A Stable Relationship Between Personality and Academic Performance from Childhood through Adolescence. An original study and replication in hundred thousandperson samples

Andersen, Simon Calmar; Gensowski, Miriam; Ludeke, Steven; John, Oliver P.

A Stable Relationship Between Personality and Academic Performance from Childhood through Adolescence.
An original study and replication in hundred-thousand-person samples

Simon Calmar Andersen*
TrygFondens Centre for Child Research
Department of Political Science
Aarhus University
Bartholins alle 7
8000 Aarhus C
Denmark
sca@ps.au.dk
phone: +45 61 66 65 01
facsimile: +45 8613 9839

Miriam Gensowski
Department of Economics
University of Copenhagen
Øster Farimagsgade 5
1353 Copenhagen K.
Denmark
Miriam.Gensowski@econ.ku.dk

Steven Ludeke
Department of Psychology
University of Southern Denmark
Campusvej 55
5230 Odense M
Denmark
lud@health.sdu.dk

Oliver P. John
Department of Psychology
University of California, Berkeley
2121 Berkeley Way
Berkeley, CA 94720-1650
USA
o_johnx5@berkeley.edu

* Corresponding author

A Stable Relationship Between Personality and Academic Performance from Childhood through Adolescence.
An original study and replication in hundred-thousand-person samples

Abstract

Objective: Many studies have demonstrated that personality traits predict academic performance for students in high school and college. Much less evidence exists on whether the relationship between personality traits and academic performance changes from childhood to adolescence, and existing studies show very mixed findings. This study tests one hypothesis—that the importance of Agreeableness, Emotional Stability and Conscientiousness for academic performance changes fundamentally during school—against an alternative hypothesis suggesting that the changing relationships found in previous research are largely measurement artifacts.

Method: We used a nationwide sample of 135,389 primary and lower-secondary students from grade 4 to grade 8. We replicated all results in a separate sample of another 127,375 students.

Results: We found that academic performance was equally strongly related to our measure of Conscientiousness at all these grade levels, and the significance of Agreeableness and Emotional Stability predominantly reflected their connections with Conscientiousness. However, age also appeared to shape the relationship between Emotional Stability and performance.

Conclusion: Amidst the replication crisis in psychology these findings demonstrate a very stable and predictable relationship between personality traits and academic performance, which may have important implications for the education of children already in primary school.

Introduction

The importance of high school and university students' personality traits for their academic performance is documented in a number of reviews and meta-analyses. Especially Conscientiousness is strongly related to performance, whereas other personality traits are less so (Almlund et al., 2011; Poropat, 2009; Trapmann et al., 2007; Vedel, 2014). Much less is known about whether younger children's different personality traits are equally important for their academic progress. In a highly cited meta-analysis, Poropat (2009) found that the relationship between personality and academic performance changed fundamentally from late childhood (around 11 years) to adolescence (around 16 years): Agreeableness and Emotional Stability were of more or less the same importance as Conscientiousness in late childhood, but then, during adolescence, they lost almost all of their relevance while Conscientiousness remained important despite a small decrease. Understanding how personality traits relate to academic performance already in childhood is not just of theoretical interest, but also highly relevant to educators and parents that care to know which children most likely need additional support to develop their full academic potential.

There could be theoretical reasons to expect a real change in the relationship between educational requirements and children's personality. For example, schooling may become more like a job as children get older—less focused on joyful play-based learning and more on diligent work, which requires high levels of Conscientiousness. If the early school grades are more play-based, then children who are socially competent (high Agreeableness) and emotionally stable may perform well. As the educational context changes from play-based learning to more diligent work, Agreeableness and Emotional Stability may become less important, whereas Conscientiousness may retain or even increase its importance (Poropat, 2009). We refer to this as the *age-trait explanation*.

However, as with most other meta-analyses, Poropat's (2009) literature review considered only one trait at a time and used different samples and study designs to estimate how the relationship

between personality and academic performance changed as children became older. Accordingly, there may be methodological explanations for apparent changes in the relative importance for academic performance of the three personality traits for which we have data in this study: Agreeableness, Emotional Stability and Conscientiousness. As discussed further below, the comparative paucity of elementary-aged samples available to Poropat (2009) highlights the possibility that these differences are measurement artifacts stemming from different samples and different measures of personality and academic performance.

One methodological explanation is that it is simply more difficult to distinguish different parts of the personality in younger children. De Pauw, Mervielde, and Van Leeuwen's (2009) comparison of Five Factor Models to models of temperament in preschool children found six components with only partial overlap with the Five Factor Model, which suggest that in younger children a Five Factor structure has not fully emerged. Soto and colleagues (2008) found that between the age of 10 and 14 respondents increasingly differentiated Agreeableness and Emotional Stability from Conscientiousness. The average inter-scale correlation dropped substantially between ages 10 and 14 for Agreeableness (approximately .37 to .20) and Conscientiousness (approximately .38 to .23), with a more modest improvement (approximately .38 to .30) for Emotional Stability (which the authors of that study discuss using the name of the trait's reversed form, Neuroticism).[1] Based on the results in Soto et al. (2008), it could be that Conscientiousness is the single most important personality trait for academic performance in both childhood and adolescence, and that the apparently high correlations academic performance has with Agreeableness and Emotional Stability in late childhood may be due to difficulties in distinguishing these traits from Conscientiousness in younger children. This alternative expectation makes sense because key facets of Conscientiousness, such as self-discipline, focus, and persistence, should facilitate learning and academic progression at all

---

[1] The numerical values presented here are inexact as they are derived by examining Figure 4 in Soto et al. (2008).

educational levels. We refer to this explanation for any changing relationship across age groups or grade levels as the *measurement artifact explanation*.

Unfortunately, few studies have directly tested the relationship between personality traits and academic performance in young students, let alone any moderating effect of age. Even though Poropat (2009) was able to include more than 70,000 students in his meta-analysis, only eight studies (with a total of 3,196 students) examined the relationship between personality and academic performance in primary school. Furthermore, as we will show in the next section, in the years since Poropat's meta-analysis very little additional evidence has accumulated for students in primary school. More studies that incorporate students representing a range of ages are needed, as relying on meta-analyses to address developmental issues has important limitations. Meta-analyses can, on the one hand, compensate for the limitations of individual studies in terms of insufficient data and statistical power to test moderation hypotheses such as the hypothesis that the relationship between personality and academic performance changes as students become older. On the other hand, meta-analyses are limited in the sense that they often must rely on different measurements across studies and different sample inclusion criteria in the individual studies. In other words, sampling, design, and measurement differences may be inherently confounded with age differences in the meta-analyses. These inherent differences make it difficult to distinguish heterogeneous subgroup effects from differences across measures and study samples. It is possible that the reason that Poropat (2009) found a changing relationship between personality and academic performance may be because studies of older students were based on other samples with different characteristics rather than because of a changing effect of personality.

In sum, further evidence is needed on (a) whether Conscientiousness, Agreeableness and Emotional Stability predict academic performance in primary school, (b) whether they are stronger predictors of academic performance in childhood than in adolescence, and (c) whether developmental

changes in measurement (i.e., the degree to which Agreeableness and Emotional Stability are confounded with Conscientiousness) play a role in these relationships. Based on existing evidence it is premature to conclude that Conscientiousness is relatively less important for younger children compared to other traits, and that it therefore deserves less attention in younger children from parents and educators. In light of the replication crisis in psychology (Klein et al., 2018; Open Science Collaboration, 2015) and Rozin's (2009, 435) call for studies of fundamental psychological phenomena, "such as functional relations that apply to the real world and have generality," what is needed in order to establish the relationship between personality and academic performance in younger children is large samples and *large replication samples* of students that differ in age, but are sampled in the same way and studied with the same measures across age groups (see also Funder et al., 2014).

The Relationship between Personality Traits and Academic Performance

The Five-Factor Model, also called "The Big Five," is the dominant model for representing the core areas of personality. The model represents the primary areas of personality differences as encapsulated by five broad domains: Agreeableness, Conscientiousness, Emotional Stability, Extraversion, and Openness. Due to limitations in available measures, we focus in this study on the first three.[2] Agreeableness relates to people's tendency to be cooperative and empathetic towards other people. Conscientiousness assesses aspects of personality that relate to the degree to which people are responsible and work carefully to get things done. Emotional Stability describes absence of characteristics such as anxiety and vulnerability to stress.

---

[2] We use items from an assessment that was not designed with the Big Five constructs in mind. In validation work conducted prior to analyses for the present study, we found, however, that we could use the available items to create meaningful measures of Conscientiousness, Emotional Stability, and Agreeableness, though not Extraversion and Openness. See further discussion in Method section.

Whereas especially Openness also predicts academic performance, Conscientiousness often turns out to be the strongest predictor (Almlund et al., 2011; Trapmann et al., 2007; Vedel, 2014). In his meta-analysis, Poropat (2009, p. 328) estimated the correlation between Conscientiousness and academic performance to be .28 in primary school (around 11 years) and .21 in secondary education (around 16 years). Overall, Conscientiousness was at least twice as strongly correlated with academic performance as any of the other Big Five traits, and it was of equivalent magnitude as the correlation between IQ and academic performance (which had a correlation of .23). In contrast, correlations between academic performance and Agreeableness dropped from .30 in primary school to .05 in secondary school. Correlations between academic performance and Emotional Stability dropped from .20 in primary school to .01 in secondary school.

Little empirical evidence on the moderating role of age or educational level on the relationship has accumulated since the publication of Poropat's (2009) meta-analysis, and even less evidence is available to help distinguish between the age-trait hypothesis and the measurement artifact hypothesis. According to ISI Web of Knowledge (on December 14, 2017), 482 articles have cited Poropat's (2009) meta-analysis. 70 of those are new publications of the relationship between personality and academic performance. However, only three studies include students in primary education (about 6-12 years old), and of these only one also includes students from secondary school, which means that there is very little new basis for evaluating the age- or educational level-moderation hypothesis.

Table 1 summarizes the findings in these three studies. The largest sample was found in Zupančič et al. (2016) with 1,618 students spanning from the age of 8 to 15. Neuenschwander et al. (2013) had the youngest sample ranging from 6 to 9 years ($M = 8.1$, N=446). Medford and McGeown (2012) included 295 students between 9 and 11 years. The studies reported far from all correlations and significance tests between the traits themselves and with academic performance. The reported

correlations between academic performance and each of the traits were generally smaller than what Poropat (2009) found. Correlations with Conscientiousness ranged between .10 and .19, compared to Poropat's .28 in primary school. Correlations with Agreeableness ranged from -.06 to .09, which is much smaller than Poropat's .30. Correlations with Neuroticism (the reverse of Emotional Stability) ranged from -.04 to -.15, which is also smaller than Poropat's .20. However, the correlations between the traits were relatively high, wherever reported, ranging between .31 and .44. These high inter-trait correlations could indicate that the high correlations that Poropat found for Agreeableness and Emotional Stability may be confounded by correlations with Conscientiousness.[3]

[Table 1 here]

The changing relationship between traits and academic performance may also relate to the measurement of academic performance. Many studies use teacher-based assessments such as grade point averages to measure academic performance. If teachers' grading is influenced not only by academic performance, but also by the behavior and personality of the students, then changes in the relationship between grades and personality traits as students become older may reflect, for instance, that in primary school teachers value students that are high in Agreeableness and Emotional Stability, or at least more so than they do in secondary school, where teachers' may put more emphasis on work habits related to Conscientiousness. This highlights a potential advantage of using objective assessments rather than teacher-based criteria such as grades when evaluating the associations between personality and academic performance across different ages: Any apparent change in these associations across ages would more clearly reflect shifts in how personality contributes to learning

---

[3] Out of the 482 studies citing Poropat (2009), only two directly tested the age-trait hypothesis. However, they did not include primary school students. Results for the age-trait hypothesis were mixed and inconsistent (Caprara et al., 2011; Zhang and Ziegler, 2016).

outcomes, as opposed to age-related changes in what traits teachers find desirable in students. At the same time, there is some reason to think the use of objective assessments would result in lower associations between personality and academic performance: Neuenschwander et al. (2013) found that both Conscientiousness and Emotional Stability had stronger correlations with grades than with standardized achievement tests (see also Chamorro-Premuzic & Furnham, 2006).[4] Different academic performance measures thus reflect different trade-offs: The use of objective assessments such as standardized tests may provide more straightforwardly interpretable results across ages, but these relationships could potentially be attenuated compared to teacher grades.

Similarly, the changing relationship between traits and academic performance may relate to the measurement of the traits. Larger inter-trait correlations in younger children may be due to limited capacity to accurately report about their own traits. Poropat (2014, p. 247) noted the possibility that "children's self-ratings are affected by their academic ability, with children who are performing better at school being more able to accurately comprehend and respond to personality assessments and in turn being quicker to recognize how best to describe themselves for self-presentation purposes." Younger children may therefore be unable to reliably respond to questions about their own personality and that may conflate different personality traits in younger children's self-reports.

Studies concerned with this issue have sought to compare how the personality-performance relationship is affected by who is providing the personality assessment. Zupančič et al. (2016) compared mother- and self-rated personality measures. They generally found stronger correlations with the grades for mother- than self-rated measures. However, Poropat (2014, 2016) found in a new meta-analysis that differences between self- and other-reported personality measures varied in their correlation with academic performance between students in primary and secondary/tertiary

---

[4] Studies of intelligence also support the notion that the measurement of academic performance matters – though, interestingly, in a way that exhibits results opposite those for personality: Intelligence seems more able to predict academic performance when that performance is measured objectively (e.g., Deary, Strand, Smith, & Fernandes, 2007) rather than when using teacher grades (Poropat, 2009).

education. Especially, self-reported Agreeableness showed a stronger correlation with academic performance in primary education than other-reported, but vice versa in secondary/tertiary education.

In sum, Poropat (2009) found that Agreeableness and Emotional Stability were correlated equally as much with academic performance in late childhood as Conscientiousness, but much less so in adolescence. This result is uncertain for several reasons. Individual studies were typically underpowered for assessing differences in the personality-performance relationship across ages, and the modest number of studies looking at younger populations exhibited consequential heterogeneity in the measurement of both personality and academic performance across these studies. Meta-analytic methods cannot easily overcome such heterogeneity with such a limited pool of primary data, pointing to the need for new work with larger samples and a single study method across ages.

Additional considerations: The replication crisis

The degree to which psychological research can produce replicable findings has become one of the central concerns of the field. The issue of greatest concern is, of course, the rate at which psychological findings replicate in any form and to any degree. Evidence to date on this issue is extremely dismaying, with even studies published in highly-regarded journals failing to replicate as often as not (Klein et al., 2018; Open Science Collaboration, 2015). But accompanying the failure rate is equally dismal news: Even many of the studies with replications that were ostensibly "successful" had replication effect sizes that were considerably more modest, with approximately three-quarters of the studies explored by both the Open Science Collaboration and Many Labs 2 projects showing lower effect sizes in the replication study (Klein et al., 2018; Open Science Collaboration, 2015).

Fortunately, failures to replicate are not random, as previous research has identified several features that increase a study's likelihood of replicating. Successful replication is more likely for well-

powered studies that have a relatively straightforward and easily copied design, and for studies where the effects do not appear particularly counterintuitive (Open Science Collaboration, 2015; see also Funder et al., 2014). Furthermore, studies that use correlational designs in the course of focusing specifically on the estimation of effect sizes rather than the presence or absence of statistically significant effects might be expected to replicate relatively well – and, crucially, to produce effect sizes that are replicable (Tackett et al., 2017). Next we describe how the present study meets these criteria.

The Present Study

The two different explanations for the developing relationship between personality traits and academic performance lead to three different expectations.

First, based on the existing evidence, especially Poropat's (2009) meta-analysis, one potential hypothesis is the age-trait explanation, which suggests that the bivariate correlations between academic performance and both Agreeableness and Emotional Stability will be smaller for older students in grade 8 than younger students in grade 4 because the influence of personality traits on academic abilities changes fundamentally as students become older.

A second alternative hypothesis is a methodological explanation, following from Soto and colleagues' (2008) findings in their study of children and adolescents' self-reported personality. The measurement artifact explanation suggests that Agreeableness and Emotional Stability are correlated more strongly with Conscientiousness in 10-year-olds than in 14-year-old students. If this explanation is correct, Conscientiousness should be an equally strong predictor of academic performance throughout schooling, and once controlling for Conscientiousness, correlations between academic performance and Agreeableness and Emotional stability will stay at lower, but stable levels.

Alternative methodological explanations relate to the measurement of the traits (where students' self-reports may be influenced by their ability to accurately comprehend and respond to personality assessments). If self-report capabilities play an important role, evidence of this might be found in comparing scale reliability coefficients for different student populations, such as younger vs older students, and higher- versus lower-performing students.

To test and compare these explanations, we need data on first and foremost students in primary school. As mentioned, very little evidence exists on the relationship between personality and academic performance among these young students. Secondly, to study any different relationships between younger and older students and whether such relationships could partly be driven by changes in the correlations between traits themselves, we need a very large sample of students to be confident that inadequate statistical power is not the reason for any insignificant age-personality differences. Furthermore, we need data that include the full range of students in terms of academic performance in order to make sure that lack of age-personality differences are not due to range restrictions on the academic performance measure. We also need to use comparable measures of both personality and academic performance across students from late childhood to early adolescence to make sure that changes in associations are not driven by changes in measurement tools. To rule out the explanation relating to teachers' grading of the students, we need standardized tests which are blind to the personality traits and other characteristics of the students. Finally, we would like to replicate any results on a separate sample with a separate data collection to in order to establish whether what we find may be what Rozin (2009) might call fundamental, general functional relationship between personality and educational performance already at the age of 10 and onwards.

We used a large, national survey, "The Danish Wellbeing Survey" (DWS), from 2016 on more than 100,000 public-school students in Denmark in grades 4-8, that is, a sample size that exceeds the number of participants available for Poropat's (2009) combined meta-analytic sample. Through a

number of validation analyses described below, we were able to identify eight questions in this survey that measure Conscientiousness, Emotional Stability, and (somewhat less effectively) Agreeableness. Furthermore, we replicated our analyses on a separate sample of another more than 100,000 students in a data collection one year later in 2017.[5]

We were able to connect the survey data with computer scored and population-normed measures of reading performance in grades 4, 6, and 8 (corresponding to approximately 10, 12, and 14 years of age). The IT-based scoring of the test scores ensured that they reflected objective academic performance and not factors that confound the interpretation of academic grades such as negative teacher opinions based on classroom misbehavior. Administrative register data collected by the authorities further allowed us to merge the student data with information about their parents. This provided us with a highly reliable measure of parents' education and immigrant status, which we use to examine rates of non-response to the DWS. The combined data allowed us to estimate with high precision how strongly personality traits predict academic performance for students in grade 4 through 8.

For research purposes, Statistics Denmark has anonymized the data and stored them on a secured server. The Danish Data Protection Agency, Statistics Denmark and the Ministry of Education have granted access to the data for this research project.[6]

---

[5] Students in grade 1-3 respond to a shorter survey. The validity and reliability of self-reported personality measures for children younger than about 10 years are questionable (Borgers, De Leeuw & Hox, 2000; De Leeuw, 2011) and some of the items used here were not completed by the younger children, so we did not use these data in this study. The survey was introduced in 2015, but reflecting challenges in the roll-out of the survey response rates were lower in the first year, so we use the 2016 and 2017 data instead. Further, since students complete the reading performance tests every other year, students taking the exams in 2015 will (unless they have graduated or switched out of public schools) also be represented in the 2017 sample, whereas the 2016 and 2017 samples consist of different students.
[6] Researchers seeking to reproduce and extend these results with this data should apply with Statistics Denmark for access. Statistical code to reproduce all results will be provided to researchers through the Statistics Denmark system.

Method

Participants

The target population consisted of all students in public schools in Denmark in grades 4, 6 and 8 in 2016 and 2017, for the main and replication studies, respectively. Response rates to the DWS were 92.5% (2016) and 92.7% (2017). Non-respondents included 7.6 (6.4) percent more boys in 2016 (2017), 3.5 (2.8) percentage points more immigrants, and their parents' education was .19 (.19) SD below those of the respondents in 2016 (2017). Their reading skills were .50 (.47) SD below the respondents in 2016 (2017). However, even though the non-respondents were disadvantaged in several ways, the study sample represented the full range of students on all these dimensions. To illustrate this, Figure 1 shows the number of respondents and non-respondents by reading scores. The figure shows that even in the lowest-performing group (Group 1, with reading scores more than 1.75 of a SD below the mean), the study sample included almost 5,000 students.[7] So even though non-respondents were more likely to come from this low-reading skill group, there was still a very large number of low-reading-ability students that were included in the study sample.

[Figure 1 here]

Data

Danish authorities collect administrative data for separate purposes (e.g., taxation). All citizens and legal residents have a personal identification number, which enabled us to merge the different datasets. The personal identification number reveals the individual's gender, which we could use to compare boys and girls in our analyses. Permissions and monitoring pertaining to ethical and legal use of the data is obtained by registering the project with The Danish Data Protection Agency

---

[7] The figure shows the 2016 data set. Distribution and number of respondents and non-respondents within each group is very similar in the 2017 data set; see Appendix Figure A.1

(registration no. [masked for review]), which is the independent authority that supervises compliance with the rules on protection of personal data.

The Danish Wellbeing Survey. The survey was administered electronically by the schools, with each school designating one person as responsible for collecting the data (typically the school principal). Each classroom had a designated teacher responsible for conducting the survey in the class. The Ministry of Education had only few formal requirements in terms of how to conduct the data collection and how to introduce the survey to the students. However, the ministry asked teachers to have all students in a classroom respond to the survey during the same lesson, and to inform them that the purpose was to improve the wellbeing of all students at the school. Teachers were also to tell students that their responses would not be shown to their parents, teacher, or other persons at the school, that they should respond honestly, and that they could have the questions read aloud if they had reading problems or be helped in other ways. The designated teachers had the authority to assess whether students with special needs would be incapable of responding and therefore exempted. Schools were free to use whatever software to administer the questionnaire they wanted, but should upload data according to certain standards, which included that all questionnaires should be linked to the students' national identification number. We were therefore able to combine the survey data with data on test scores and administrative records.

The numbers of respondents in the 2016 (2017) DWS were 46,062 (46,584) in fourth grade, 48,250 (48,272) in sixth grade, and 44,234 (42,564) in eighth grade. The reduced sample size in grade 8 compared to grade 6 is caused both by more students shifting from public to private schools in higher grades and by lower response rates among the older students. However, we still have data on many students with high and low scores on both the reading skill test and personality measures even in grade 8. So, range restrictions should not meaningfully attenuate the correlations we are able to estimate. We return to this issue in the discussion.

Validation survey. We did a separate data collection in order to validate the measures in the DWS. In the validation study, which was merged to our study samples using national identification numbers, students responded to validated Big Five instruments. The appendix presents validation analyses. Results showed that the items that were included in our measures correlated well with the relevant items from the Big Five Inventory (BFI; John & Srivastava, 1999). We performed principal-component factor analyses for each of the traits separately, retaining a single factor (as indicated by scree plots) for the combined set of relevant items from the BFI and the DWS. The mean (un-rotated) factor loading for the Conscientiousness factor was .63 for the DWS items compared to .56 for the BFI items (Table A.1). For Agreeableness, the mean factor loading for the survey items was lower, .47, compared to .58 for the BFI items (Table A.2), whereas for Emotional Stability survey items on average loaded .68 compared to .52 for the BFI items (excluding a malfunctioning BFI item) (Table A.3). Our trait measures also had satisfying levels of convergent and discriminant validity with International Personality Item Pool (IPIP; Goldberg, 1999) measures of several facets with the Big Five (Table A.4). Finally, we found that the overall bivariate correlations with reading test scores when using the full BFI scales from the validation data were all within .01 of the overall results when using our DWS personality scales (Table A.5; see appendix for all details).

Tests of academic performance. All students in public schools in Denmark were required to participate in the Danish National Tests. They were comprised of reading tests in grade 2, 4, 6, and 8, and math tests in grades 3 and 6.[8] They took place towards the end of the school year, after the Danish Wellbeing Survey was completed. The tests were IT-based and scored objectively by an algorithm and therefore blind to the students' personality in the sense that test scores only depended on the students' responses to the tasks (for a detailed description of the test system, see Nandrup and Beuchert-Pedersen, 2018).

---

[8] In 2018 a math test was added to the program for grade 8, but data for this is not yet available.

Measures

Conscientiousness. We used three items from the DWS to measure Conscientiousness: "How often can you complete what you set out to do?" "Can you concentrate during class?" and "If interrupted during lessons, I can quickly concentrate again." Scale reliability coefficients are adequate, especially considering the scale's brevity ($\alpha_{2016} = .69$; $\alpha_{2017} = .70$).

Emotional Stability. We used three items to measure Emotional Stability: "Do you feel lonely?" "Other students accept me as I am.," and "How often do you feel secure at school?" ($\alpha_{2016} = .71$; $\alpha_{2017} = .71$).

Agreeableness. Two items measured Agreeableness: "I try to understand my friends' feelings when they are sad or upset," and "I am good at collaborating with others." With only two items, the reliability was considerably more modest ($\alpha_{2016} = .40$; $\alpha_{2017} = .42$).

To ease interpretation, the personality measures were standardized first by standardizing each item (i.e. M = 0 and SD = 1), then calculating the average of the standardized items and finally standardizing the total score.

Reading performance. The reading test assessed students' performance in three subdomains: Decoding, language comprehension, and reading comprehension. Items used to assess decoding included tasks that asked the student to separate a sequence of letters into three meaningful words. An example of a language comprehension tasks is to identify the meaning of a single word with multiple-choice response categories. An example of reading comprehension would have asked students to read a paragraph and afterwards select the appropriate description of one character in the paragraph, also using multiple-choice response categories.

The test system estimated student ability on each of the subdomains using a continuous logit scale ranging from -7 to 7. Following Nandrup and Beuchert-Pedersen (2018), we first standardized

the test scores from each of the subscales ($M = 0$ and $SD = 1$). Then we used the mean of the three standardized measures and standardized this total score.

The sample of students with information on all measures contained 135,389 students in 2016 and 127,375 in 2017.

Analytical Procedure

We analyzed the data using Ordinary Least Square regressions. Since the dependent and explanatory variables were standardized, the regression coefficients can be interpreted as standardized beta coefficients.

Results

To test the age-trait explanation that the predictive power of Agreeableness and Emotional Stability declines from primary to lower secondary school grades, we compared the bivariate correlations between each of the three personality traits and academic performance for students in grades 4, 6, and 8. Table 2 shows that the correlations of Emotional Stability with academic performance in 2016 (2017) declined by .07 (.09) from .14 (.15) to .07 (.06). (Note that because our estimates are extraordinarily precise, the 95% CIs typically include a range of .02 and are left in the tables for readability.) The changes across grades for Agreeableness and Conscientiousness were very small. Agreeableness declined in 2016 (2017) by .02 (.01) from .19 (.17) in grade 4 to .17 (.16) in grade 8. Conscientiousness was .29 in grade 4 and .30 in grades 6 and 8. In a strict sense, the age-trait explanation is supported given that the declines in both Agreeableness and Emotional Stability are statistically significant at the $p < .05$ level. However, these declines were much smaller than Poropat's

(2009) results, where Agreeableness dropped from .30 to .05 between primary and secondary education, and Emotional Stability dropped from .20 to .01.[9]

[Table 2 here]

The measurement artifact explanation suggests that younger children will be less able to distinguish between their own personality traits, and that therefore the inter-trait correlations will be stronger in grade 4 than in grade 8. Figure 2 graphs these inter-trait correlations. We did see a decline in the correlations between Conscientiousness (C) and the other two traits, Agreeableness (A) and Emotional Stability (ES) (see panels A and B) as well as for the smaller correlation between Agreeableness and Emotional Stability (panel C). Again, the results were very similar in the original 2016 data (dashed lines) and in the 2017 replication data (solid lines). These results thereby confirm the findings by Soto and colleagues (2008), even though for Conscientiousness and Agreeableness our changes from around .44 to around .36 were smaller than their results showing a drop from around .40 to around .22. This may partly be because whereas the BFI was developed for adults, the DWS was developed explicitly for children and young adults.

[Figure 2 here]

One possible explanation for the high inter-trait correlation in younger grades may be that reading skills of younger children are lower and that they therefore are less able to accurately comprehend and respond to personality assessments. We examined this possibility by comparing

---

[9] The average correlation between Conscientiousness and test scores across grades was .29 (Table A.5), which is much stronger than what Poropat (2009) found in his meta-analysis across all ages ($r = .19$ unadjusted and .22 when correcting for scale reliability), even though Poropat's studies predominantly used GPAs, which may be more influenced by student behavior in the classroom than IT-scored tests. The average correlations across grades between test scores and both Agreeableness and Emotional Stability was more similar to those reported by Poropat: .19 (vs. .18 by Poropat) for Agreeableness, and .11 for both studies concerning Emotional Stability.

scale reliability coefficients, alpha (α), for students with very high and very low reading skills across the grades. We compared Groups 8 and 9 in Figure 1, i.e. students with reading skills more than 1.25 SD above the mean, to Groups 1 and 2, i.e. more than 1.25 SD below the mean. Figure 3 shows that for both the Conscientiousness and Agreeableness scales, reliabilities were very similar both across grades and across students with very low and very high reading skills in the 2016 data. Even though we found slightly increased differences in scale reliabilities of Emotional Stability between low and high skilled readers in 4th grade, we found almost no differences in 8th grade. The average scale reliability coefficients for Emotional Stability were .70 in 4th grade, .73 in 6th grade and .70 again in 8th grade. This is an indication that comprehension was not a general problem for items in the DWS— and that reliability was not generally increasing between 4th and 8th grade. Figure A2 in the appendix shows very similar results in the 2017 data.

[Figure 3 here]

A final question therefore is whether Conscientiousness stands out as the single most important trait for predicting reading scores across all grades from 4th to 8th when controlling for the other traits in a multivariate model; and whether it maintains its relationship with reading test scores across this time span. Table 3 shows that this was, indeed, the case. In grade 4, the beta coefficients were .25 (95% CI [.24, .26]) in both years. In grade 8 the beta coefficients were .29 (95% CI [.28, .30]) in the original 2016 data and .30 (95% CI [.29, .31]) in the 2017 replication data. Comparing these results to Table 2, the correlations between Conscientiousness and reading scores were only to a very limited extent affected by including the other two variables in the model, and most in grade 4, as proposed by the measurement artifact explanation. In grades 6 and 8, in both 2016 and 2017, the coefficients ranged between .29 and .30 when controlling for the other traits (Table 3) and between

.30 and .32 when not controlling (Table 2). In grade 4, the coefficients changed from .29 when not controlling (in both 2016 and 2017; Table 2) to .25 when controlling (in both years; Table 3).

[Table 3 here]

In contrast, the coefficients for Agreeableness were more than halved when controlling for Conscientiousness, and did not decline between grades 4, 6, and 8. In 2016 (2017) the correlation was .07 (.05) in grade 4 and .08 (.08 and .07) in grade 6 and 8. These results generally support the notion that the importance of Agreeableness did not drop dramatically between primary and lower secondary school (as found in Poropat, 2009), but rather that when controlling for Conscientiousness, Agreeableness stayed at the same steady, lower level, at around .07.

The coefficients of Emotional Stability became even closer to zero once controlling for Conscientiousness and Agreeableness. However, in contrast to Agreeableness, they showed the same declining trend from grade 4 to grade 8 as we saw in the bivariate correlations in Table 2. This is illustrated in Figure 4. The green lines depict the correlation between Emotional Stability and academic performance. The dashed lines show the bivariate correlations from Table 2, the full lines show the regression coefficients from Table 3. Comparing the dashed and the full green lines, we see that controlling for Conscientiousness and Agreeableness reduced the correlations between academic performance and Emotional Stability by around .13 in each grade, but the negative trend from grade 4 to grade 8 was the same. That is, the connection between Emotional Stability and academic performance truly does seem to shift during adolescence. Perhaps counterintuitively, Emotional Stability ended being negatively connected with reading scores in grade 8, once Conscientiousness and Agreeableness were controlled for, with standardized coefficients of -.05 (-.06) in 2016 (2017). Although it can be tempting to ignore small coefficients, the fact that these associations were

measured so precisely and replicated so closely in 2017 (95% CI [-.06,-.04] in 2016 and [-.07,-.05]) supports paying them more attention. Specifically, our results suggest there are small but detectable changes in the link between Emotional Stability and academic performance (beta coefficients becoming .06 and .09 more negative between 4th and 8th grade in 2016 and 2017, respectively), and that rather than facilitating academic performance, Emotional Stability may in fact produce a modest impairment among older students.

[Figure 4 here]

Figure 4 also makes it clear that there were no differences in grade 8 between the bivariate and the partial correlation between Conscientiousness (blue lines) and reading scores, whereas this difference between the bivariate and the multivariate correlations continued to be large for Agreeableness (red lines) and Emotional Stability (green lines). This supports the interpretation that as students become older and more adept at distinguishing their personality traits, Conscientiousness remains the most important predictor of reading scores and is largely unaffected by the students' other personality traits.

To make sure the results were not driven by large differences between boys and girls, we ran supplementary analyses that included gender as a control variable for the multivariate analyses in Table 3. They showed that the results for Conscientiousness were virtually unaffected by controlling for gender. The coefficients for Agreeableness were a bit lower when also controlling for gender (.04 instead of .07 in grade 4). Emotional Stability was slightly more positively related to reading when also controlling for gender. Instead of changing from .01 to -.05 between grade 4 and 8, it changed from .03 to -.03, but the trend was similar.

Discussion

Our analyses have shown consistently that of the three traits studied here, Conscientiousness was by far the most important predictor for educational performance, even among the younger ages that prior meta-analytic work had indicated would be comparably affected by Agreeableness and Emotional Stability (Poropat, 2009). Furthermore, Conscientiousness is as strong a predictor of academic performance in primary school (4th grade) as it is in lower secondary (8th grade). Since few studies have examined the relationship between Conscientiousness and academic performance in younger populations, this is an important result showing that already from the age of 10, Conscientiousness is a strong predictor for reading.

The *age-trait explanation* predicts a decline in the link between Agreeableness and test scores in the later grades, and our results did exhibit such a decline. However, this difference was incredibly small, and it disappeared when controlling for Conscientiousness. Instead, we found – as expected from the *measurement artifact explanation* – that the inter-trait correlation between Conscientiousness and Agreeableness dropped from grade 4 to grade 8. We found no evidence that this was driven by younger students' lack of ability to comprehend and respond to the items, as scale reliability coefficients remained stable for students with both low and high reading skills across grades. Instead, the decreased inter-trait correlations could reflect students becoming better at knowing themselves and distinguishing between different aspects of their behavior and personality as they become older. Whatever the explanation, our multiple regression results suggested that much of the apparent effect of Agreeableness on academic performance appeared attributable to its overlap with Conscientiousness.

Emotional Stability, by contrast, displayed a more clearly changing relationship with academic performance as students aged. Students with high Emotional Stability tended to perform better in the reading tests in all grades, but less so in 8th grade. However, a substantial part of this

relationship was because these students also tended to report that they had higher Conscientiousness. Once Conscientiousness was controlled for, students in 8th grade with high levels of Emotional Stability actually performed slightly worse on the test. This was a surprising finding, but it could very well be that, aside from the low Conscientiousness that tends to accompany low Emotional Stability, being a little concerned about school performance may contribute positively to performance for older students.

The results have important implications for future research on children and young adults. Especially three findings stand out: First, the early, stable and continuing importance of Conscientiousness for academic performance contradicts previous findings of declining and more modest associations in other studies of primary school students. Second, the reduced importance of Agreeableness and Emotional Stability once Conscientiousness is controlled for shows that when studies analyze correlations between single personality traits and academic outcomes, the apparent effects are likely to be inflated due to covariance among different personality traits – especially covariation with Conscientiousness. Third, the small but negative relationship between Emotional Stability and academic performance is intriguing and warrants further studies of the psychological mechanisms that could be driving this result.

Before we discuss the limitations of the study, we note some of the features that make this study stand out compared to existing studies. First, since all public-school students in the country were required to participate in both the survey measuring their personality traits and in the test measuring their reading performance, we had little self-selection into the sample and very broad range or variation in the measured traits and performance. Second, our first study had a very large sample size of more than 125,000 students – much more than the total participants in many previous meta-analyses – which gave us statistical power to estimate the relationship between personality traits and educational performance very precisely. Third, the precise estimates replicated closely in a separate

sample of another more than 125,000 students. Fourth, because the reading tests are measured by a computer system, our academic performance measure is not biased by influences such as the teacher's attitudes towards the student in ways that might occur for studies using teachers' assessments (e.g., grade point average).

Limitations and Future Directions

Despite the frequent use of brief measures of personality, the measures used in the present study have limitations that require careful consideration. While our measures appear to reflect core aspects of their respective traits, use of longer batteries may capture more variation in the traits as well as a wider range of each trait's content. Agreeableness requires particular attention, as the scale is not only the briefest of the measures used here but both of its items had comparatively modest loadings in the factor analysis together with other Agreeableness items from the Big Five Inventory. Even if the low scale reliability score of around .40 is partly influenced by the low number of items, all results based on the DWS Agreeableness measure should be interpreted with caution. At the same time, the highly similar results across two samples suggest the concern for DWS Agreeableness results should relate not to the measure being particularly noisy but instead to how precisely it assesses Agreeableness as represented in other Big Five measures. Because Agreeableness is represented in consequentially diverse ways across Big Five measures (cf. Hilbig, Moshagen, & Zettler, 2016), future research on the trait should consider using multiple Big Five inventories. Rather than equally representing diverse components of Agreeableness, correlations with different IPIP facets (reported in the Appendix) indicate that Altruism represents a particular focus of the DWS Agreeableness measure.

The IPIP facet profile of our Conscientiousness measure is also instructive (see Appendix). Correlations between our Conscientiousness measures and different facets of the trait show that our

measure is most highly related to Self-efficacy (r = .59), Self-discipline (.56), Achievement striving (.40), and Derryberry and Reed's (2002) Concentration measure (.47), with considerably lesser associations with Dutifulness (.23), Orderliness (.16) and Cautiousness (.08). A measure with this facet profile could be expected to show comparatively strong correlations with educational performance (Chamorro-Premuzic & Furnham, 2004). The fact that two of the three Conscientiousness items (as well as two of three Emotional Stability items) refer to the school context could also inflate the correlations between these measures and test scores. These potential contributors to increased associations with outcomes are presumably offset to some degree by attenuations due to imperfect reliability that is inherent to abbreviated measures such as those from the DWS. Supplementary analyses using the Big Five Inventory data collected (Ns = 292 – 312), however, suggest that these various potential biases may have cancelled each other out, as the correlations between academic performance and BFI scores from our validation survey were all within .01 of the result observed using the DWS measures (see Appendix). This rather stunning degree of similarity in results between the two measures is likely somewhat of a fluke, because Big Five measures are expected to exhibit larger divergences in their relationships with outcomes across scales (e.g., Hilbig, Moshagen, & Zettler, 2016); still, it does speak against the possibility that the DWS measures markedly overstate the relationship between personality and academic performance.

Another limitation of the present study is that it lacks assessments of Openness and Extraversion. Future work should address these limitations by measuring all the Big Five personality traits with full instruments. However, since Conscientiousness has been found to be the most important personality trait for educational achievement in previous research (Almlund et al., 2011; Poropat, 2009; Trapmann et al., 2007; Vedel, 2014), and Conscientiousness tends to be highly independent of Openness (the second best Big Five predictor of academic performance; Soto and John, 2017), this is not expected to change the findings substantially.

As mentioned, the sample size falls between grade 6 and 8, which relates partly to a lower response rate among older students and partly to some students leaving the public school system for private schools in older grades. There are still about 40,000 students in grade 8 alone, so there should be little cause for concern that range restriction could limit the estimated relationships between personality traits and educational performance. Furthermore, if the relationships between the individual traits and reading scores are the same for all students (what may be thought of as homogeneous treatment effects of the traits), a biased sample would not bias the estimate of this relationship. Only if the personality traits had different importance for different subgroups of students, and if some of these subgroups are underrepresented in our sample, would the estimates be biased. We explored one major subgroup of interest – gender – and found no evidence that the results meaningfully differed between groups. Yet, future research should examine whether personality traits measured in 6th grade predict students' non-response in 8th grade, and whether personality traits predict a shift from public to private school. Besides the methodological interest in these questions, they would be of substantial interest for research in the different kinds of consequences of young adolescence' personality traits.

The present study's use of cross-sectional data might be seen as limiting the ability to infer that these are truly developmental changes within the individual rather than cohort effects. However, the very precise replication of results in a sample one year younger might be taken to imply that the differences between students in grades 4, 6, and 8 are more likely to reflect developmental than cohort effects. Nevertheless, future studies may also collect longitudinal data to study changing relationships within individual children as they become older. This lack of longitudinal data is a limitation of the present study that we share with the two existing studies on the age-trait relationship from childhood to adolescence.

Data limitations meant that this study only examined the relationship of personality traits with reading scores. Future research should examine whether the same relationships exist for other school subjects. Conscientiousness may also be helpful for learning math and science, since strong work habits can be helpful for studying any subject. Similarly, a small negative effect of Emotional Stability could also be expected for performance in various subjects, not just reading.

Finally, even though this study used a nationwide sample of students, all of those students are from the same Scandinavian country. Another question for future research would therefore be whether the results found in Denmark would replicate in other countries. The Big Five personality model has proven to be adequate across a broad range of countries and cultures (John & Srivastava, 1999) and recently also in a validation study from Denmark (Vedel et al., 2019), so the relationship between traits and educational performance is expected to be consistent across countries, but that needs to be examined.

Conclusion

In 2009, the same year that Poropat (2009) published his seminal meta-analysis of the relationship between personality and academic performance, Rozin (2009, 435) argued that the "discovery of fundamental phenomena, such as functional relations that apply to the real world and have generality, should have a higher priority in psychology. Such findings have been the basis for theoretical advances in other natural sciences." Within the domain of personality and education, Poropat's main finding of the supreme importance of Conscientiousness for educational performance in older students has largely been confirmed in subsequent research. However, the more specific results from late childhood through adolescence showing unstable relationships between educational

performance with Agreeableness and Emotional Stability have not been clearly replicated afterwards, highlighting the need for further study.

A pair of extremely large samples were used in this study to demonstrate highly replicable relationships between personality and academic performance, where Conscientiousness in particular showed substantial and stable connections to test scores throughout the age ranges studied. Conscientiousness accounted for much of the apparent relationship between test scores and both Agreeableness and Emotional Stability. While Agreeableness mirrored Conscientiousness in exhibiting stable connections to test scores throughout development, Emotional Stability lost its positive association with test scores as children aged.

Despite limitations in the available data to measure especially Agreeableness, finding such stable functional relations makes it all the more relevant in future work to study to what extent the development of these traits can be supported through targeted interventions in primary and secondary school. That could potentially be very helpful for children from family environments with little parental support.

References

Almlund, M., Duckworth, A. L., Heckman, J. J. & Kautz, T. (2011). Personality Psychology and Economics. *Handbook of the Economics of Education*, 4, 4-181.

Borgers, N., De Leeuw, E., & Hox, J. (2000). Children as respondents in survey research: Cognitive development and response quality 1. *Bulletin de methodologie Sociologique*, 66, 60-75.

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences*, 113, 13354-13359; DOI: 10.1073/pnas.1601135113.

Caprara, G.V., Vecchione, M., Alessandri, G., Gerbino, M., & Barbaranelli, C. (2011). The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology*, 81, 78-96.

Chamorro-Premuzic, T., & Furnham, A. (2004). A possible model for understanding the personality-intelligence interface. *British Journal of Psychology*, 95(Pt 2), 249–264.

Chamorro-Premuzic, T., & Furnham, A. (2006). Intellectual competence and the intelligent personality: A third way in differential psychology. *Review of General Psychology*, 10, 251–267.

De Leeuw, E. D. (2011). Improving data quality when surveying children and adolescents: cognitive and social development and its role in questionnaire construction and pretesting. In *Report prepared for the Annual Meeting of the Academy of Finland: Research Programs Public Health Challenges and Health and Welfare of Children and Young People*, May (pp. 10-12).

De Pauw, S. S. W., Mervielde, I., & Van Leeuwen, K. G. (2009). How Are Traits Related to Problem Behavior in Preschoolers? Similarities and Contrasts Between Temperament and Personality. *Journal of Abnormal Child Psychology*, 37(3), 309-325. doi:10.1007/s10802-008-9290-0

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1), 13-21.

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology*, 111, 225-236.

Duckworth, A. L., Quinn, P. D., & Tsukayama, E.. (2012). What No Child Left Behind Leaves Behind: The Roles of IQ and Self-Control in Predicting Standardized Achievement Test Scores and Report Card Grades. *Journal of educational psychology* 104, 439–451. PMC. Web. 15 Nov. 2017.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement. *Developmental Psychology* 43, 1428-1446. DOI:10.1037/0012-1649.43.6.1428

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. (2014). Improving the Dependability of Research in Personality and Social Psychology: Recommendations for Research and Educational Practice. *Personality and Social Psychology Review*, *18*, 3-12. https://doi.org/10.1177/1088868313507536

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower level facets of several Five-Factor models. *Personality Psychology in Europe*, 7, 7–28

Hilbig, B. E., Moshagen, M., & Zettler, I. (2016). Prediction consistency: A test of the equivalence assumption across different indicators of the same construct. *European Journal of Personality*, *30*(6), 637–647. https://doi.org/10.1002/per.2085

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual issues. In O. P. John, R. W. Robins, &

L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory - Versions 4a and 54.* Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. Pervin and O.P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed.). New York: Guilford

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490.

Maples, J. L., Guan, L., Carter, N. T., & Miller, J. D. (2014). A test of the International Personality Item Pool representation of the Revised NEO Personality Inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment*, 26, 1070.

Medford, E., & McGeown, S. P. (2012). The influence of personality characteristics on children's intrinsic reading motivation. *Learning and Individual Differences,* 22, 786-791. doi:10.1016/j.lindif.2012.06.002.

Nandrup, A. B., & Beuchert-Pedersen, L. V. (2018). The Danish national tests at a glance. *Nationaloekonomisk Tidsskrift*, 1.

Neuenschwander, R., Cimeli, P., Röthlisberger, M., & Roebers, C. M. (2013). Personality factors in elementary school children: Contributions to academic performance over and above executive functions? *Learning and Individual Differences*, 25, 118-125. doi:10.1016/j.lindif.2012.12.006

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Poropat, A. E. (2009). A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance. *Psychological Bulletin* 135, 322-338.

Poropat, A. E. (2014). A meta-analysis of adult-rated child personality and academic performance in primary education. *British Journal of Educational Psychology*, 84(2), 239-252. doi:10.1111/bjep.12019

Poropat, A. E. (2016). Beyond the shadow: The role of personality and temperament in learning. In L. Corno & E. Anderman (Eds.), *Handbook of Educational Psychology* (pp. 172-185). Washington D.C.: American Psychological Association, Division 15 - Educational Psychology.

Romano, E., Babchishin, L., Pagani, L. S., & Kohen, D. (2010). School readiness and later achievement: Replication and extension using a nationwide Canadian survey. *Developmental Psychology* 46, 995-1007. DOI:10.1037/a0018880.

Rozin, P. (2009). What Kind of Empirical Research Should We Publish, Fund, and Reward? A Different Perspective. *Perspectives on Psychological Science*, 4, 435-439.

Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 21–50). New York: Guilford.

Spearman, C. (1927). *The Abilities of Man*. London: Macmillan.

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

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., … Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*(5), 742–756. https://doi.org/10.1177/1745691617690042

Trapmann, S., Hell, B., Hirn, J.-O. W., & Schuler, H. (2007). Meta-Analysis of the Relationship Between the Big Five and Academic Success at University. *Zeitschrift für Psychologie/Journal of Psychology* 215, 132-151.

Vedel, A. (2014). The Big Five and Tertiary Academic Performance: A Systematic Review and Meta-Analysis. *Personality and Individual Differences* 71(Supplement C), 66-76.

Vedel, A., Wellnitz, K. B., Ludeke, S., Soto, C., John, O., & Andersen, S. C. (2019) Development and validation of the Danish Big Five Inventory-2: Domain- and facet-level structure, construct validity, and reliability. *European Journal of Psychological Assessment*. (Accepted/In press).

Webb, E. (1915). *Character and Intelligence; an Attempt at an Exact Study of Character*, in collab. with Wellesley College Library. Cambridge, University Press.

Zhang, J., & Ziegler, M. (2016). How do the big five influence scholastic performance? A big five-narrow traits model or a double mediation model. *Learning and Individual Differences*, 50, 93-102. doi:10.1016/j.lindif.2016.07.011

Zupančič, M., Kavčič, T., Slobodskaya, H. R., & Akhmetova, O. A. (2016). Broad and Narrow Personality Traits Predicting Academic Achievement Over Compulsory Schooling: A Cross-Sectional Study in Two Countries. *The Journal of Early Adolescence*, 36, 783-806.

Tables and Figures

Table 1

Studies of primary school students citing Poropat (2009)

Reference	Sample			Correlations											
	Age, mean	Age, range	N	C-Perf	p	A-Perf	p	N-Perf	p	C-A	p	C-N	p	A-N	p
Neuenschwander et al. (2013)	8.1	approx.6-9	446	.19	.01	.09	.07	-.15	<.001	0.37	<.001	0.31	<.001	0.47	<.001
Zupančič et al. (2016)	11.9	8-15	1618	0.19		0.04		-0.04		NA		NA		NA	
Medford and McGeown (2012)	10.6	approx. 9-11	295	0.1	>.05	-0.06	>.05	NA		0.44	<.01	NA		NA	

Note: "C" = Conscientiousness, "A" = Agreeableness, "N" = Neuroticism (the reverse of Emotional Stability), "Perf." = Academic performance.

Table 2

Bivariate correlations with reading skills

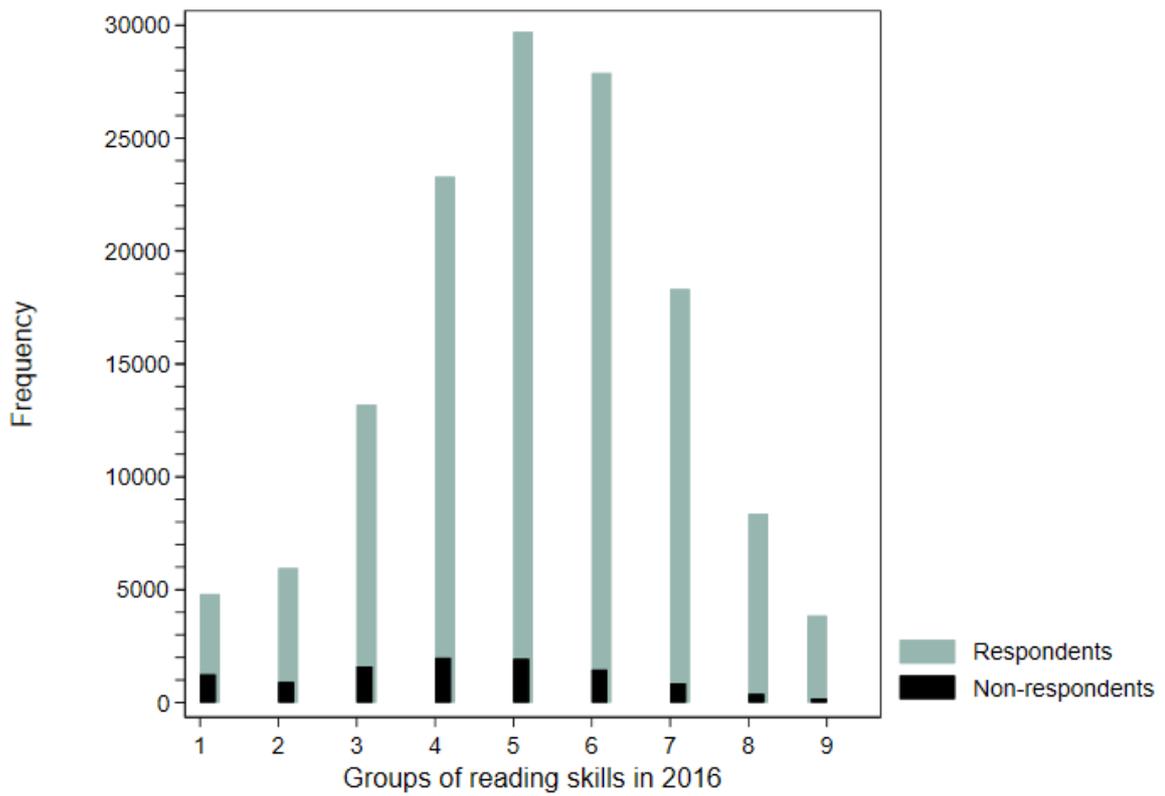|	Grade 4		Grade 6		Grade 8	
	2016	2017	2016	2017	2016	2017
Conscientiousness	.29***	.29***	.31***	.32***	.30***	.30***
	[.28, .30]	[.28, .29]	[.30, .32]	[.31, .33]	[.29, .31]	[.29, .31]
Agreeableness	.19***	.17***	.20***	.20***	.17***	.16***
	[.18, .19]	[.16, .18]	[.19, .20]	[.19, .21]	[.16, .18]	[.15, .18]
Emotional Stability	.14***	.15***	.11***	.10***	.07***	.06***
	[.14, .15]	[.14, .16]	[.10, .12]	[.09, .11]	[.06, .08]	[.05, .07]
Observations	45,135	43,963	47,305	45,881	42,949	37,531

Note: 95% confidence intervals in brackets. * $p < 0.05$, ** $p < 0.01$, * $p < 0.001$

Table 3

Standardized regression coefficients from multivariate analysis

	Grade 4		Grade 6		Grade 8	
	(1)	(2)	(3)	(4)	(5)	(6)
	2016	2017	2016	2017	2016	2017
Conscientiousness	.25***	.25***	.29***	.30***	.29***	.30***
	[.24,.26]	[.24,.26]	[.28,.30]	[.29,.31]	[.28,.30]	[.29,.31]
Agreeableness	.07***	.05***	.08***	.08***	.08***	.07***
	[.06,.08]	[.04,.06]	[.07,.09]	[.07,.09]	[.07,.09]	[.06,.09]
Emotional Stability	.01**	.03***	-.03***	-.03***	-.05***	-.06***
	[.00,.02]	[.02,.04]	[-.03,-.02]	[-.04,-.02]	[-.06,-.04]	[-.07,-.05]
Observations	45,135	43,963	47,305	45,881	42,949	37,531

Note: 95% confidence intervals in brackets. * $p < 0.05$, ** $p < 0.01$, * $p < 0.001$

Figure 1: Personality respondents and non-respondents based on reading performance. Group 5 is centered around the mean level of performance (-0.25 to 0.25 *SD*). Remaining groups represent additional half-*SD* steps away from mean performance, except groups 1 and 9, which run between 1.75 and 4 *SD* below and above the mean, respectively.
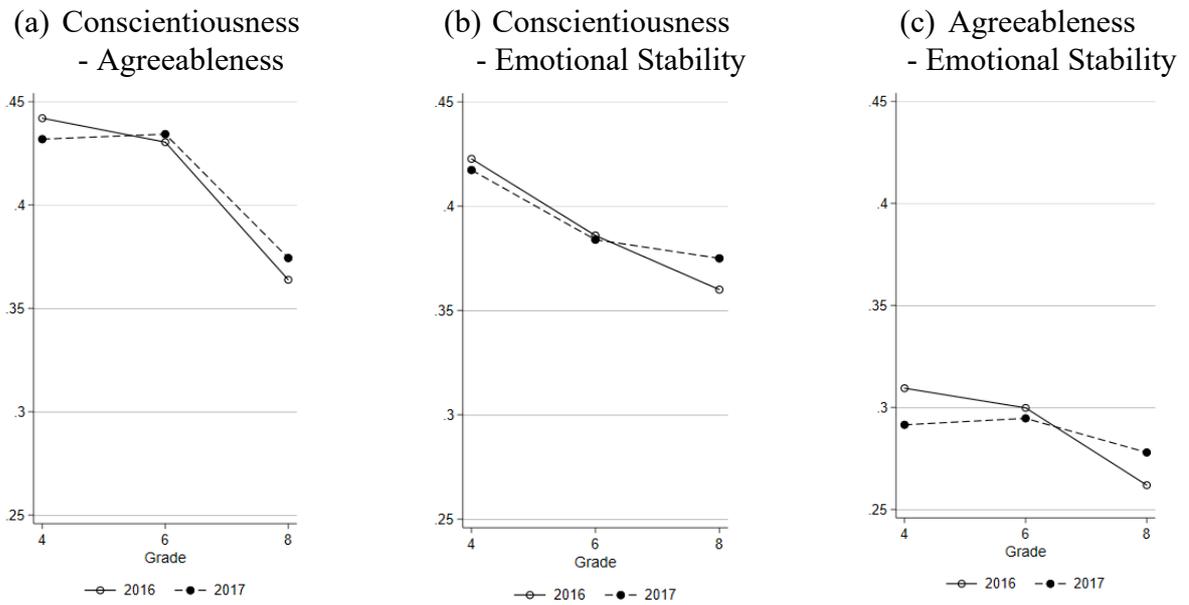
| (a) Conscientiousness - Agreeableness | (b) Conscientiousness - Emotional Stability | (c) Agreeableness - Emotional Stability |

Figure 2: Inter-trait correlations of the Danish Wellbeing scales across grades 4, 6, and 8
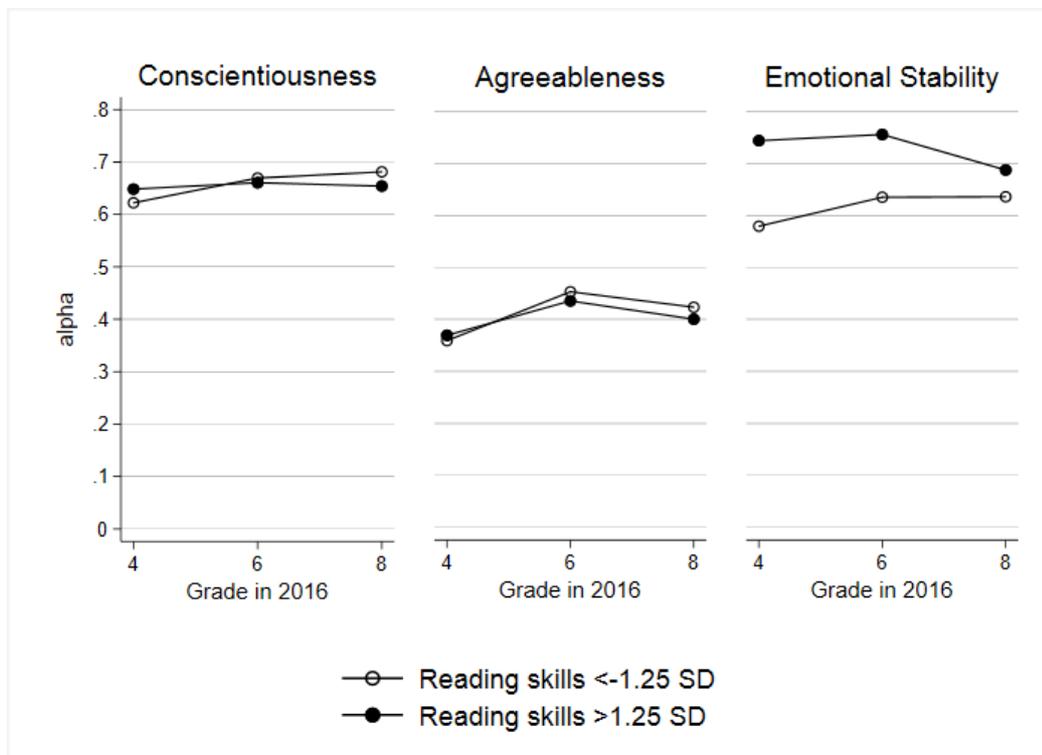

Figure 3: Scale reliability coefficients (alpha) by grade and reading skill levels.
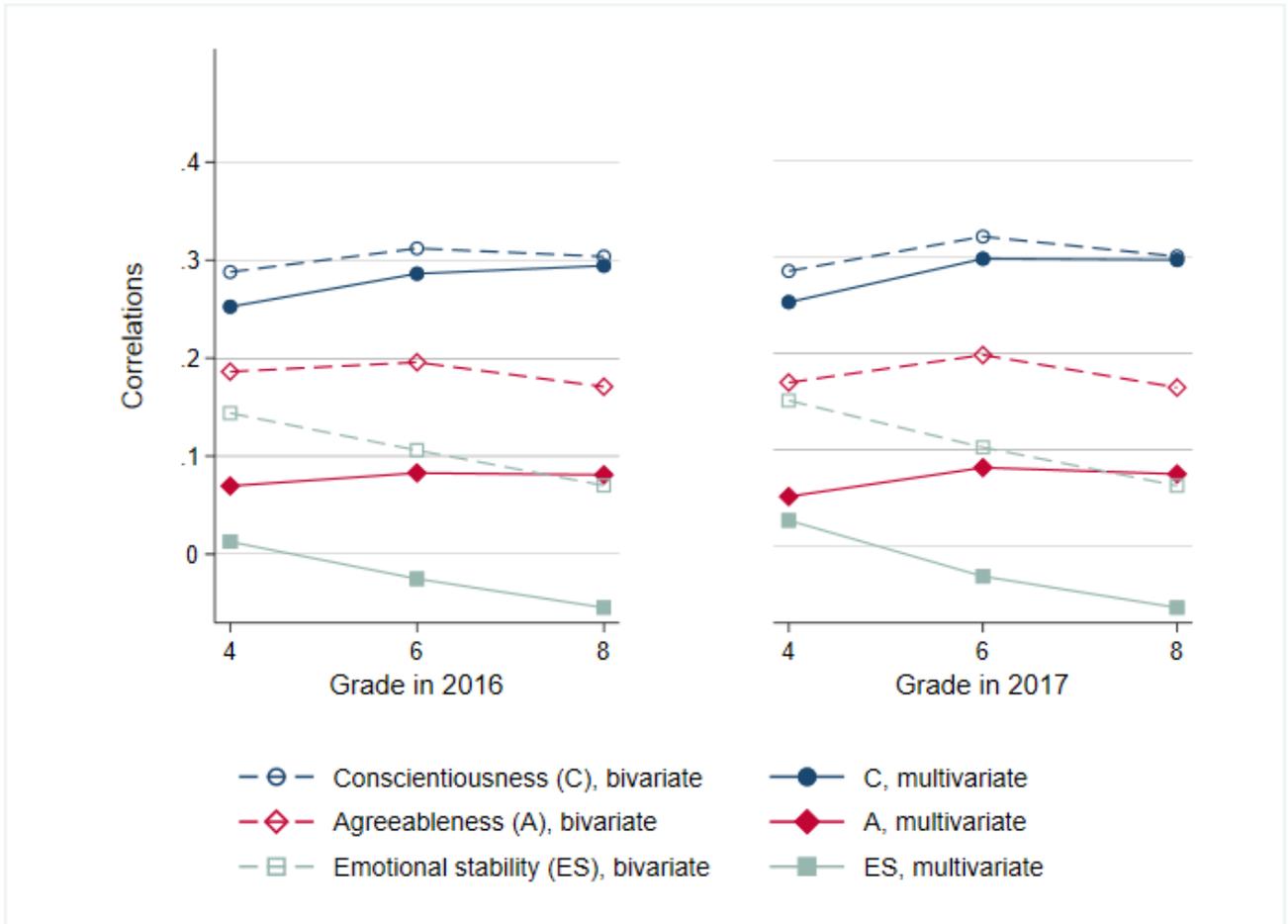
Figure 4: Bivariate and multivariate correlations with reading skills. Confidence intervals too small to be visually represented (see Tables 2 and 3), but all results significantly different from 0 at $p < .01$.

Online Appendix for

A Stable Relationship Between Personality and Academic Performance from Childhood through Adolescence.
An original study and replication in hundred-thousand-person samples

Simon Calmar Andersen, Miriam Gensowski, Steven Ludeke,
and Oliver P. John

Participants in 2017

Figure A.1 shows respondents and non-respondents divided by reading test score groups in the 2017 replication sample.
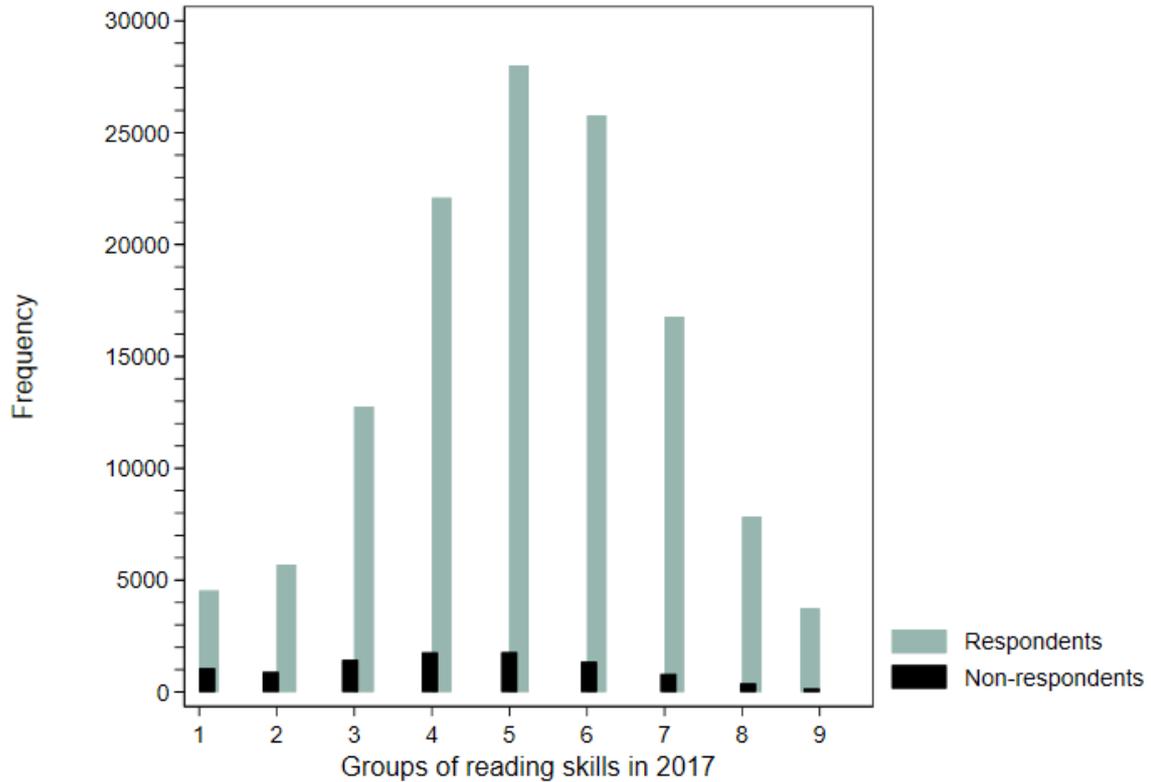
Figure 1: Personality respondents and non-respondents based on reading performance. Group 5 is centered around the mean level of performance (-0.25 to 0.25 *SD*). Remaining groups represent additional half-*SD* steps away from mean performance, except groups 1 and 9, which run between 1.75 and 4 *SD* below and above the mean, respectively.

Validation of the personality measures

To validate the items used to measure Agreeableness, Conscientiousness, and Emotional Stability, we surveyed 1,768 public school students in a questionnaire that included validated scales from the Big Five Inventory (BFI-C; John et al., 2008 and John et al., 1991) and from the International Personality Item Pool (IPIP; Maples et al., 2002). To reduce the number of items that each individual

student should respond to, we randomly assigned students to three different surveys, with each three versions of the item ordering, or nine versions in total. There were complete responses for 413, 422, and 425 of the BFI scales for Conscientiousness, Agreeableness, and Neuroticism, respectively. The survey was conducted in the summer 2015. By use of students' personal identification numbers, we were able to match the validation survey to the Danish Wellbeing Surveys (DWS) that are described in the main text. In addition to matching it to the DWS of 2016, we also match it to the first round of the DWS that took place in 2015. The procedures for this round were the same as for 2016 and 2017. The first wave in 2015 was, however, closer to the validation data than was 2017 data, which is why we use the years 2015 and 2016 for the validation study in this appendix.

Convergent validity

We assess the convergent validity of the items from the DWS by running principal component factor analyses that include the BFI items from the appropriate personality traits. Tables A1-A3 show the outcome from the factor analyses. The BFI items are measured in the summer 2015. The two columns use the personality items from the DWS in 2015 and 2016.

Table A1 shows that the items in general have high loadings and, notably, the three items from the DWS on average have higher loadings than the original BFI items. This confirms that the Conscientiousness items tap into the basic core of trait. In Table A2 we find that the two Agreeableness items from the DWS on average had somewhat smaller loadings than the BFI Agreeableness items, but still loaded at least .38 in both 2015 and 2016. The Emotional Stability items presented in Table A3 have on average higher loadings than the BFI loadings, even if we exclude from the BFI mean the item "Can be moody", which have low factor loadings.

Table A.1

Factor loadings of BFI items and Danish Wellbeing Survey items measuring Conscientiousness

Item	2015	2016
Keeps working until things are done	0.74	0.71
Does things carefully and completely	0.70	0.65
Is easily distracted; has trouble paying attention [1]	0.64	0.67
Does things efficiently (quickly and correctly)	0.63	0.63
Is a reliable worker	0.60	0.66
Tends to be lazy [1]	0.54	0.55
Makes plans and sticks to them	0.37	0.37
Tends to be disorganized [1]	0.28	0.28
Mean loadings, BFI	*0.56*	*0.57*
If interrupted during lessons, I can quickly concentrate again.	0.68	0.67
Can you concentrate during class?	0.63	0.66
How often can your complete what you set out to do?	0.58	0.55
Mean loadings, Danish Wellbeing Survey items	*0.63*	*0.62*
N	367	373
Mean age	12.2	13.2
Age range	9-15	10-16

Note: Entries present (unrotated) loadings from a principal-component factor analysis retaining a single factor. Big Five Inventory (BFI) items measured in June 2015. Questionnaire items measured in January to March 2015 and January to March 2016.
[1] Reverse coded

Table A.2

Factor loadings of BFI items and Danish Wellbeing Survey items measuring Agreeableness

Item	2015	2016
Is considerate and kind to almost anyone	0.70	0.72
Is helpful and unselfish with others	0.67	0.70
Has a forgiving nature [1]	0.62	0.60
Can be cold and distant with others [1]	0.61	0.58
Is sometimes rude to others	0.59	0.57
Is generally trusting	0.56	0.59
Likes to cooperate; goes along with others	0.55	0.53
Starts quarrels with others	0.53	0.55
Tends to find fault with others [1]	0.40	0.41
Mean loadings, BFI	*0.58*	*0.58*
I try to understand my friends when they are sad or upset.	0.50	0.38
I am good at collaborating with others.	0.45	0.43
Mean loadings, Danish Wellbeing Survey items	*0.47*	*0.40*
N	377	380
Mean age	12.1	13.1
Age range	9-15	10-16

Note: Entries present (unrotated) loadings from a principal-component factor analysis retaining a single factor. Big Five Inventory (BFI) items measured in June 2015. Questionnaire items measured in January to March 2015 and January to March 2016.

[1] Reverse coded

Table A.3

Factor loadings of BFI items and Danish Wellbeing Survey items measuring Emotional Stability

Item	2015	2016
Is depressed, blue	0.74	0.71
Worries a lot	0.61	0.68
Stays calm in tense situations [1]	0.55	0.61
Is relaxed, handles stress well [1]	0.50	0.58
Gets nervous easily	0.48	0.51
Can be tense	0.40	0.30
Doesn't get easily upset, emotionally stable [1]	0.36	0.44
Can be moody	0.15	0.15
Mean loadings, BFI	*0.47*	*0.50*
Mean loadings, BFI without "Can be moody"	*0.52*	*0.55*
Do you feel lonely?	0.71	0.66
Other students accept me as I am. [1]	0.69	0.62
How often do you feel secure at school? [1]	0.65	0.66
Mean loadings, Danish Wellbeing Survey items	*0.68*	*0.65*
N	377	367
Mean age	12.1	13.1
Age range	9-15	10-16

Note: Entries present (unrotated) loadings from a principal-component factor analysis retaining a single factor. Big Five Inventory (BFI) items measured in June 2015. Questionnaire items measured in January to March 2015 and January to March 2016.
[1] Reverse coded

Discriminant validity

In order to assess the discriminant validity of the personality scales measured by the DWS we compare the scales to several Big Five facets measured by IPIP. Table A4 presents the results. The Conscientiousness scale is highly correlated with the Productiveness facet of Conscientiousness as measured by IPIP scales for Self-efficacy, Self-discipline and Achievement striving, and the Derryberry concentration scale (Derryberry and Reed, 2002). The Conscientiousness measure is also capable of discriminating against measures of the other Big Five traits.

Our Agreeableness scale correlates mostly with IPIP Altruism and discriminate against the other Big Five facets. The Emotional Stability scale correlates with IPIP scales for Depression, Vulnerability and Anxiety and also discriminate against the other Big Five traits and facets.

Table A.4
Correlations between the Danish Wellbeing Survey measures of Conscientiousness (C), Agreeableness (A), and Neuroticism (N; reverse of Emotional Stability) and the Big Five facets

	C	A	N
C facets			
IPIP Selfefficacy	**0.59**	0.27	-0.28
IPIP Selfdiscipline	**0.56**	0.28	-0.25
Concentration Derryberry	**0.47**	0.16	-0.25
IPIP Achievement striving	**0.40**	0.22	-0.19
IPIP Dutifulness	0.23	0.26	-0.19
IPIP Orderliness	0.16	0.04	-0.10
IPIP Cautiousness	0.08	0.03	0.01
A facets			
IPIP Altruism	0.32	**0.41**	-0.12
IPIP Trust	0.28	0.29	-0.28
IPIP Coorporation	0.27	0.27	-0.06
N facets			
IPIP Depression	-0.30	-0.15	**0.52**
IPIP Vulnerability	-0.39	-0.17	**0.36**
IPIP Anxiety	-0.24	-0.10	**0.32**
IPIP Selfconsciousness	-0.35	-0.12	0.26
E facets			
IPIP Assertiveness	0.26	0.07	-0.05
O facets			
IPIP Artistic interest	0.18	0.07	0.08

Note: Ns: 350-418, except IPIP Orderliness and IPIP Artistic Interest N: 135-142. Facets measured in June 2015. Questionnaire items measured in January to March 2015.

Concurrent validity

To assess the concurrent validity of the DWS personality scales, we compared how the BFI scales from the validation survey and our DWS scales correlated with the reading test score outcomes. Due to the small sample size of the validation survey, we used the overall bivariate correlations across grades 4, 6 and 8. The DWS survey are compared to the reading tests taken closest to the time of the survey, i.e. 2015 for the validation survey and 2016 for the 2016 DWS.

Table A5 shows that correlations are within .01 of each other.

Table A.5

Bivariate correlations with reading test scores

	(1) BFI validation survey	(2) DWS population study
Conscientiousness	.30***	.29***
	[.20,.40]	[.29,.30]
Agreeableness	.20***	.19***
	[.10,.29]	[.18,.19]
Emotional Stability	.12*	.11***
	[.22,.02]	[.11,.12]

Note: Grades 4, 6, and 8 are pooled due to the small sample size of the validation survey. Observations (N) in correlations based on Big Five Inventory (BFI): 292-312. Observations (N) in correlations based on Danish Wellbeing Survey (DWS): 139,127-143,266. Validation survey is from 2015, DWS data are from 2016. 95% confidence intervals in brackets. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Reliability

Figure A2 presents scale reliability coefficients for the 2017 sample by grade and reading skills for each of the three trait measures. Results are similar to those in the 2016 sample presented in Figure 3.
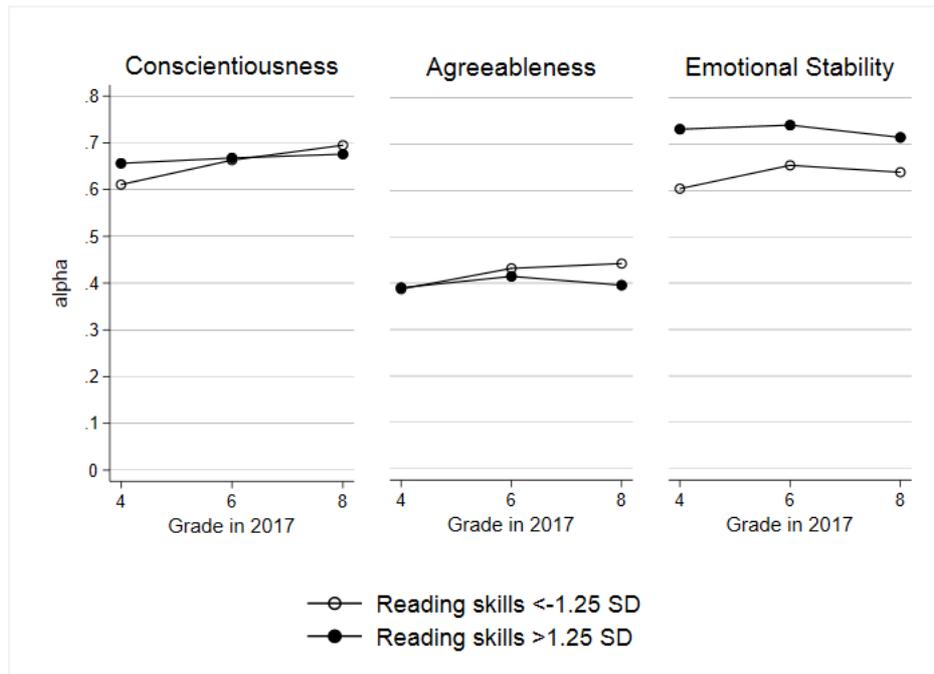

Figure A.2: Scale reliability coefficients (alpha) by grade and reading skill levels. 2017 data.