twenge & Campbell, 2001), and narcissism (e.g., Twenge & Foster, 2008; Twenge, Konrath, Foster, Campbell, & Bushman, 2008a, 2008b). Moreover, some studies have found birth cohort effects on personality factors. The relationship between birth cohort and the Big Five personality traits has been addressed in a number of studies, including the studies of Twenge, (2000, 2001a); Sutton-Smith, Rosenberg, and Morgan (1961); Mroczek and Spiro (2003); and Terracciano, McCrae, Brant, and Costa (2005). These studies showed increases over time in Extraversion (Mroczek & Spiro, 2003; Twenge, 2001a) and Conscientiousness (Terracciano et al., 2005) and increases (Sutton-Smith et al., 1961; Twenge, 2000) or decreases (Mroczek & Spiro, 2003; Terracciano et al., 2005) over time in Neuroticism.

An explanation for the differences found in direction of the changes over time in Neuroticism may lie in the fact that these studies concerned populations that differed in age. The studies of Twenge (2000) and Sutton-Smith and colleagues (1961) were based on samples of college students (Twenge, 2000, Study 1) and children, ages 9 to about 14 (Sutton-Smith et al., 1961; Twenge, 2000, Study 2), whereas the studies of Mroczek and Spiro (2003) and Terracciano and colleagues (2005) were based on samples of adults with mean ages around 65. These differences in age group of the samples might have contributed to differences in direction of the changes over time in Neuroticism. Nevertheless, the studies of Twenge, (2000, 2001a), Sutton-Smith and colleagues, Mroczek and Spiro, and Terracciano and colleagues all found evidence for modest to large birth cohort effects on some personality factors. These effects were interpreted by Twenge (2000, 2001a) as generational effects and by Terracciano (2010) as either generational or period effects. The studies of Trzesniewski and Donnellan (e.g., Donnellan & Trzesniewski, 2009; Donnellan, Trzesniewski, & Robins, 2009; Trzesniewski & Donnellan, 2009, 2010; Trzesniewski, Donnellan, & Robins, 2008a, 2008b), on the other hand, studying cohort effects on personality-related variables such as narcissism and self-esteem, found no evidence for the hypothesis that there may are generational changes in personality traits. These contradictory results provided the impetus for a debate between Twenge (e.g., Twenge & Campbell, 2010) and Trzesniewski and Donnellan (2010) on the reasonableness of the evidence for generational changes in personality traits. This debate emphasized the importance both of establishing changes over time in personality factors and of doing so in a methodologically sound way.

Investigating Cohort Effects

Studies of cohort effects are characterized by specific methodological and design challenges. First, studying cohort effects requires a time-lag study, which involves the examination of successive cohorts of people of the same age but different birth years (Schaie, 1965). This implies a study that may have to span many years, depending on the desired time interval between the cohorts and the desired number of cohorts. Simple cross-sectional designs are not well suited due to the confounding of age and cohort (Nesselroade & Baltes, 1974). Second, it is important to establish whether a given personality test does indeed measure the same latent variables (e.g., neuroticism, anxiety) in the successive cohorts. This can be established by showing that the test is measurement invariant with respect to cohorts.

Measurement Invariance

A test is measurement invariant if it measures the same latent trait across groups. This implies that two persons in different groups with the same latent trait must have the same expected score on a test (Mellenbergh, 1989). In the case of measuring cohort differences in personality traits like Extraversion, this implies that the affirmative response to a given Extraversion item (e.g., "I like to meet new people") should be equally attractive among testees who are in different cohorts but who are identically extravert, as expressed on the latent variable Extraversion. Thus, measurement invariance must be established to be able to interpret cohort differences in terms of the latent variables of interest.

Issue of Measurement Invariance in Studying Cohort Effects

The issue of measurement invariance is often ignored, and cohort-related changes in observed scores on a test are simply interpreted as manifestations of changes in the latent variables, which the test is designed to measure. For instance, from the finding that "self-reports of anxiety/neuroticism have increased substantially from the 1950s to the early 1990s" (Twenge, 2000; p. 1017), Twenge (2000) inferred that "the larger sociocultural environment . . . has a considerable effect on a major personality trait" (p. 1017). However, bearing in mind that successive cohorts may be separated by many years, one may ask whether an item administered to young adults in the 1950s and early 1990s will necessarily measure the same construct. For instance, the response to an item purporting to measure Conscientiousness, such as "It is important to dress to the occasion," is likely to be dictated by the sociocultural perceptions of what is appropriate with respect to dress code. Changes in the responses to such an item may have more to do with changes in such perceptions than with the latent variable Conscientiousness per se.

In addition to specific item content, cohort-related differences in response styles may also be a source of bias. For instance, Twenge (e.g., 2000, 2001a) warned that cohort differences might be due to respondents' willingness to describe themselves in terms of the item content due to changes in the social desirability of the personality trait. Such tendencies, which exert a general influence on the responses, are known as response styles (e.g., Cheung & Rensvold, 2000; Messick, 1991; van Herk, Poortinga, & Verhallen, 2004). They will violate measurement invariance, in so far as the cohort-related changes in response styles are largely unrelated to changes in the latent variables of interest. However, the violations may be harder to identify and interpret than violations due to changes in item content; one expects them to exert a general influence and not necessarily to be associated with specific item content. That is, these violations can only be detected if they are expressed in a number or a subset of items; response tendencies that exert an equal influence on all items cannot be examined.

To justify the interpretation of differences in observed test scores as attributable to differences with respect to the latent variables that the tests purport to measure, one has to establish that the test is measurement invariant or unbiased with respect to cohort. This can be done using any suitable psychometric (measurement) model, such as the well-known item response theory models (e.g., Embretson & Reise, 2000) or the linear factor model (e.g., Mellenbergh, 1994). Measurement invariance or unbiasedness is considered to be an important measurement ideal that is attainable by reasonable approximation (see, e.g., Meredith, 1993). However, we note that violations of measurement invariance or unbiasedness may be informative in their own right. For instance, if one can identify items as being biased with respect to cohort, one can study the item content to gain insight into the possible sources of the bias. In addition, the subset of items that do satisfy measurement invariance may still afford a solid basis for the interpretation of cohort differences in terms of the latent variables (see Byrne, Shavelson, & Muthén, 1989).

Notwithstanding its importance to the interpretation of cohort differences in personality, we are not aware of any investigation of measurement invariance of personality test with respect to cohort. We suspect that this is largely attributable to the fact that raw data are required to carry out the appropriate analyses. Understandably, given the demands of time-lag cohort studies, such data are rare. Therefore many researchers have relied on published summary statistics, which do not contain sufficient information to fit the desired measurement models.

Overview

The aim of the present study was to study cohort changes in the Big Five factors Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness, as measured by the Dutch instrument Five Personality Factors Test (5PFT; Elshout & Akkerman, 1975). The 5PFT is among the oldest Big Five scales in the world. The sample consisted of male and female first-year psychology students, ages 18 to 25 years. The study spanned a period from 1982 to 2007. The terms *measurement invariance* and *unbiasedness* are used interchangeably; henceforth, for the sake of clarity, we refer to the term *measurement (in)variance* in relation to the test, and we refer to the term *(un)biasedness* in relation to the items.

In this article, we first present the results of the analysis toward the tenability of measurement invariance with respect to cohort separately in males and females. We did not pool the data in view of possible item bias with respect to sex (see, e.g., Stommel, Given, Given, & Kalaian, 1993). Next, we discuss the effects of the biased items on the interpretation of differences in cohort on the Big Five personality factors. Then, we present the results of the effect of cohort on the Big Five personality factors. Subsequently, we present the results of secondary analysis on data from 1971 to 2007 of females and males of different ages together. Finally, we discuss the value of establishing measurement invariance in interpreting cohort differences in personality. On the basis of previous findings in samples of young adults, we expected an increase over time in Extraversion and Neuroticism. We had no explicit expectations about the direction of possible cohort differences in Agreeableness, Conscientiousness, and Openness to Experience.

Method

Participants

From 1971 onward, all freshman psychology students (N = 12,479) at the University of Amsterdam (Amsterdam, the Netherlands) completed the 5PFT (Elshout & Akkerman, 1975). The 5PFT is a personality questionnaire designed to measure the Big Five personality factors. Students completed this questionnaire together with other tests and questionnaires during a group testing program as part of course requirements. The 5PFT was administered over these years in unaltered form and under similar circumstances. Our main analyses are limited to the cohorts of 1982 onward because between 1971 and 1981, the sex and/or age of participants were often not recorded or are no longer available. Data of students' sex and age are required to be able to analyze the data for male and female students separately and to be able to select students of a narrow age range. Barring 1987, the data from 1982 to 2007 are almost complete and comprise the students' 5PFT scores, sex, and age (N = 10,864). The data of 1987 (350 cases; 3.2%) were excluded from the analysis due to missing information on age and sex. To obtain cohorts of sufficient sizes for the purpose of the present analyses, we formed cohorts by aggregating the data from two successive years, which yielded a total of 13 cohort groups. To ensure that the cohorts did not differ in age range, we retained only the data of students in the age range of 18 to 25 years. This resulted in a total sample size of N = 9,070, with a mean age of 20.18 years.

Although the test was administered in an examlike setting with corresponding discipline, there was no absolute control over the students during testing. To achieve high quality control, we excluded cases on the basis of the following conservative exclusion criteria: Students with more than half of the items missing (i.e., 36 items or more) were excluded (excluded: 21 cases; 0.2%). Furthermore, students with extreme response tendencies were excluded (excluded: 88 cases; 1%).[1] Finally, students missing information concerning sex were excluded from analysis (excluded: 7 cases; 0.1%). Following this selection, the resulting total sample consisted of N = 8,954 students, of whom 2,752 were male and 6,202 were female. To assess possible differences in age across cohort and sex, age groups were submitted to analyses of variance (ANOVAs). On average, students have become slightly younger over the years—main effect of cohort on age, $F(12, 8928) = 17.7$, $p < .001$, partial $\eta^2 = .023$—and female students were slightly younger on average than male students—main effect of sex on age, $F(1, 8928) = 145.8$, $p < .001$, partial $\eta^2 = .016$; $M_{\text{male}} = 20.59$, 95% CI [20.52, 20.66]; $M_{\text{female}} = 20.00$, 95% CI [19.95, 20.05].

[1] A student was identified as having an extreme response tendency when he or she met the following two criteria of outliers: (a) For the raw scores, the frequencies of scores per response category were recorded per factor and per student. A student was considered to be an outlier if his or her frequency in a particular response category was larger than 1.5 times the interquartile range (IQR; i.e., 1.5 × IQR) beyond the quartiles of the frequencies, on all five factors. (b) For the mirrored scores, the standard deviations of scores were computed across the 14 items, per factor and per subject. A student was considered to be an outlier if his or her standard deviation was larger than 1.5 × IQR beyond the quartiles of the standard deviations across subjects per factor, on one or more of the five factors.

However, we consider these effects of no concern, as the effect sizes are quite small. No Sex × Cohort interaction on age was found, $F(12, 8928) = 0.9$, $p = .54$. Table 1 contains a breakdown of the total sample of our main analyses over sex and cohort.

The data between 1971 and 1981 were used in secondary analyses in which we assessed the trend on data from 1971 to 2007 of females and males of different ages together. The secondary analyses could not be restricted to the target population of young adults, nor was it possible to analyze these data for male and female students separately because data of subjects' sex and/or age were often missing in the records between 1971 and 1982. Subjects were excluded if they did not meet the exclusion criteria as in our main analyses: Students with more than half of the items missing were excluded (excluded: 28 cases; 0.2%). Furthermore, students with extreme response tendencies were excluded (excluded: 111 cases; 0.9%; see footnote 1). The resulting total sample consisted of $N = 12{,}340$ students. We formed cohorts by aggregating the data from two successive years, as we did in our main analyses, which yielded a total of 17 cohort groups (designated Cohorts a–q in this article).

To consider whether possible selection effects could have influenced the results, we considered the stability of our population. Data of admission policies and admission criteria indicated that there were no significant changes in admission policies and admission criteria for studying psychology in the period from 1982 to 2007, the period explored in the main analyses. A diploma from the highest level in secondary school (called preuniversity education or VWO) is sufficient to enroll at Dutch universities, and there were no additional requirements to enroll in psychology. There have been slight modifications in the enrollment criteria for a minority of students who do not have a VWO degree, but these mostly concern age requirements that hardly affected our analyses because the data considered in our primary analysis were restricted in age range. In addition, there has been a slight increase over the years in the number of students who already have completed another college degree at lower levels than university (i.e., bachelor-type college degree). However, the number of students who do so is still fairly small. Moreover, these students are typically older students, and our analyses were restricted to those below the age of 25.

Universities in the Netherlands have comparable reputations and consequently score quite similarly in various global university rankings. The reputation of the University of Amsterdam does not appear to have changed since 1982. Universities in the Netherlands have the same admission policies and nationally regulated tuitions that are relatively modest compared to those at high-ranked universities in the United States. Before 1986, students from lower and middle socioeconomic status backgrounds received financial aid from the Dutch government. Since 1986, all students in the Netherlands receive such financial aid, the height of which remains dependent on parental income. The modal height of the financial aid has decreased somewhat since the second half of the 1990s, which has supposedly led to an increase in the number of students who have paid jobs alongside their education. At the same time, the regulations concerning study progress have become more stringent, so that a lack of progress over the course of study means that the financial aid will be changed to a loan. However, these issues hardly affect the constitution of the cohorts in our study because these were composed of students who participated early in their freshmen year, that is, at a time when progress was equal for all.

The number of students enrolled in psychology at the University of Amsterdam has increased over the years, thereby suggesting a gain in popularity of the topic as well as a general population trend toward increasing enrollment at universities. Census data from Statistics Netherlands (Centraal Bureau voor de Statistiek, 2010) indicate that there has been an increase in enrollment at universities from 181,982 students in 1990 (the earliest year with available data) to 212,713 students in 2007 (a 16.9% increase). This increase is almost entirely due to an increase in gross enrollment at universities among females (a 39.6% increase, compared to 0.6% for males). Enrollment in psychology at all Dutch universities has increased by around 80% during the same time, with females showing a considerably larger increase over this time span (a 96.4% increase) compared to males (47.0%). So, the growth in enrollment at the University of Amsterdam reflects both a general trend among the wider population and a more specific trend that indicates that psychology has gained in popularity (especially among females). To conclude, data of external factors including admission policies and admission criteria suggest that there have been no significant changes in the population studied.

Test

The Big Five personality factors were measured using a Dutch instrument, the 5PFT, developed by Elshout and Akkerman (1973, 1975; see Elshout, 1999). As discussed by Wicherts and Vorst (2010), the 5PFT is related to the NEO Personality Inventory questionnaire that was developed by Costa and McCrae (1985) in the mid-1980s.

The 5PFT consists of 70 items, 14 for each of the factors Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Culture. Culture is also referred to as Intellect or Openness to Experience. We use the latter term, as this one is commonly used

Table 1
Breakdown of Total Sample by Cohort and Sex (Age Range in All Cells: 18–25 Years)

	Cohort													
Breakdown	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
Year	'82 & '83	'84 & '85	'86 & '88	'89 & '90	'91 & '92	'93 & '94	'95 & '96	'97 & '98	'99 & '00	'01 & '02	'03 & '04	'05 & '06	'07	
N males	100	134	180	255	313	234	207	258	241	223	281	203	123	2,752
N females	156	246	390	529	686	428	431	594	731	506	657	544	304	6,202
N total	256	380	570	784	999	662	638	852	972	729	938	747	427	8,954

in contemporary literature. Each item consists of a short description of some behavior. The participants rate, on a 7-point Likert-type scale, the degree to which they judge the descriptive as applicable to them. Cronbach's alphas of the five personality factors of the 5PFT range from .77 to .87 (Elshout & Akkerman, 1973, 1975; Reumerman, 1993). Wicherts and Vorst (2010) found that the 5PFT scales showed good convergent validity with the Dutch NEO Personality Inventory—Revised (Hoekstra, Ormel, & de Fruyt, 1996) in a sample of psychology freshmen ($N = 500$): Extraversion $r = .76$, Neuroticism $r = .82$, Agreeableness $r = .71$, Openness to Experience $r = .55$, and Conscientiousness $r = .71$.

Analyses

Here, we describe the procedure of our main analyses. The procedure of our secondary analyses is comparable to that of the main analyses, with the exceptions that we could not restrict these analyses to subjects of a narrow age range and that we could not conduct these analyses separately for males and females. First, we investigated whether measurement invariance with respect to cohort was tenable. To address the issue of measurement invariance, we fitted increasingly constrained 13-group one-common-factor models to the 14 items of each Big Five factor. We did this for males and females separately because item bias with respect to gender might have been present and because the gender composition of the cohort was subject to change, with the percentages of females increasing from 61% to 71% over the 25-year period. Measurement invariance with respect to sex and possible Sex × Cohort interactions are beyond the scope of the present article. To reduce computational burden and model complexity, we fitted the factor models separately to each of the five sets of 14 items.[2] This approach rests on the assumption that each item within the set of 14 items is related only to the presumed factor, not to any of the four remaining factors.

Examining measurement invariance. Before explaining the methodological procedure, we first introduce the approach we used to examine measurement invariance and present the specific model we applied (for an extensive literature on measurement invariance in the common-factor model, see Byrne et al. 1989; Horn & McArdle, 1992; Little, 1997; Meredith, 1993; Millsap & Everson 1991; for literature on measurement invariance in the field of personality, see, e.g., Chen, 2008).

Measurement invariance with respect to group implies that the relation between the observed test scores and the latent trait scores is the same across the groups. Those relations are expressed via the parameters of a measurement model. A test is said to be measurement invariant if the parameters of the measurement model are equal across groups. Possible differences between groups are captured in the latent trait scores. If measurement invariance holds, differences in observed scores can be attributed to differences between groups in the trait of interest. The tenability of measurement invariance can be studied by comparing the fit of the models with and without the restriction that parameters are equal across groups. The constraints are introduced in a sequence of analyses to be able to evaluate the changes in goodness of fit that are associated with the successive constraints. This is a common strategy for identifying specific sources of misfit. The successive constraints are summarized in Table 2 and are now described briefly. Subsequently, the goodness-of-fit measures are discussed.

Configural invariance model. The first step in establishing measurement invariance is fitting the model without constraints. The model without constraints is also called the configural invariance model. In this study, the unconstrained model is the 13-group one-common-factor model, for each of the five factors. We refer to this model as Model A in the tables below. The common-factor model is a linear regression model in which the scores on several items are regressed upon the scores on the latent trait (e.g., Extraversion). For each item, the measurement parameters of the common-factor model are a factor loading (λ), a residual term (ε), and an intercept (τ). In the configural invariance model, the factor loadings, residual variances, and intercepts of the items are allowed to vary across groups.

Metric invariance model. In the first step toward a measurement invariance model, we constrain the factor loadings to be equal across cohorts ($\Lambda = \Lambda_1 = \ldots = \Lambda_{13}$), which results in the metric invariance model. We refer to this model as Model B in the tables below. The requirement of equal factor loadings is a necessary condition for a test to be measurement invariant across groups. In establishing measurement invariance, often the main (if not exclusive) focus is on the equality of the factor loadings (Wicherts & Dolan, 2010). However, this equality constraint is insufficient to satisfy full measurement invariance (Mellenbergh, 1989; Meredith, 1993).

Equal residual variances. In the next model in the sequence, we add equality constraints on covariance matrices of the residuals $\Theta = \Theta_1 = \ldots = \Theta_{13}$). We refer to this model as Model C in the tables below. Thus, in the second step toward a measurement invariance model, both the factor loadings and the residual variances are restricted to be equal across groups.

Strict factorial invariance model. Adding equality constraints on the intercepts results in the final model, which is called the strict factorial invariance model ($\tau = \tau_1 = \ldots = \tau_{13}$). We refer to this model as Model D in the tables below. Intercepts have to be equal across groups to ensure that the test scores do not systematically undervalue or overestimate the subjects' latent ability in some of the groups. The strict factorial invariance model satisfies Mellenbergh's (1989) definition of measurement invariance (Meredith, 1993). So, if this model holds, we maintain that biasedness with respect to cohort is absent, and we may interpret the observed cohort differences as attributable to the differences in the underlying common factor.

Evaluation of model fit. We introduced the constraints in a sequence of analyses as outlined in the previous section. Each model was evaluated by considering a number of fit indices.

Fit indices. We assessed the fit of the models by considering the root-mean-square error of approximation (RMSEA; Browne & Cudeck, 1993; Steiger, 1990) and the sample-size adjusted Bayesian information criterion (ABIC; see Muthén & Muthén, 1998–2007). The prevailing convention is that an RMSEA value of .08 or smaller is indicative of acceptable fit (e.g., Schermelleh-Engel, Moosbrugger, & Müller, 2003). The ABIC provides a rank ordering of models: the lower the ABIC, the better the model fit.

[2] It is easier to detect model violation in a 13-group (cohort) analysis of 14 items, which should satisfy a one-factor model, than in a 13-group (cohort) analysis of 70 items in a five-factor model.

Table 2
Steps of Equality Constraints Across Groups

Model	Description	Factor loadings	Residual variances	Intercepts
A	Configural invariance	Λ free	Θ free	τ free
B	Metric invariance	Λ invariant	Θ free	τ free
C	Equal residual variances	Λ invariant	Θ invariant	τ free
D	Strict factorial invariance	Λ invariant	Θ invariant	τ invariant

Note. Cohort 1 is the reference group. Throughout Models B–D, we allowed factor variances to differ across Cohorts 2–13. In addition, in Model D, we allowed factor means to differ across Cohorts 2–13.

For the first model and for the nested models for which the more parsimonious model results in a clear deterioration in model fit, we inspected the modification indices[3] to determine the cause of the misfit. A modification index of 20 or greater is viewed as a cause for concern. We added these parameters to the model if they resulted in appreciable improvement in fit and were readily interpretable.

Secondary fit indices. We also report the chi-square, degrees of freedom (*df*), and the ratio χ^2/df (see Schermelleh-Engel et al., 2003). Because the models in the sequence (described above) are nested, we also report log-likelihood differences ($\Delta\chi^2$), Δdfs, and ratio of these ($\Delta\chi^2/\Delta df$). The chi-square and also the χ^2/df ratio are reported for completeness, but it is generally recognized that these have little value as formal test statistics. Specifically, trivial model approximation error will readily result in large values, especially if the total sample size is large (as is the case here). We did not consider incremental fit indices such as the Tucker-Lewis index or the comparative fit index, which entail a comparison of the model with a model in which all items are uncorrelated. The use of such standard comparative fit measures may be problematic in models with mean structure (Widaman & Thompson, 2003), as we have here.

Methodological procedure. We conducted the analysis as outlined in the previous section. First, we fitted the configural invariance model. We inspected the modification indices to determine whether there were any correlated residuals. We added these correlated residuals to the model if they resulted in appreciable improvement in fit and were readily interpretable in terms of item content. Next, we fitted the metric invariance model, and we constrained the residual covariance matrices to be equal over cohort. Finally, we fitted the strict factorial invariance model. We assessed the fit of the models by considering the RMSEA and the ABIC. If the progression from a given model (say, Model A) to a more parsimonious model (say, Model B) was accompanied by a clear deterioration in model fit, we inspected the modification indices associated with the latter to determine the cause of the misfit.

We carried out the analyses separately for the male and female students and separately for the five sets of 14 items, resulting in 10 analyses. The item responses were given on 7-point Likert-type scales. They included a very small number of missing data (mean percentage missing values ranged from 0.1% to 2.8%) and were approximately normal, as indicated by the skewness and kurtosis values (skewness values ranged from −.7 to .6, and mean kurtosis values ranged from −.8 to .5).[4] We fitted the models using (normal theory) maximum-likelihood estimation in the Mplus program (Muthén & Muthén, 1998–2007). Other estimation procedures are available (Wirth & Edwards, 2007), but Dolan (1994) established by means of a simulation study that maximum-likelihood estimation performs adequately for 7-point scales, provided that they are not too skewed (for an overview of maximum-likelihood estimation performances, see Forero & Maydeu-Olivares, 2009).

Results

Measurement Invariance With Respect to Cohort

We now present the results of our main analyses, starting with the analyses to assess measurement invariance for each personality factor.

Extraversion. First, we fitted the configural invariance model (Model A; see Table 3). This model did not fit adequately in either the male or the female sample (RMSEAs = .101 and .103 in male and female samples, respectively). By inspecting the modification indices, we detected four correlated residuals, which we judged to be interpretable in the light of the item content. For instance, in one case, the correlated residual concerned the items "Behaves in a low-keyed, modest manner" and "Likes to be the center of attention." We consider these items to be almost mirror images of each other. We attribute the correlation among these residuals to method effects. That is, the excess correlation was due to using virtually the same method (in the present case, items with highly comparable item content) to measure the construct. These correlated residuals were detected in several cohorts in both the male and female samples, which suggests that they are a structural feature of certain items. Adding these four correlated residuals to each of the 13 factor models resulted in appreciable improvement in model fit (Model A2: RMSEAs = .080 and .083 in male and female samples, respectively).

Subsequently, we added, in succession, the constraints of equal factor loadings (Model B), equal residual covariance matrices (Model C), and equal intercepts (Model D) across groups. Judging by the fit indices in Table 3, we consider the strict factorial invariance model (Model D) to be acceptable in both samples (RMSEAs = .070 and .072 in male and female samples, respectively). Moreover, this model was associated with the smallest ABIC in both samples. Evaluating all goodness-of-fit indices, we conclude that the strict factorial model fits to reasonable approximation in both male and female samples. Therefore, we conclude that the Extraversion subtest is measurement invariant with respect to cohort in both the male and female samples.

Agreeableness. First, we fitted the configural invariance model (Model A; see Table 4). This model did not fit adequately (RMSEAs = .093 and .094 in male and female samples, respectively). Inspecting the modification indices, we detected three correlated residuals. These were observed in several cohorts and in

[3] The modification index provides an indication of the change in chi-square that may be expected if a single parameter restriction is relaxed.

[4] Detailed information on the distribution of the scores is available on request to Iris A. M. Smits.

Table 3
Model Fit Results for 14 Extraversion Items

Model	χ^2	*df*	χ^2/df	$\Delta\chi^2$	Δdf	$\Delta\chi^2/\Delta df$	ABIC	RMSEA
Male								
A. No constraints	3,182.0	1,001	3.2				128,970.6	0.101
A2. No constraints[a]	2,231.7	949	2.4	950.3	52	18.3	128,267.0	0.080
B. Λ equal	2,402.6	1,105	2.2	170.9	156	1.1	127,698.0	0.074
C. Λ Θ equal	2,722.6	1,321	2.1	319.9	216	1.5	126,993.5	0.071
D. Λ Θ τ equal	3,002.4	1,477	2.0	279.8	156	1.8	126,533.4	0.070
Female								
A. No constraints	6,037.3	1,001	6.0				281,351.5	0.103
A2. No constraints[a]	4,059.5	949	4.3	1,977.8	52	38.0	279,662.5	0.083
B. Λ equal	4,260.8	1,105	3.9	201.3	156	1.3	278,997.3	0.077
C. Λ Θ equal	4,740.1	1,321	3.6	479.3	216	2.2	278,276.7	0.074
D. Λ Θ τ equal	5,133.5	1,477	3.5	393.4	156	2.5	277,803.5	0.072

Note. ABIC = sample-size adjusted Bayesian information criterion; RMSEA = root-mean-square error of approximation.
[a] Four off-diagonals freed.

both samples. We judged these to be interpretable in light of the item content. For example, one pair of items concerned "Trusting, having faith in people" and "Having little faith in people, cynical." Barring the negation of the latter item, these are very similar in content. Allowing these residuals to correlate resulted in an improvement in model fit (Model A2: RMSEAs = .078 and .077 in male and female samples, respectively).

Subsequently, we added, in succession, the constraints of equal factor loadings (Model B), equal residual covariance matrices (Model C), and equal intercepts (Model D) across groups. We judged the strict factorial model (Model D; see Table 4) to fit adequately in both samples (RMSEAs = .072 and .069 in male and female samples, respectively). This model also has the smallest ABIC in the sequence of models in both samples. Evaluating the fit indices, we conclude that in both male and female samples, the Agreeableness subtest is measurement invariant with respect to cohort.

Conscientiousness. First, we fitted the configural invariance model (Model A in Table 5). This model did not fit adequately (RMSEAs = .114 and .111 in male and female samples, respectively). Inspecting the modification indices, we detected four correlated residuals, which we judged to be interpretable in the light of the item content. For instance, one pair of items referred to persevering and persisting on the one hand and to a lack of persistence and persevering on the other. A second item pair referred to neat and tidy on the one hand and to sloppiness on the other. So, again, we judged the item content to be very similar in content. These four correlated residuals were detected in several cohorts in both samples. Allowing these residuals to correlate resulted in an improvement in model fit (Model A2: RMSEAs = .090 and .087 in male and female samples, respectively).

Subsequently, we added, in succession, the constraints of equal factor loadings (Model B), equal residual covariance matrices (Model C), and equal intercepts (Model D) across groups. The strict factorial model (Model D; see Table 5) did not fit well in either the male or the female sample. Specifically, the step from Model C to Model D resulted in a general deterioration in fit (see Table 5). By inspecting the modification indices associated with the intercepts, we detected, in some cohorts, two offending items in the male sample and six offending items in the female sample. The item content is shown in Table 6. Allowing for between-group differences in these items, the models fitted adequately (see Table 5; RMSEAs = .081 and .076 in male and female samples, respectively; the models have the smallest ABIC in both samples). Thus,

Table 4
Model Fit Results for 14 Agreeableness Items

Model	χ^2	*df*	χ^2/df	$\Delta\chi^2$	Δdf	$\Delta\chi^2/\Delta df$	ABIC	RMSEA
Male								
A. No constraints	2,838.6	1,001	2.8				122,107.8	0.093
A2. No constraints[a]	2,193.1	962	2.3	645.5	39	16.6	121,647.2	0.078
B. Λ equal	2,399.8	1,118	2.1	206.7	156	1.3	121,114.0	0.074
C. Λ Θ equal	2,771.7	1,322	2.1	372.0	204	1.8	120,518.5	0.072
D. Λ Θ τ equal	3,105.5	1,478	2.1	333.8	156	2.1	120,112.4	0.072
Female								
A. No constraints	5,253.7	1,001	5.2				262,897.4	0.094
A2. No constraints[a]	3,686.7	962	3.8	1,567.0	39	40.2	261,547.0	0.077
B. Λ equal	3,911.3	1,118	3.5	224.6	156	1.4	260,905.1	0.072
C. Λ Θ equal	4,311.2	1,322	3.3	399.9	204	2.0	260,171.8	0.069
D. Λ Θ τ equal	4,840.2	1,478	3.3	529.0	156	3.4	259,834.2	0.069

Note. ABIC = sample-size adjusted Bayesian information criterion; RMSEA = root-mean-square error of approximation.
[a] Three off-diagonals freed.

Table 5
Model Fit Results for 14 Conscientiousness Items

Model	χ^2	*df*	χ^2/df	$\Delta\chi^2$	Δdf	$\Delta\chi^2/\Delta df$	ABIC	RMSEA
Male								
A. No constraints	3,744.6	1,001	3.7				132,272.0	0.114
A2. No constraints[a]	2,574.4	949	2.7	1,170.1	52	22.5	131,348.5	0.090
B. Λ equal	2,780.3	1,105	2.5	205.8	156	1.3	130,814.4	0.085
C. Λ Θ equal	3,062.5	1,321	2.3	282.2	216	1.3	130,072.2	0.079
D. Λ Θ τ equal	3,626.5	1,477	2.5	564.0	156	3.6	129,896.4	0.083
D2. Λ Θ τ equal[b]	3,470.7	1,453	2.4	155.8	24	6.5	129,854.3	0.081
Female								
A. No constraints	6,936.1	1,001	6.9				288,378.0	0.111
A2. No constraints[a]	4,410.8	949	4.6	2,525.3	52	48.6	286,141.5	0.087
B. Λ equal	4,582.7	1,105	4.1	171.9	156	1.1	285,446.9	0.081
C. Λ Θ equal	4,995.4	1,321	3.8	412.7	216	1.9	284,659.7	0.076
D. Λ Θ τ equal	6,100.4	1,477	4.1	1,105.0	156	7.1	284,898.1	0.081
D2. Λ Θ τ equal[c]	5,237.4	1,405	3.7	863.0	72	12.0	284,435.1	0.076

Note. ABIC = sample-size adjusted Bayesian information criterion; RMSEA = root-mean-square error of approximation.
[a] Four off-diagonals freed. [b] Two intercepts freed. [c] Six intercepts freed.

due to two biased items in the male sample and six biased items in the female sample, the test for Conscientiousness in its original form is not measurement invariant with respect to cohort. However, the Conscientiousness subtest in the male and female samples is measurement invariant with respect to cohort if these biased items are taken into account.

Table 6
Uniformly Biased Items

Factor	Item content
Male	
Extraversion	None
Agreeableness	None
Conscientiousness	Item: Casual, will do things when it suits me. May forget an appointment.
	Item: Thinks it is important to dress to fit the occasion.
Neuroticism	None
Openness to Experience	None
Female	
Extraversion	None
Agreeableness	None
Conscientiousness	Item: Thinks it important to abide by prevailing norms.
	Item: Thinks it is important to dress to fit the occasion.
	Item: Responsible, puts general interest above personal interest.
	Item: Tidy, everything has its place.
	Item: Casual, will do things when it suits me. May forget an appointment.
	Item: Principle, stick to one's word.
Neuroticism	Item: Worrisome, worries a lot.
	Item: Feeling down, melancholic.
	Item: Emotionally stable, even tempered.
	Item: Often very emotional (angry, sad, elated, in love, afraid, etc.).
	Item: Occupied with own health; thinks regularly that something is wrong.
Openness to Experience	None

Note. Items are taken from the Five Personality Factors Test (5PFT) and have been translated from Dutch into English by us. The copyright on the 5PFT is held by Libbe Mulder, who has granted us permission to reproduce these translated test items in this article.

Neuroticism. First, we fitted the configural invariance model (see Table 7; Model A). This model did not fit adequately (RMSEAs = .101 and .102 in male and female samples, respectively; see Table 7). Inspecting the modification indices, we detected four correlated residuals. In the light of the item content, the correlated residuals are comprehensible, as the relevant items appear to be quite closely interrelated. In one case, one pair of items concerns nervousness on the one hand and tension on the other. A second pair concerns emotional stability on the one hand and being emotional on the other. The correlated residuals were detected in several cohorts and in both samples. Allowing these residuals to correlate resulted in an improvement in model fit (Model A2: RMSEAs = .084 and .087 in male and female samples, respectively).

Subsequently, we added, in succession, the constraints of equal factor loadings (Model B), equal residual covariance matrices (Model C), and equal intercepts (Model D) across groups. In the male sample, we judge the strict factorial invariance model (Model D) to be acceptable in the light of the fit measures (RMSEA = .079; the model was associated with the smallest ABIC; see Table 7). We conclude that the Neuroticism subtest is measurement invariant with respect to cohort in the male sample.

In the female sample, the restriction of equal intercepts across groups was accompanied by a drop in model fit indicated by the RMSEA (see Table 7). We identified five biased items in some cohorts. The item content is shown in Table 6. Relaxing the equality constraints on the intercepts of these items resulted in a reasonably well-fitting model (RMSEA = .076), which was also associated with the smallest ABIC.

Openness to Experience. First, we fitted the configural invariance model (see Table 8; Model A). This model did not fit adequately (RMSEAs = .101 and .100 in male and female samples, respectively; see Table 8). Inspecting the modification indices, we identified four correlated residuals. Again, these correlated residuals seemed quite comprehensible in the light of the item content. For instance, one item pair related reading

Table 7
Model Fit Results for 14 Neuroticism Items

Model	χ^2	*df*	χ^2/df	$\Delta\chi^2$	Δdf	$\Delta\chi^2/\Delta df$	ABIC	RMSEA
Male								
A. No constraints	3,148.1	1,001	3.1				126,907.0	0.101
A2. No constraints[a]	2,373.2	949	2.5	774.9	52	14.9	126,378.8	0.084
B. Λ equal	2,608.6	1,105	2.4	235.4	156	1.5	125,874.2	0.080
C. Λ Θ equal	3,023.4	1,321	2.3	414.9	216	1.9	125,264.7	0.078
D. Λ Θ τ equal	3,420.9	1,477	2.3	397.4	156	2.5	124,922.2	0.079
Female								
A. No constraints	5,924.5	1,001	5.9				283,315.8	0.102
A2. No constraints[a]	4,381.5	949	4.6	1,543.0	52	29.7	282,061.6	0.087
B. Λ equal	4,573.2	1,105	4.1	191.7	156	1.2	281,386.8	0.081
C. Λ Θ equal	4,999.7	1,321	3.8	426.5	216	2.0	280,613.5	0.076
D. Λ Θ τ equal	5,837.6	1,477	4.0	837.9	156	5.4	280,584.8	0.079
D2. Λ Θ τ equal[b]	5,291.8	1,417	3.7	545.8	60	9.1	280,372.2	0.076

Note. ABIC = sample-size adjusted Bayesian information criterion; RMSEA = root-mean-square error of approximation.
[a] Four off-diagonals freed. [b] Five intercepts freed.

little on the one hand and being cultivated and reading a lot on the other. Here, we consider the reference to reading to be the source of the covariance among the residuals. The modification indices were observed in several cohorts and in both samples. As shown in Table 8, allowing these residuals to correlate resulted in an improvement in model fit (Model A2: RMSEAs = .084 and .080 in male and female samples, respectively).

Subsequently, we added, in succession, the constraints of equal factor loadings (Model B), equal residual covariance matrices (Model C), and equal intercepts (Model D) across groups. Judging by the fit indices in Table 8, we consider the strict factorial invariance model (Model D) to be acceptable in both samples (RMSEAs = .076 and .070 in male and female samples, respectively). Moreover, this model was associated with the smallest ABIC in both samples. Evaluating the goodness-of-fit indices, we conclude that the Openness subtest is measurement invariant with respect to cohort in both the male and female samples.

The Effect of Biased Items on the Trend

From the results above, we conclude that some of the items of the factors Conscientiousness and Neuroticism display uniform bias, especially in the female sample. Removal of the biased items leaves a sufficient number of unbiased items to create interpretable sum scores. Now, an important question is how these biased items affect the trend over time. To address this issue, we conducted repeated measures ANOVAs with biasedness as a within-subject factor (two levels: 14 item vs. unbiased item versions) and cohort as a between-subject factor (13 levels) for the factors in which bias was present: Conscientiousness in the male and female samples and Neuroticism in the female sample. We were interested in the Cohort $\times$ Biasedness interaction because we wanted to know how the biased items affect the trend over time. The main effects of biasedness are not considered here because they reflect only the difference in number of items on which the sum scores are based. As shown in Table 9, the interactions were consistently significant, but the effect sizes were very small for Conscientiousness in males (η^2 = .013) and Neuroticism in females

Table 8
Model Fit Results for 14 Openness Items

Model	χ^2	*df*	χ^2/df	$\Delta\chi^2$	Δdf	$\Delta\chi^2/\Delta df$	ABIC	RMSEA
Male								
A. No constraints	3,151.1	1,001	3.1				123,693.8	0.101
A2. No constraints[a]	2,367.0	949	2.5	784.1	52	15.1	123,156.3	0.084
B. Λ equal	2,559.3	1,105	2.3	192.3	156	1.2	122,608.8	0.079
C. Λ Θ equal	2,879.1	1,321	2.2	319.8	216	1.5	121,904.1	0.075
D. Λ Θ τ equal	3,303.3	1,477	2.2	424.2	156	2.7	121,588.4	0.076
Female								
A. No constraints	5,768.3	1,001	5.8				274,043.3	0.100
A2. No constraints[a]	3,820.8	949	4.0	1,947.5	52	37.5	272,384.6	0.080
B. Λ equal	4,050.9	1,105	3.7	230.1	156	1.5	271,748.2	0.075
C. Λ Θ equal	4,415.2	1,321	3.3	364.3	216	1.7	270,912.7	0.070
D. Λ Θ τ equal	4,970.8	1,477	3.4	555.6	156	3.6	270,601.7	0.070

Note. ABIC = sample-size adjusted Bayesian information criterion; RMSEA = root-mean-square error of approximation.
[a] Four off-diagonals freed.

Table 9
Main and Interaction Effects of Biasedness and Biasedness × Cohort on Mean Scores

Factor	*F* value	$df_{hypothesis}$	df_{error}	*p* value	Effect size (partial η^2)
Conscientiousness					
Male					
Biasedness	31,581.915	2,739	1	<.0005	.920
Biasedness × Cohort	3.006	2,739	12	<.0005	.013
Female					
Biasedness	156,253.223	6,189	1	<.0005	.962
Biasedness × Cohort	37.494	6,189	12	<.0005	.068
Neuroticism					
Female					
Biasedness	73,891.895	6,189	1	<.0005	.923
Biasedness × Cohort	3.216	6,189	12	<.0005	.006

($\eta^2 = .006$) and slightly larger, but still quite modest, for Conscientiousness in females ($\eta^2 = .069$). To obtain an impression of how the trends actually differed, see Figures 1 and 2.

Cohort Effects in Personality Factors

Taking biased items into account, what can we say about changes in personality factors over time? Is personality affected by a broader sociocultural factor, as indexed by cohort? The mean scores in the cohorts from 1982 to 2007 are depicted for males and females in Figures 1 and 2, respectively. For each sex and each Big Five factor, we performed a one-way ANOVA on the unbiased items with cohort as a between-subject factor (13 levels). Given an alpha of 0.001, our analyses revealed small main effects of cohort on Agreeableness, $F_{male}(12, 2739) = 6.33$, $p < .001$, partial $\eta^2 = .027$; $F_{female}(12, 6189) = 15.30$, $p < .001$, partial $\eta^2 = .029$; Conscientiousness, $F_{male}(12, 2739) = 7.07$, $p < .001$, partial $\eta^2 = .030$; $F_{female}(12, 6189) = 27.40$, $p < .001$, partial $\eta^2 = .050$; and Neuroticism, $F_{male}(12, 2739) = 3.03$, $p < .001$, partial $\eta^2 = .013$;

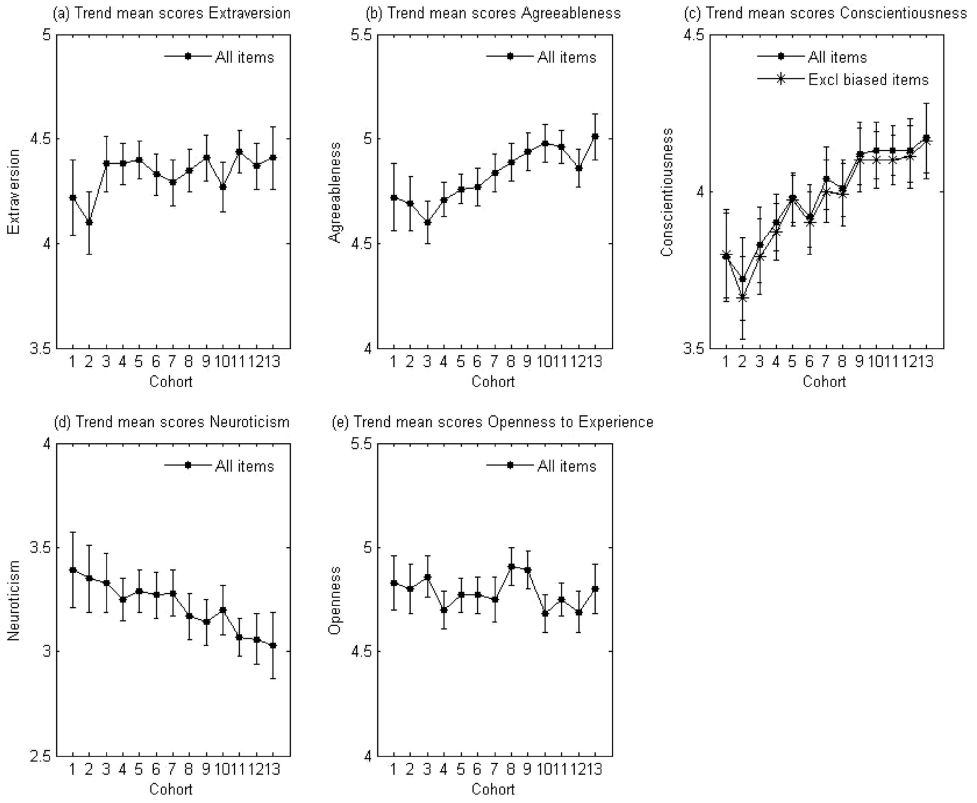

Figure 1. Male mean trends of mean scores with 95% confidence intervals (represented by error bars) in 13 cohorts from 1982 (Cohort 1) to 2007 (Cohort 13), based on all items and on excluding (Excl) biased items. Biased items are detected only in the Conscientiousness items.

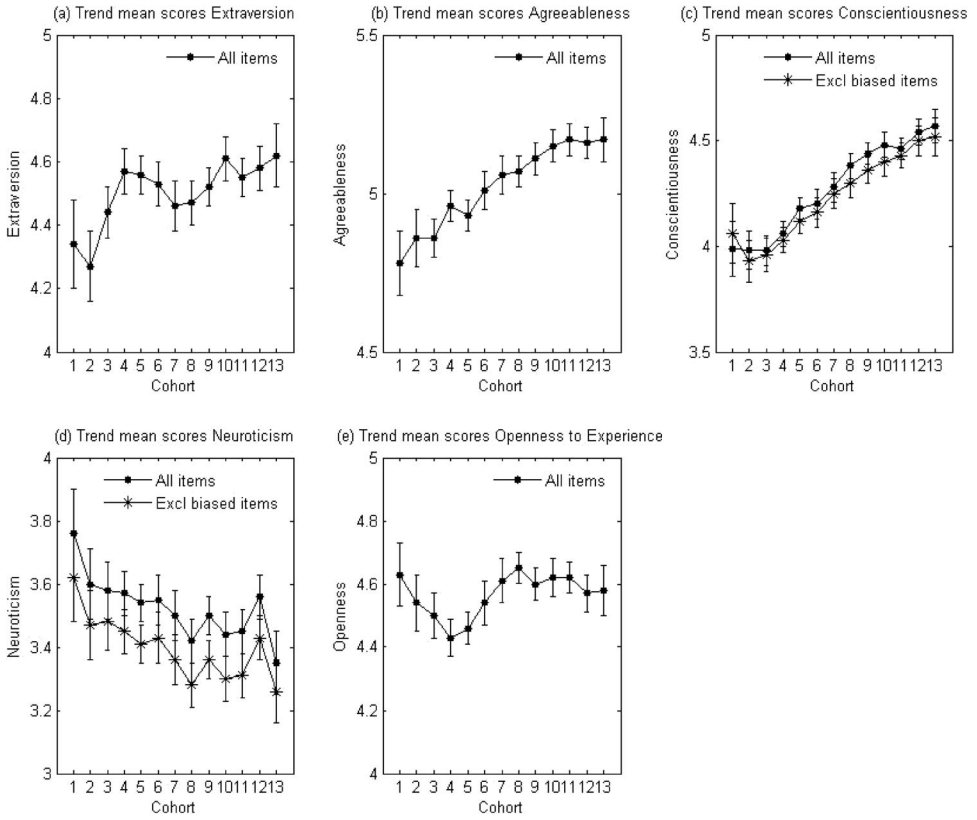

Figure 2. Female mean trends of mean scores with 95% confidence intervals (represented by error bars) in 13 cohorts from 1982 (Cohort 1) to 2007 (Cohort 13), based on all items and on excluding (Excl) biased items. Biased items are detected only in the Conscientiousness and Neuroticism items.

$F_{female}(12, 6189) = 4.17$, $p < .001$, partial $\eta^2 = .008$. In the female sample but not in the male sample, we found, given an alpha of 0.001, small main effects of cohort on Extraversion, $F_{female}(12, 6189) = 4.84$, $p < .001$, partial $\eta^2 = .009$; $F_{male}(12, 2739) = 1.96$, $p = ns$, and Openness, $F_{female}(12, 6189) = 5.13$, $p < .001$, partial $\eta^2 = .010$; $F_{male}(12, 2739) = 2.39$, $p = ns$. Testing polynomial contrasts up to the fourth order, we found, given an alpha of 0.01, linear trends over time in the Extraversion ($p_{female} < .01$, $p_{male} < .01$), Agreeableness ($p_{female} < .01$, $p_{male} < .01$), Conscientiousness ($p_{female} < .01$, $p_{male} < .01$), and Neuroticism ($p_{female} < .01$, $p_{male} < .01$) factors. In the Openness to Experience factor, we observed significant ($p < .01$) third-order polynomial contrasts in the female sample and no significant contrast in the male sample.

Trend Cohort Effects Findings 1971–2007

To get an indication of the trend over time in personality factors over the complete time span, we conducted secondary analyses of the data from 1971 to 2007. Because of the missing information on students' sex and age, these analyses could not be restricted to the target population of young adults, nor was it possible to analyze these data for male and female students separately. Therefore, the analyses were conducted on data from 1971 to 2007 of females and males of different ages together.

We now present the results of the analyses to assess measurement invariance for the five personality factors. The same procedure was applied as in the main analyses. First the configural invariance model (Model A) was fitted. Subsequently, the constraints of equal factor loadings (Model B), equal residual covariance matrices (Model C), and equal intercepts (Model D) across groups were added. For all five factors, the configural invariance model (Model A) did not fit adequately. Inspecting the modification indices, we detected around seven correlated residuals in each of the five personality factors (6, 7, 7, 8, and 7 correlated residuals in the factors Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness, respectively), which we judged to be interpretable in the light of the item content. Adding these correlated residuals to each of the 17 factor models resulted in appreciable improvement in model fit (RMSEAs = .075, .071, .082, .077, and .075 in the factors Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness, respectively). Next, we added the constraints of equal factor loadings, equal residual covariance matrices, and equal intercepts across groups. Judging the fit indices, we consider the strict factorial invariance model (Model D) to be acceptable in the Extraversion subtest (RMSEA = .067; smallest ABIC). In the four other factors, the strict factorial model did not fit well. Specifically, the step from Model C to Model D (the succeeding addition of equal intercepts) resulted in a general deterioration in fit. By inspecting the modification indices associated with the intercepts, we detected six, nine, seven, and nine offending items in the factors Agreeableness, Conscientiousness, Neuroticism, and Openness, respectively. Allowing between-group

differences in these items, the models fitted adequately (RMSEAs= .062, .070, .067, and .066 in the factors Agreeableness, Conscientiousness, Neuroticism, and Openness, respectively).

Taking biased items into account, what can we say about changes in personality factors over the complete time span? In Figure 3, the mean scores on the unbiased items in the cohorts from 1971 to 2007 are presented. Conducting ANOVAs on the unbiased items (with $\alpha = .001$), the analyses revealed a small main effect of cohort on all factors: Extraversion, $F(16, 12323) = 28.50$, $p < .001$, partial $\eta^2 = .036$; Agreeableness, $F(16, 12323) = 25.85$, $p < .001$, partial $\eta^2 = .032$; Conscientiousness, $F(16, 12323) = 30.10$, $p < .001$, partial $\eta^2 = .038$; Neuroticism, $F(16, 12323) = 14.95$, $p < .001$, partial $\eta^2 = .019$; and Openness, $F(16, 12323) = 11.47$, $p < .001$, partial $\eta^2 = .015$. The results of the secondary analysis indicate that a linear trend is visible comparable to those in the main analysis (see Figure 3).

Discussion

The aim of this study was to evaluate the trend in the mean personality test scores of the Big Five questionnaire 5PFT (Elshout & Akkerman, 1975) over a period of about 25 years in a Dutch sample of young adults, ages 18 to 25 years. Therefore, we first assessed measurement invariance with respect to cohort over this period. In our analyses of measurement invariance, we identified 11 and two items as uniformly biased with respect to cohort in the female and male samples, respectively. It is notable that we detected the most biased items in the female sample. This is probably due to the fact that the total female sample is appreciably larger than the male sample (6,202 vs. 2,752) and thus confers greater power to detect uniformly biased items.

The changes over time in the responses to the biased items cannot be accounted for adequately by the changes with respect to the latent variables, which the items are supposed to measure. To obtain some idea of why this might be the case, we inspected the item content (see Table 6). We considered the items associated with the Conscientiousness subscales in the males and females, as here the effect of the bias is the largest (see Table 9). With respect to the Conscientiousness items (six in the females, two in the males), we suspect that a number of the items display bias because they are directly or indirectly related to social conventions (e.g., dressing properly). We guess that students in the 1980s were less sensitive to such social conventions than students in the 2000s. This is consistent with the lower degree of increase in Conscientiousness once these items have been removed (see Figures 1 and 2).

Overall, once the biasedness was taken into account by relaxing the equality constraints on the intercepts (having previously added correlated residuals where necessary), the factor models showed reasonable fit. Given that the number of biased items is relatively small (11 and two of the 70 items in the female and male samples,

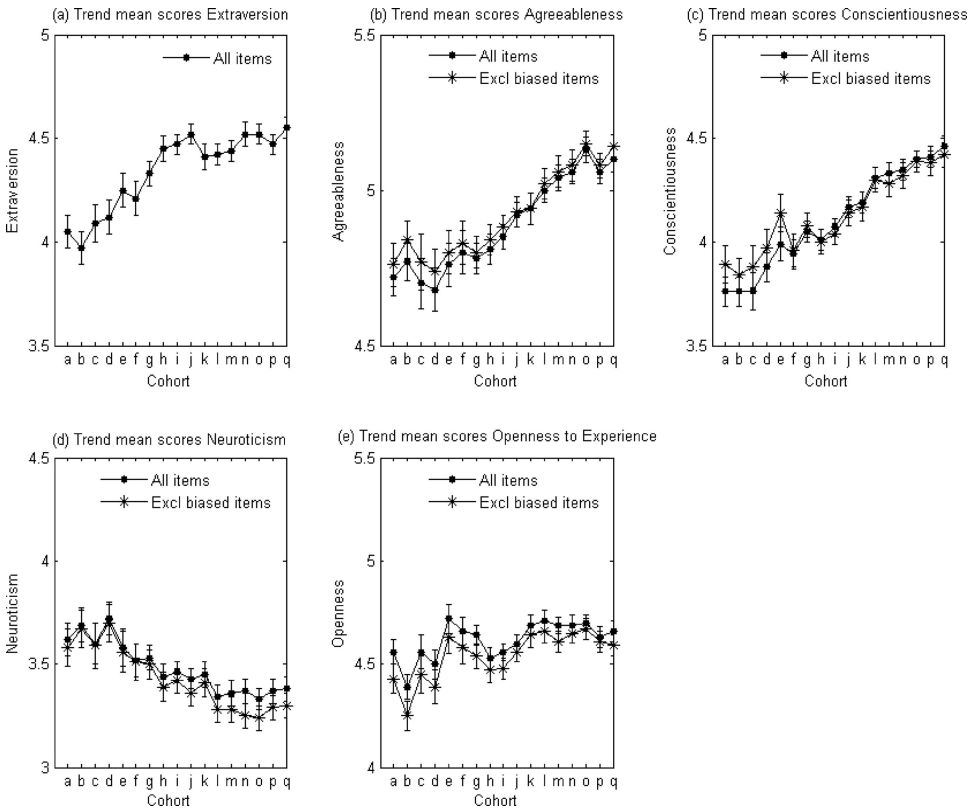

Figure 3. Overall mean trends of mean scores with 95% confidence intervals (represented by error bars) in 17 cohorts from 1971 (Cohort a) to 2007 (Cohort q), based on all items and on excluding (Excl) biased items. Biased items are detected in the Agreeableness, Conscientiousness, Neuroticism, and Openness items.

respectively), we conclude that the 5PFT is largely measurement invariant over cohorts. However, we do believe it is important to take the biased items into account in view of the Biasedness × Cohort interaction. Taking the biased items into account results in a more precise evaluation of cohort differences in terms of the latent dimensions, which the 5PFT purports to measure.

Considering the results in terms of the unbiased items, our results indicate that first-year psychology students between 18 and 25 years old appeared to be more agreeable, more conscientious, and less neurotic in the course of years from 1982 to 2007 (see Figures 1 and 2). The increase in the Extraversion factor is less clear, and no clear result emerged with respect to the Openness to Experience factor. The results of our secondary analyses indicate that a linear trend is visible for Agreeableness, Conscientiousness, Neuroticism, and Extraversion comparable to those in our main analyses if we analyze the data from 1971 to 2007 in a sample of students differing in age and sex (see Figure 3). However, we underline that these results only show a global trend over time because they are confounded by sex and age.

Comparing our results with previous findings in studies examining cohort differences in a sample of young adults, we note that our results are consistent with a previous finding of an increase in Extraversion over time and are inconsistent with a previous finding of an increase in Neuroticism over time (Twenge, 2000). Also, this study indicates much smaller effects of cohort on Extraversion and Neuroticism than previously found in studies examining cohort differences in young adults (Twenge, 2000, 2001a). A first explanation for these differences is that data were collected in different populations. Our study is based on a sample of Dutch college students. The studies of Twenge are based on a sample of American college students (Twenge, 2000, Study 1; Twenge, 2001a). Since cohort effects, being generated by the sociocultural environment, are dependent on culture, the difference between our results and those of previous studies could be a reflection of the cultural differences between the populations studied. We take the view that is very hard, if not impossible, to unravel the mechanisms behind cohort effects in personality traits. However, we notice that the observed cohort effects are embedded in a specific cultural context influenced by general processes seen in the Netherlands in the past 25 years, such as an increasing individualization and considerable changes in the welfare system.

A second explanation for the discrepancies between our and previous studies is that previous studies may have suffered from a lack of measurement invariance. Unfortunately, measurement invariance could not be explored in previous studies in samples of young adults since these data analyses were based on summary statistics and raw data are required to establish measurement invariance.

The analyses of measurement invariance indicate that violations of measurement invariance or unbiasedness can be informative in their own right. With respect to the Conscientiousness items, inspection of the item content of the biased items with respect to cohort indicated that, in particular, items which directly or indirectly referred to social conventions are biased with respect to cohort. This suggests that items of which the item content is more embedded in a sociocultural context are probably more sensitive to being biased with respect to cohort than items of which the item content is more neutrally stated. Such a conclusion is not surprising, but it underlines what is already known: that it is difficult to construct survey valid questions. Moreover, it indicates that prudence is in order in using (old) questionnaires, of which the interpretation of the items may differ between contemporary samples and the original norm samples.

This is the first study exploring cohort shifts in a European population. Strong aspects of this study are the study design and extensive analysis of raw data, including the evaluation of measurement invariance between the different cohorts. The design led to a large data set covering an extensive period of time. The assessment of measurement invariance is an important contribution to earlier work, precluding yet another possible type of attributive error. A limitation of this study is the fact that the sample consists of psychology-class freshmen. We make no attempt to generalize the present results to the general population. However, we have no reason to expect that this sample will differ appreciably from the general population in susceptibility to the broad sociocultural environment.

Summarizing, this study has shown slight but significant effects of the passing years, the changing of sociocultural surroundings, on Big Five personality factors. For the first time in this research field, validity of attribution considering measurement invariance has been established.

References

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466. doi:10.1037/0033-2909.105.3.456

Centraal Bureau voor de Statistiek. (2010). *StatLine* [Data file]. Retrieved from http://statline.cbs.nl

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95,* 1005–1018. doi:10.1037/a0013193

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31,* 187–212. doi:10.1177/0022022100031002003

Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory manual.* Odessa, FL: Psychological Assessment Resources.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47,* 309–326.

Donnellan, M. B., & Trzesniewski, K. H. (2009). How should we study generational "changes"—or should we? A critical examination of the evidence for "Generation Me." *Social and Personality Psychology Compass, 3,* 775–784. doi:10.1111/j.1751-9004.2009.00204.x

Donnellan, M. B., Trzesniewski, K. H., & Robins, R. W. (2009). An emerging epidemic of narcissism or much ado about nothing? *Journal of Research in Personality, 43,* 498–501. doi:10.1016/j.jrp.2008.12.010

Duncan, L. E., & Agronick, G. S. (1995). The intersection of life stage and social events: Personality and life outcomes. *Journal of Personality and Social Psychology, 69,* 558–568. doi:10.1037/0022-3514.69.3.558

Eaves, L. J., Eysenck, H. J., & Martin, N. G. (1989). *Genes, culture and personality: An empirical approach.* San Diego, CA: Academic Press.

Elder, G. H., Jr. (1998). The life course as developmental theory. *Child Development, 69,* 1–12. doi:10.2307/1132065

Elshout, J. J. (1999). De vijf persoonlijkheidsfactoren test (5PFT) 1973–1999 [The Five Personality Factor Test (5PFT), 1973–1999]. *Nederlands Tijdschrift Voor De Psychologie En Haar Grensgebieden, 54,* 195–207.

Elshout, J. J., & Akkerman, A. E. (1975). *Vijf persoonlijkheidsfaktoren test 5PFT: Handleiding* [The Five Personality Factor Test (5PFT): Manual]. Nijmegen, the Netherlands: Berkhout B.V.

Elshout, J. J., & Akkerman, T. E. (1973). Een nederlandse test voor vijf persoonlijkheidsfactoren, de 5 PFT [A Dutch test for five personality factors, the 5PFT]. In P. J. D. Drenth, P. J. Willems, & C. J. de Wolff (Eds.), *Arbeids-, en organisatiepsychologie* (pp. 49–56). Deventer, the Netherlands: Kluwer.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn effect.* New York, NY: Cambridge University Press.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14,* 275–299. doi:10.1037/a0015825

Gentile, B., Twenge, J. M., & Campbell, W. K. (2010). Birth cohort differences in self-esteem, 1988–2008: A cross-temporal meta-analysis. *Review of General Psychology, 14,* 261–268. doi:10.1037/a0019919

Hoekstra, H. A., Ormel, J., & de Fruyt, F. (1996). *Handleiding NEO persoonlijkheids-vragenlijsten NEO-PI-R en NEO-FFI* [NEO Personality Inventories: NEO-PI-R and NEO-FFI—Dutch manual]. Lisse, the Netherlands: Swets Test Services.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144.

Kovacs, M., & Gatsonis, C. (1994). Secular trends in age at onset of major depressive disorder in a clinical sample of children. *Journal of Psychiatric Research, 28,* 319–329. doi:10.1016/0022-3956(94)90014-0

Lewinsohn, P. M., Rohde, P., Seeley, J. R., & Fischer, S. A. (1993). Age-cohort changes in the lifetime occurrence of depression and other mental disorders. *Journal of Abnormal Psychology, 102,* 110–120. doi:10.1037/0021-843X.102.1.110

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32,* 53–76. doi:10.1207/s15327906mbr3201_3

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

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29,* 223–236. doi:10.1207/s15327906mbr2903_2

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

Messick, S. (1991). Psychology and methodology of response styles. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 161–200). Hillsdale, NJ: Erlbaum.

Millsap, R. E., & Everson, H. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research, 26,* 479–497. doi:10.1207/s15327906mbr2603_6

Mroczek, D. K., & Spiro, A. (2003). Modeling intraindividual change in personality traits: Findings from the Normative Aging Study. *Journals of Gerontology: Series B. Psychological Sciences and Social Sciences, 58,* P153–P165. doi:10.1093/geronb/58.3.P153

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Author.

Neisser, U. (1998). *The rising curve: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association. doi:10.1037/10270-000

Nesselroade, J. R., & Baltes, P. B. (1974). Adolescent personality development and historical change: 1970–1972. *Monographs of the Society for Research in Child Development, 39*(1, Serial No. 154), 1–79. doi:10.2307/1165824

Reumerman, R. (1993). *Analyse van de 5PFT op basis van de schendersstheorie* [Analysis of the 5PFT based on violator theory]. Amsterdam, the Netherlands: University of Amsterdam.

Ryan, N. D., Williamson, D. E., Iyengar, S., & Orvaschel, H. (1992). A secular increase in child and adolescent onset affective disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 31,* 600–605. doi:10.1097/00004583-199207000-00004

Ryder, N. B. (1965). The cohort as a concept in the study of social change. *American Sociological Review, 30,* 843–861. doi:10.2307/2090964

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin, 64,* 92–107. doi:10.1037/h0022371

Schaie, K. W., & Elder, G. H., Jr. (Eds.). (2005). *Historical influences on lives & aging.* New York, NY: Springer Publishing Company.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research, 8,* 23–74.

Sherman, P. J., & Spence, J. T. (1997). A comparison of two cohorts of college students in responses to the Male–Female Relations Questionnaire. *Psychology of Women Quarterly, 21,* 265–278. doi:10.1111/j.1471-6402.1997.tb00112.x

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180. doi:173-180.1991-03242-001

Stewart, A. J., & Healy, J. M. (1989). Linking individual development and social changes. *American Psychologist, 44,* 30–42. doi:10.1037/0003-066X.44.1.30

Stommel, M., Given, B. A., Given, C. W., & Kalaian, H. A. (1993). Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). *Psychiatry Research, 49,* 239–250. doi:10.1016/0165-1781(93)90064-N

Sutton-Smith, E., Rosenberg, B. G., & Morgan, E. F., Jr. (1961). Historical changes in the freedom with which children express themselves on personality inventories. *Journal of Genetic Psychology, 99,* 309–315.

Terracciano, A. (2010). Secular trends and personality: Perspectives from longitudinal and cross-cultural studies—Commentary on Trzesniewski & Donnellan (2010). *Perspectives on Psychological Science, 5,* 93–96. doi:10.1177/1745691609357017

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

Trzesniewski, K. H., & Donnellan, M. B. (2009). Reevaluating the evidence for increasingly positive self-views among high school students: More evidence for consistency across generations (1976–2006). *Psychological Science, 20,* 920–922. doi:10.1111/j.1467-9280.2009.02361.x

Trzesniewski, K. H., & Donnellan, M. B. (2010). Rethinking "Generation Me": A study of cohort effects from 1976–2006. *Perspectives on Psychological Science, 5,* 58–75. doi:10.1177/1745691609356789

Trzesniewski, K. H., Donnellan, M. B., & Robins, R. W. (2008a). Do today's young people really think they are so extraordinary? An examination of secular trends in narcissism and self-enhancement. *Psychological Science, 19,* 181–188. doi:10.1111/j.1467-9280.2008.02065.x

Trzesniewski, K. H., Donnellan, M. B., & Robins, R. W. (2008b). Is "Generation Me" really more narcissistic than previous generations? *Journal of Personality, 76,* 903–918. doi:10.1111/j.1467-6494.2008.00508.x

Twenge, J. M. (1997). Changes in masculine and feminine traits over time: A meta-analysis. *Sex Roles, 36,* 305–325. doi:10.1007/BF02766650

Twenge, J. M. (2000). The age of anxiety? Birth cohort change in anxiety

and neuroticism, 1952–1993. *Journal of Personality and Social Psychology, 79,* 1007–1021. doi:10.1037/0022-3514.79.6.1007

Twenge, J. M. (2001a). Birth cohort changes in extraversion: A cross-temporal meta-analysis, 1966–1993. *Personality and Individual Differences, 30,* 735–748. doi:10.1016/S0191-8869(00)00066-0

Twenge, J. M. (2001b). Changes in women's assertiveness in response to status and roles: A cross-temporal meta-analysis, 1931–1993. *Journal of Personality and Social Psychology, 81,* 133–145. doi:10.1037/0022-3514.81.1.133

Twenge, J. M., & Campbell, W. K. (2001). Age and birth cohort differences in self-esteem: A cross-temporal meta-analysis. *Personality and Social Psychology Review, 5,* 321–344. doi:10.1207/S15327957PSPR0504_3

Twenge, J. M., & Campbell, W. K. (2010). Birth cohort differences in the Monitoring the Future dataset and elsewhere: Further evidence for Generation Me—Commentary on Trzesniewski & Donnellan (2010). *Perspectives on Psychological Science, 5,* 81–88. doi:10.1177/1745691609357015

Twenge, J. M., & Foster, J. D. (2008). Mapping the scale of the narcissism epidemic: Increases in narcissism 2002–2007 within ethnic groups. *Journal of Research in Personality, 42,* 1619–1622. doi:10.1016/j.jrp.2008.06.014

Twenge, J. M., Konrath, S., Foster, J. D., Campbell, W. K., & Bushman, B. J. (2008a). Egos inflating over time: A cross-temporal meta-analysis of the Narcissistic Personality Inventory. *Journal of Personality, 76,* 875–902. doi:10.1111/j.1467-6494.2008.00507.x

Twenge, J. M., Konrath, S., Foster, J. D., Campbell, W. K., & Bushman, B. J. (2008b). Further evidence of an increase in narcissism among college students. *Journal of Personality, 76,* 919–928. doi:10.1111/j.1467-6494.2008.00509.x

Twenge, J. M., & Nolen-Hoeksema, S. (2002). Age, gender, race, socioeconomic status, and birth cohort difference on the Children's Depression Inventory: A meta-analysis. *Journal of Abnormal Psychology, 111,* 578–588. doi:10.1037/0021-843X.111.4.578

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35,* 346–360. doi:10.1177/0022022104264126

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

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence, 32,* 509–537. doi:10.1016/j.intell.2004.07.002

Wicherts, J. M., & Vorst, H. C. M. (2010). The relation between specialty choice of psychology students and their interests, personality, and cognitive abilities. *Learning and Individual Differences, 20,* 494–500. doi:10.1016/j.lindif.2010.01.004

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8,* 16–37. doi:10.1037/1082-989X.8.1.16

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58–79. doi:10.1037/1082-989X.12.1.58

Woodruff, D. S., & Birren, J. E. (1972). Age changes and cohort difference in personality. *Developmental Psychology, 6,* 252–259. doi:10.1037/h0032086

Received March 4, 2010
Revision received December 13, 2010
Accepted January 13, 2011 ■