Andrea Westphal | Rebecca Christine Lazarides | Miriam Vock

Are some students graded more appropriately than others? Student characteristics as moderators of the relationships between teacher-assigned grades and test scores in mathematics

Suggested citation referring to the original publication:
British journal of educational psychology 91 (2020) 3, Art. e12397 pp. 865 - 881
DOI: https://doi.org/10.1111/bjep.12397
ISSN: 0007-0998, 2044-8279

Journal article | Version of record

Secondary publication archived on the Publication Server of the University of Potsdam:
Zweitveröffentlichungen der Universität Potsdam : Humanwissenschaftliche Reihe 853
ISSN: 1866-8364
URN: https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-563330
DOI: https://doi.org/10.25932/publishup-56333

the british psychological society
promoting excellence in psychology

British Journal of Educational Psychology (2021), 91, 865–881

www.wileyonlinelibrary.com

Are some students graded more appropriately than others? Student characteristics as moderators of the relationships between teacher-assigned grades and test scores in mathematics

Andrea Westphal*, Rebecca Lazarides and Miriam Vock
Department of Education, University of Potsdam, Germany

Background. Building on the Realistic Accuracy Model, this paper explores whether it is easier for teachers to assess the achievement of some students than others. Accordingly, we suggest that certain individual characteristics of students, such as extraversion, academic self-efficacy, and conscientiousness, may guide teachers' evaluations of student achievement, resulting in more appropriate judgements and a stronger alignment of assigned grades with students' actual achievement level (as measured using standardized tests).

Aims. We examine whether extraversion, academic self-efficacy, and conscientiousness moderate the relations between teacher-assigned grades and students' standardized test scores in mathematics.

Sample. This study uses a representative sample of $N = 5{,}919$ seventh-grade students in Germany (48.8% girls; mean age: $M = 12.5$, $SD = 0.62$) who participated in a national, large-scale assessment focusing on students' academic development.

Methods. We specified structural equation models to examine the inter-relations of teacher-assigned grades with students' standardized test scores in mathematics, Big Five personality traits, and academic self-efficacy, while controlling for students' socioeconomic status, gender, and age.

Results. The correlation between teacher-assigned grades and standardized test scores in mathematics was $r = .40$. Teacher-assigned grades more closely related to standardized test scores when students reported higher levels of conscientiousness ($\beta = .05$, $p = .002$). Students' extraversion and academic self-efficacy did not moderate the relationship between teacher-assigned grades and standardized test scores.

Conclusions. Our findings indicate that students' conscientiousness is a personality trait that seems to be important when it comes to how closely mathematics teachers align their grades to standardized test scores.

Teacher-assigned grades are ubiquitous in students' school lives and are highly relevant for students' educational trajectories. Educational decisions often rely on teacher-assigned grades for ability grouping (Hallinan, 1992), grade retention (Westphal, Vock, & Lazarides, 2020), and college admissions. Whether or not teacher-assigned grades adequately

**Correspondence should be addressed to Andrea Westphal, Department of Education, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany (email: andrea.westphal@uni-potsdam.de).*

DOI:10.1111/bjep.12397

represent students' achievement is a question of ongoing debate (e.g., Randall & Engelhard, 2010). While students' standardized test scores explain 25 to 35% of variance in teacher-assigned grades (Bowers, 2011), a number of other student characteristics, such as students' Big Five personality traits – especially conscientiousness – (Spengler, Lüdtke, Martin, & Brunner, 2013; Tetzner, Becker, & Brandt, 2020), academic self-efficacy (Caprara, Vecchione, Alessandri, Gerbino, & Barbaranelli, 2011), and socio-demographic characteristics (Hochweber, Hosenfeld, & Klieme, 2014) are also substantially predictive of grades. On the other hand, scores in standardized tests may be biased by test anxiety (Lang & Lang, 2010; von der Embse, Jester, Roy, & Post, 2018). Thus, both, teacher-assigned grades and standardized test score are inevitably containing biases and even errors.

Artelt and Rausch (2014) proposed the idea that Funder's Realistic Accuracy Model (Funder, 1995) could be applied to study the conditions under which teacher-assigned grades more strongly reflect students' standardized test scores. The model implies the idea that certain 'characteristics of individuals [. . .] help or hinder judgeability' (Human & Biesanz, 2013, p. 252). However, research has not yet identified which student characteristics help teachers align their grades more closely to more objective measures of student achievement, such as standardized test scores. In the present study, we use a large and representative sample of seventh-grade students from the German National Educational Panel Study (NEPS) to test whether students' extraversion, academic self-efficacy, and conscientiousness moderate the association between teacher-assigned grades and standardized test scores. We focus on mathematics as a core domain in secondary school, for which 'demands and competence models are clearer' and teachers are more likely to 'have a shared understanding of what constitutes mathematical proficiency' than in other school subjects (Artelt & Rausch, 2014, p. 35).

Teachers' grading practices

The question 'of what it is that grades may be assessing' (Bowers, 2011, p. 143) has been the topic of study, discussion, and even controversy, for several decades. Textbooks used in teacher training instruct teachers to depend on students' achievement in class when awarding report-card grades (e.g., Brookhart, 2004; Linn & Miller, 2005). Using teacher self-reports, empirical research focusing on the information underlying teachers' achievement assessments has shown that teachers in fact rely on a wide range of student characteristics when assigning grades (Brookhart, 1993; McMillan, 2001; Randall & Engelhard, 2009). In addition to the degree to which students have achieved the learning goals set in class, teachers reported, for example, that they also rely on students' effort and their work habits in their evaluation of student achievement (Brookhart, 1993; McMillan, 2001; Randall & Engelhard, 2009). Research also emphasizes, however, that teachers continue to 'primarily assign grades on the basis of student achievement' (Randall & Engelhard, 2009, p. 1; see also McMillan, 2001). These survey findings are corroborated by empirical research that examines the actual relationships between teacher-assigned grades and student characteristics – for instance, the question of whether students who are more conscientious do indeed receive better grades than less conscientious students (Spengler et al., 2013). Although we can show that different aspects of student personality, behaviour, and demographics do indeed explain teacher-assigned grades (e.g., Hochweber et al., 2014; Krejtz & Nezlek, 2016; Spengler et al., 2013), the most substantial amount of variance in teacher-assigned grades can still be attributed to students' standardized test scores (Bowers, 2011).

The Realistic Accuracy Model

The *Realistic Accuracy Model* of personality judgement (Funder, 1995) could serve as a good framework to explain the circumstances under which teacher-assigned grades more closely relate to students' actual standardized test scores (Artelt & Rausch, 2014). Funder (1995) identified four criteria that are crucial for appropriate personality judgements – and which can be applied to judgements of student achievement (see also Artelt & Rausch, 2014) – namely the *relevance*, *availability*, *detection*, and *utilization* of cues. Funder (1995) outlined that appropriate judgements require that relevant cues of the target person are available to the judge and, in addition, the judge detects and utilizes these cues when making their judgements.

Individual characteristics that enable appropriate judgements have been addressed extensively in the context of personality research (Human & Biesanz, 2013). In their review, Human and Biesanz (2013) concluded that psychologically well-adjusted individuals and individuals with a higher social status reveal more relevant information about their personalities and, consequently, their personality is judged more appropriately. This supports the experimental study by Hall, Rosip, LeBeau, Horgan, and Carter (2006), in which pairs of individuals were nominated as either equal-power or unequal-power partners and had to non-verbally transmit messages with positive, negative, or neutral content. The interactions were videotaped and subsequently decoded by a third group of participants. Hall et al. (2006) was able to show that participants in subordinate roles expressed themselves less clearly than participants in equal or dominant roles (for similar results in naturalistic social-status settings and verbal interactions, see Garcia, Hallahan, & Rosenthal, 2007; Gross & John, 2003). Based on their review, Human and Biesanz (2013) suggest that psychologically well-adjusted individuals and individuals with a higher social status tend to express their emotions and opinions in a more authentic, open, and dominant way, thereby providing cues about their personality that are more *relevant* and making these cues *available*.

Artelt and Rausch (2014) have posited that the Funder's Realistic Accuracy Model (Funder, 1995) might also be employed to study the adequacy of teachers' judgements and thus the extent to which teacher judgements or teacher-assigned grades reflect students' standardized test scores. The authors outlined that teachers' judgements of student achievement can be aided by students expressing their comprehension or lack of comprehension (relevant cues) in a way that is observable to the teacher – for instance, by being attentive to the teacher's instruction and by participating actively in the classroom discourse (available cues). The teacher needs to then be able to detect these diagnostically relevant pieces of information, which can, however, be compromised by, for instance, noisy environmental conditions (detection of cues). The teacher must then utilize this diagnostic information when judging the given student's achievement (utilization of cues). Whether relevant diagnostic cues about a student's achievement are available may strongly depend on individual student characteristics, given that students differ systematically in the extent to which they are able to communicate information about their actual abilities.

Taken together, whereas the hypothesis that certain individual characteristics enable adequate judgements to be made has been largely validated in the context of personality research (Human & Biesanz, 2013), little is known about its implications for teachers' grading practices. It might, however, be assumed that the theoretical tenets outlined above could also be used in the context of teachers' grading, where the *relevance*, *availability*, *detection,* and *utilization* of cues may play a similar role for teacher judgements and the extent to which teachers align their judgements with standardized test scores (Artelt & Rausch, 2014).

Potential moderators of the adequacy of teacher-assigned grades

When it comes to teachers' grading, it is not yet clear which student characteristics may contribute to more appropriate grading, that is to say grading that exhibits a closer relationship between teacher-assigned grades and students' standardized test scores. As Human and Biesanz' review (Human & Biesanz, 2013) focused on the adequacy of personality judgements, their findings are not directly transferrable to the context of grading. Moreover, some of the research on the adequacy of personality judgements is based on judgements made on first meeting (e.g., Paulhus & Morgan, 1997). In a study by Paulhus and Morgan (1997), for instance, participants had to judge the intelligence of previously unacquainted partners in a discussion group. The authors found that participants underestimated the intelligence of shy partners after having met them only twice. However, after seven meetings the trait of shyness was no longer relevant to how they judged their partners' intelligence. Consequently, studies on cues that are salient when first meeting people may not be relevant to most school contexts, in which teachers and students are engaged in a sustained relationship over an extended period of time.

Teachers rely on oral and written information about students' achievement when assigning grades (Martínez, Stecher, & Borko, 2009). Students who are more engaged and participate more during class, in other words are more talkative, consequently provide teachers with more information about their understanding of the classroom material. One of the Big Five personality traits, extraversion, has been associated with talkativeness (Mehl, Gosling, & Pennebaker, 2006) and individuals who exhibit a higher degree of extraversion also seem to have a higher speech rate and hesitate less when speaking under stress (Dewaele & Furnham, 1999, 2000). It is therefore highly likely that more extraverted students provide teachers with more cues about their comprehension.

In addition to this, students' self-efficacy may be crucial in their level of engagement in classroom conversations. Bandura's self-efficacy theory postulates that self-efficacy, the beliefs in one's own competence, strongly affects our effort and persistence (e.g., Bandura, 1986, 1989). Relying on the theory of planned behaviour (Fishbein & Ajzen, 2010), Girardelli, Patel, and Martins-Shannon (2017) outlined that perceived self-efficacy is a central prerequisite for students' behavioural intention to participate in class. Their findings showed that students with a higher academic self-efficacy in English also reported a higher intention to participate in class, when controlling for attitudes, subjective norms, and anxiety. In line with these results, a number of other studies have found a link between students' active participation in class and their academic self-efficacy (Gao, Lochbaum, & Podlog, 2011; Girardelli & Patel, 2016; Sánchez-Rosas, Takaya, & Molinari, 2016). Thus, students with higher academic self-efficacy may make more diagnostic cues available to their teachers.

Whether the diagnostic information that students provide, be it in oral or written forms, is a reliable indicator of these same students' actual achievement appears to be heavily reliant on another one of the Big Five personality traits, namely their conscientiousness (Kappe & van der Flier, 2010). Kappe and van der Flier (2010) examined the extent to which the Big Five personality traits are differentially relevant for specific formats of assessment in school (i.e., tests on lectures, short reports on skills trainings, team projects, evaluations in on-the-job training, and a written thesis; Kappe & van der Flier, 2010). Whereas students' extraversion and neuroticism correlated to their performance in skill trainings, while students' openness correlated to their performance in team projects, only conscientiousness was consistently related to all five assessment formats (Kappe & van der Flier, 2010). It therefore seems likely that class assessments of students' achievement may be more adequately reflective of their actual achievement if

they are thorough and diligent workers. Therefore, students' conscientiousness may have an influence on the adequacy of teachers' grading.

Present study

In this study, we aim to extend the current knowledge about teachers' grading practices by applying Funder's Realistic Accuracy Model (Funder, 1995) to the school context, aiming to deepen prior knowledge about how grading practices are affected by student characteristics. Based on a review of Funder's Realistic Accuracy Model, its implications for teacher-assigned grades, and the potential moderators for the adequacy of teacher-assigned grades, the following hypotheses guided our study.

Hypothesis 1. Students' extraversion moderates the relationship between teacher-assigned grades and students' standardized test scores. Specifically, we expect that teacher-assigned grades of more extraverted students will be more closely associated with these students' standardized test scores (than teacher-assigned grades of less extraverted students).

Hypothesis 2. Students' domain-specific self-efficacy moderates the relationship between teacher-assigned grades and students' standardized test scores. Specifically, we hypothesize that teacher-assigned grades of students with higher levels of self-efficacy are more closely associated with these students' standardized test scores (than teacher-assigned grades of students who experience lower levels of self-efficacy).

Hypothesis 3. Students' conscientiousness moderates the relationship between teacher-assigned grades and students' standardized test scores. Specifically, we expect that teacher-assigned grades of more conscientious students are more closely associated with these students' standardized test scores (than teacher-assigned grades of less conscientious students).

Methods

Sample

We used a sample of $N = 8{,}317$ seventh-grade students from the NEPS (Blossfeld, Roßbach, & von Maurice, 2011), a longitudinal multi-cohort study administered in all 16 German federal states that focuses on educational processes and competence development. We used data from the third panel wave, which took place at the beginning of seventh grade (November 2012 to January 2013).[1] For all our analyses, we excluded students from remedial schools, students for whom class IDs were missing,[2] or who were in classes with only five students or less, as well as students with missing data for each of

[1] *With the exception of teacher-assigned grades (final report cards in seventh grade) that were collected at the beginning of eighth grade (fourth panel wave: November 2013 to February 2014).*

[2] *To account for the nested data structure, we adjusted the standard errors in all our analyses (using TYPE = COMPLEX in* Mplus*), which is only possibly if class IDs are available.*

our study variables. Our final sample comprised of N = 5,919 students in 457 classes. Students were on average 12.5 years old (SD = 0.62) and 48.8% were female.

Measures

We applied the following measures in data collection.

Teacher-assigned grades in mathematics

We obtained final report-card grades in mathematics from seventh-grade students' self-reports. In the German school system, teachers award numeric grades ranging from one (denoting excellent achievement) to six (reflecting unsatisfactory achievement). For our analyses, grades were reverse-coded so that higher grades reflect better achievement.

Test scores in mathematics

Test scores for mathematics were assessed using 23 items from the content areas quantity (five items); space and shape (five items); change and relationships (seven items); and data and chance (six items), which captured different cognitive components distributed across the items (arguing, communicating, modelling, problem-solving, representing, and applying technical skills; Schnittjer & Gerken, 2017). The response format was multiple-choice (with the exception of one item requiring a short constructed response). A unidimensional partial credit model was found to fit the data well (Schnittjer & Gerken, 2017). For our analyses, we used WLE scores estimated using ConQuest (the syntax used for estimating the scores is provided in Schnittjer & Gerken, 2017). The WLE reliability was high (.72; Schnittjer & Gerken, 2017).

Big Five personality traits

Students' Big Five personality traits were assessed based on the 10-item version (BFI-10; Rammstedt & John, 2007) of the Big Five Inventory (John, Donahue, & Kentle, 1991), which has been demonstrated to be valid, reasonably stable (test–retest reliabilities; Gosling, Rentfrow, & Swann, 2003), and applicable as an alternative to the longer original Big Five Inventory. It uses items from the Big Five Inventory (with each scale comprising one positively poled and one negatively poled item). The BFI-10 asks about the extent to which a person is 'outgoing, sociable' versus 'reserved' (extraversion); 'tends to be lazy' versus 'does a thorough job' (conscientiousness); is 'relaxed, handles stress well' versus 'gets nervous easily' (emotional stability); has 'an active imagination' versus 'few artistic interests' (openness to experience); 'is generally trusting' versus 'tends to find fault with others' (agreeableness). Responses were given on a 5-point scale (1 = does not apply at all to 5 = fully applies).

Academic self-efficacy

Students' academic self-efficacy was measured using four items of the self-efficacy scale originally developed by O'Neil and Herl (1998) and adapted for mathematics in the Programme for International Student Assessment (Ramm et al., 2006). The scale is based on the theory of self-efficacy by Bandura (1989; e.g., 'In math, I'm sure that I can

understand really difficult subject matter as well.'). Students responded on a 4-point scale (1 = does not apply at all to 4 = applies completely).

Families' socioeconomic status

Based on student reports of their family backgrounds, we used the International Socio-Economic Index of Occupational Status (ISEI-08; see Ganzeboom, 2010) to measure the socioeconomic status of the students. The theoretical range of the ISEI-08 reaches from 11.74 (low SES; e.g., manual workers in agricultural sectors who lack schooling) to 88.96 (high SES; e.g., judges). For our analyses, we used the higher value of both parents (HISEI) (see Table 1 for mean value, standard error and range).

Statistical analyses

To test our hypotheses, we specified structural equation models (SEMs) with *MPlus* 7.4 (Muthén & Muthén, 2015). Initially, we specified an exploratory structural equation model (ESEM), which combines confirmatory (CFA) and exploratory factor analysis (EFA) in the same model (Marsh, Morin, Parker, & Kaur, 2014). Thus, we modelled the Big Five personality traits as EFA factors in order to ease the assumption that the secondary loadings of all items equal zero. To identify all factors, we fixed their variance to one and estimated all loadings freely. To account for acquiescence ('yes-saying'), we modelled a response style factor (Aichholzer, 2014). We applied oblique geomin rotation on the Big Five factors. Beyond the latent Big Five personality factors, we used manifest indicators to capture the latent trait of students' self-efficacy in mathematics, in order to account for measurement error. We used maximum-likelihood estimation for all SEMs (MLR-SEM). We also controlled for dependency in the data that resulted from the nested data structure (with students clustered within classes). We therefore adjusted the standard errors for the model parameters using the option type = complex in *Mplus*. To begin with, we specified a model that included grades in mathematics as dependent variable and students' standardized test scores, Big Five personality traits, and self-efficacy in mathematics as independent variables. We controlled for students' gender, age, and SES, all of which that have been shown to be associated with teacher-assigned grades (Westphal et al., 2016; Cobley, McKenna, Baker, & Wattie, 2009). We then included the interaction terms between students' standardized test scores and, firstly, students' extraversion, secondly, students' self-efficacy in mathematics, and, thirdly, students' conscientiousness. On

Table 1. Descriptive statistics and intercorrelations for observed and latent constructs

	M	*SE*	Scale	2	3	4	5
(1) Grades in math[a]	4.29	0.02	1–6	.40***	.53***	.20***	−.12***
(2) Test score math	0.03	0.03	−4.44 to 3.30	.33***	.34***	−.20***	
(3) Self-efficacy math	2.69	0.02	1–4			.08***	−.03
(4) SES	55.82	0.59	11.74–88.96			−.21***	
(5) Age	12.51	0.01					

Note. For self-efficacy, latent means and standard errors are reported. For observed variables (variables no. 1–2 and 4–5), means and standard errors are reported.
SES = Socioeconomic Status.
[a]Grades are reverse-coded so that higher grades reflect better achievement.

Table 2. Relations between student characteristics, test scores, and teacher-assigned grades

	β	p	95% CI
Intercept	−.05	.073	[−.10, .00]
Test score math	.23	.000	[.20, .26]
Self-efficacy math	.45	.000	[.42, .48]
Extraversion	.00	.825	[−.04, .03]
Conscientiousness	.10	.000	[.06, .14]
Openness	−.01	.682	[−.05, .03]
Agreeableness	−.06	.037	[−.11, −.00]
Emotional stability	.00	.993	[−.04, .04]
SES	.08	.000	[.04, .11]
Gender[a]	.06	.000	[.03, .10]
Age	−.03	.009	[−.06, −.01]
R^2	36.7%		

Note. Coefficients are standardized.
SES = Socioeconomic Status.
RMSEA = .033. CFI = .976. SRMR = .014.
[a]Reference: male.

average, the percentage of missing values on all observed variables was 8.4%. We applied the full-information maximum-likelihood approach (FIML; Enders, 2001) to acquire appropriate estimates and standard errors.

Results

Our ESEM analyses showed low to moderate correlations between the five dimensions ($r \leq .31$).[3] The factor loading matrix of the Big Five personality traits is reported in supplementary Table 1. Initially, we found a correlation of $r = .40$ between teacher-assigned grades and standardized test scores in mathematics. All coefficients from the model that included teacher-assigned grades in mathematics as dependent variable are depicted in Table 2. The results show that standardized test scores in mathematics ($\beta = .23$, $p < .001$) and self-efficacy in mathematics ($\beta = .45$, $p < .001$) were strongly related to teacher-assigned grades in mathematics. Specifically, students whose test scores in mathematics were one *SD* above the average received mathematics grades that were one *SD* above the average, when controlling for all the other variables in the models. Students who reported higher academic self-efficacy were graded more favourably. Students' conscientiousness ($\beta = .10, p < .001$) and agreeableness ($\beta = -.06, p = .037$) were also statistically significantly related to teacher-assigned grades in mathematics. Thus, more conscientious students received better grades as did less agreeable students. The remaining Big Five personality traits – extraversion, openness, and emotional stability – did not statistically significantly relate to teacher-assigned grades in mathematics. In addition, the effects of gender, age, and SES were also statistically significant and ranged from $\beta = .03$ to $\beta = .08$ (see Table 2). Altogether, the model explained 36.7% of the variance in teacher-assigned grades in mathematics.

[3] *Descriptive results for the Big Five personality traits are not reported as latent means cannot be estimated within an ESEM.*

Do extraversion, self-efficacy, and conscientiousness moderate the relationship between teacher-assigned grades and standardized test scores?

In order to test our hypotheses that the relationship between teacher-assigned grades and standardized test scores are moderated by students' extraversion, self-efficacy, and conscientiousness, we included interaction effects between these variables in our SEM. Coefficients of this model are reported in Table 3. The results showed a statistically significant interaction effect between students' conscientiousness and standardized test scores in mathematics ($\beta = .05$, $p = .002$). Thus, teacher-assigned grades and standardized test scores were more closely related, when students' conscientiousness was higher. However, the relationship between teacher-assigned grades and standardized test scores was statistically significant at all levels of students' conscientiousness ($\beta_{\text{conscientiousness } 1SD\text{below mean}} = .176$, $p < .001$; $\beta_{\text{average conscientiousness}} = .230$, $p < .001$; $\beta_{\text{conscientiousness } 1SD\text{above mean}} = .284, p < .001$). Figure 1 illustrates this moderating effect by depicting the relationship between teacher-assigned grades and students' standardized test scores in mathematics (residualized by the other predictors of the regression model) at different levels of students' conscientiousness (1 *SD* below average, average, 1 *SD* above average). In contrast, the interaction effect between students' extraversion and standardized test scores in mathematics was not statistically significant ($\beta = -.02$, $p = .228$), neither was the interaction between students' self-efficacy and standardized test scores in mathematics ($\beta = -.01$, $p = .548$).

Discussion

The question 'what's in a grade' (Bowers, 2011, p. 141) has occupied researchers for several decades. Textbooks on assessing students advise teachers to base teacher-assigned grades primarily on students' achievement in class (e.g., Brookhart, 2004; Linn & Miller, 2005). However, certain characteristics of students' personalities and behaviour may 'help or hinder' (Human & Biesanz, 2013, p. 252) teachers in aligning teacher-assigned grades to students' actual achievement. Based on Funder's Realistic Accuracy Model (Funder, 1995) and its application on teacher-assigned grades (Artelt & Rausch, 2014), the present study examined whether some students are graded more appropriately than others. Specifically, we investigated whether students' extraversion, self-efficacy, or conscientiousness moderate the relationship between teacher-assigned grades and standardized test scores.

Initially, we found that teacher-assigned grades and standardized test scores in mathematics were correlated to a moderate degree ($r = .40$). Thus, the strength of the association is somewhat lower than that reported for other countries, such as the United States ($.50 \leq r \leq .60$; Bowers, 2011). Previous studies conducted in Germany already reported correlations between grades and standardized test scores in mathematics that lay outside of this range (e.g., $r = .34$ in a sample of German secondary school students, Hochweber et al., 2014). Moreover, teacher-assigned grades and written comparison tests – designed to assess the components from the German national educational standards for mathematics in secondary school – are correlated at about 0.45 to 0.49 (Nachtigall, 2015, 2018). For the present study, we used standardized test scores assessed within the NEPS. The standardized mathematics test applied in the NEPS combines the German national educational standards for mathematics and the framework of the PISA studies (Neumann et al., 2013). Thus, differences in the strength of the association between grades and standardized test scores in our study, as compared to the findings of Nachtigall (2015,

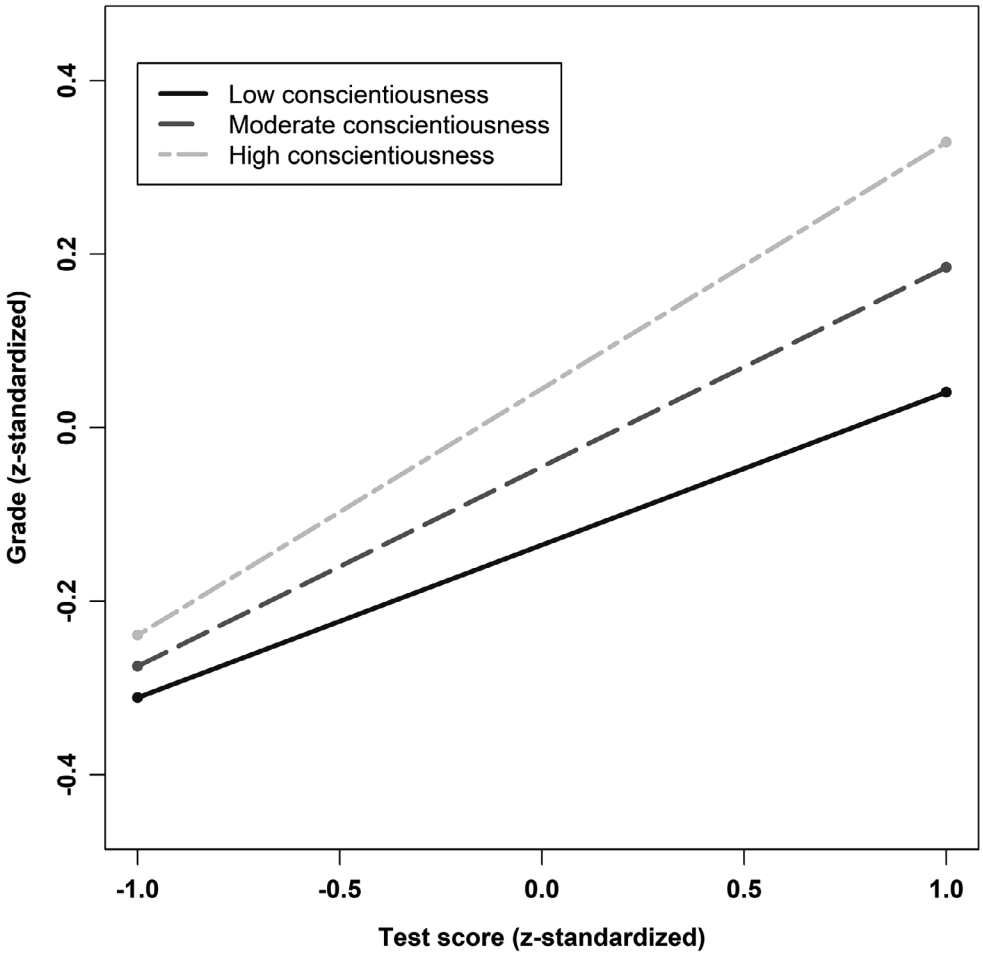

Figure 1. Conscientiousness as a moderator of the relationship between test scores and teacher-assigned grades (residualized for the other factors in the regression model). *Note*. Low conscientiousness = one standard deviation below average. Moderate conscientiousness = average. High conscientiousness = one standard deviation above average.

2018), may be explained by differences in the curricular validity of the standardized tests (see also Bowers, 2011).

Confirming one of our hypotheses, we found that teacher-assigned grades in mathematics were more closely aligned with students' standardized mathematics test scores for students who reported higher levels of conscientiousness. Students who are more conscientious exhibit more effortful control (De Pauw, Mervielde, & Van Leeuwen, 2009) and self-control (MacCann, Duckworth, & Roberts, 2009). Therefore, it is plausible that the oral and written coursework that teachers rely on when assigning grades (Martínez et al., 2009) may better reflect students' actual achievement the more conscientious the students are. Although our interaction effect was small, it may indicate, on the basis of Funder's Realistic Accuracy Model (Funder, 1995), that more conscientious

Table 3. Moderation of the test score–grade relationship

	Extraversion as moderator			Self-efficacy as moderator			Conscientiousness as moderator		
	β	p	95% CI	β	p	95% CI	β	p	95% CI
Intercept	−.05	.087	[−.10, .01]	−.04	.108	[−.10, .01]	−.05	.088	[−.10, .01]
Test score math	.23	.000	[.20, .27]	.23	.000	[.20, .27]	.23	.000	[.20, .26]
Self-efficacy math	.58	.000	[.54, .62]	.58	.000	[.54, .62]	.59	.000	[.54, .63]
Extraversion	.00	.802	[−.04, .03]	−.01	.790	[−.04, .03]	.00	.858	[−.04, .03]
Conscientiousness	.10	.000	[.06, .14]	.10	.000	[.06, .14]	.09	.000	[.05, .13]
Openness	−.01	.688	[−.05, .03]	−.01	.647	[−.05, .03]	−.01	.619	[−.05, .03]
Agreeableness	−.05	.043	[−.10, −.00]	−.05	.048	[−.10, .00]	−.05	.061	[−.10, .00]
Emotional stability	.00	.959	[−.04, .04]	.00	.984	[−.04, .04]	.00	.967	[−.04, .04]
SES	.08	.000	[.04, .11]	.08	.000	[.04, .11]	.08	.000	[.04, .11]
Gender[a]	.13	.000	[.06, .19]	.13	.000	[.06, .19]	.13	.000	[.06, .19]
Age	−.03	.012	[−.06, −.01]	−.03	.011	[−.06, −.01]	−.03	.010	[−.06, −.01]
Test score–grade relationship (slope)									
Extraversion	−.02	.228	[−.06, .01]						
Self-efficacy math				−.01	.548	[−.04, .02]			
Conscientiousness							.05	.002	[.02, .09]

SES = Socioeconomic Status.

[a]Reference: male.

students reveal a greater amount of relevant diagnostic cues about their achievement in mathematics, which teachers in turn are able to use in their grading.

In contrast, neither students' extraversion nor students' self-efficacy in mathematics moderated the relationship between standardized test scores and teacher-assigned grades. On the basis of Funder's Realistic Accuracy Model (Funder, 1995), we argued that more extraverted students and more self-efficient students may provide more information about themselves to their teachers (*availability of cues*). There is some indication that more extraverted individuals talk more than more introverted individuals (Mehl et al., 2006), have a higher speech rate, and hesitate less when speaking in stressful situations (Dewaele & Furnham, 1999, 2000). Similarly, students with higher self-efficacy may participate more often in class (Gao et al., 2011; Girardelli & Patel, 2016; Girardelli et al., 2017; Sánchez-Rosas et al., 2016). In the case of academic self-efficacy, future studies should aim to illuminate whether a more active involvement in class activities, as measured by student self-reports, is indeed visible to teachers and, thus, truly implies a higher *availability of diagnostic cues*. In the case of extraversion, future studies should aim to clarify whether extraverted students do indeed engage more intensively in conversations concerning class material or merely talk more about non-class-related topics. This raises the question of whether these cues provided by more extraverted students are indeed reliable indicators of students' actual achievement (*relevance of cues*).

Limitations and implications for future research

Our study has several limitations. Firstly, we assessed students' Big Five personality traits using the BFI-10 (Rammstedt & John, 2007). Rammstedt and John (2007) provided evidence on the reliability and validity of the BFI-10 and found that it captured 70% of the variance of the long version of the Big Five Inventory. At the same time, our study used a narrower measure of students' conscientiousness, which may potentially cause an underestimation of its role in educational outcomes (Credé, Harms, Niehorster, & Gaye-Valentine, 2012). Secondly, there is an alternative interpretation for the moderating effects of students' conscientiousness. Thus, we cannot completely rule out the possibility that more conscientious students achieve scores in standardized tests that more adequately reflect their actual achievement, which in turn explains why these students' test scores are more closely aligned with teacher grading in mathematics. However, students' test scores in mathematics were not correlated with conscientiousness (r = .01; see Table 1) and therefore this interpretation seems questionable. Thirdly, we used a sample of seventh-grade students and focused on the subject of mathematics, which is why our results are not generalizable to older students or to students in primary school or who are studying other subjects. Therefore, more extensive research in more diverse samples and for different subjects would be enlightening, alongside the application of broader measures of students' conscientiousness.

Practical implications and conclusions

The present study represents a decisive contribution to better understanding teachers' grading practices, by applying Funder's Realistic Accuracy Model (Funder, 1995) to grading practice in schools. We were able to show that specific student personality traits may affect students' 'judgeability' (Human & Biesanz, 2013, p. 252), by moderating to some extent the relation between teacher-assigned grades and students' actual

standardized test scores. Our findings indicate a need to further examine under which conditions teachers are able to most appropriately evaluate the achievement of their students and which factors lead to potential judgement biases. While our results need to be replicated in future studies, they may suggest that teacher training should put more emphasis on judging student achievement and bias and sensitize teachers and trainee teachers to those characteristics in students that may complicate appropriate grading. In terms of theory-development, our study contributes to a further validation of Funders' model by showing that theory-based predictions can be confirmed in different contexts.

Acknowledgments

We thank Ben Fergusson for his editorial assistance. This study used data from the National Educational Panel Study (NEPS) Starting Cohort Grade 5 (https://doi.org/10.5157/NEPS:SC3:6.0.1). From 2008 to 2013, NEPS data were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Conflicts of interest

All authors declare no conflict of interest.

Author contributions

Andrea Westphal (Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing) Rebecca Lazarides (Writing – review & editing) Miriam Vock (Writing – review & editing).

Data Availability Statement

We used data from the German National Educational Panel Study (NEPS). The data of the German NEPS is prepared and disseminated in the form of Scientific Use Files to the scientific community by the Research Data Center at the Leibniz Institute for Educational Trajectories (RDC-LIfBi). The data sets are available for download from the study website: https://www.neps-data.de/Data-Center/Data-Access/Download

References

Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality*, *53*, 1–4. https://doi.org/10.1016/j.jrp.2014.07.001

Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgments. In S. Krolak-Schwerdt, S. Glock & M. Böhner (Eds.), *Teachers' professional development* (pp. 27–43). Rotterdam, Netherlands: SensePublishers.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1989). Regulation of cognitive processes through perceived self-efficacy. *Developmental Psychology*, *25*, 729–735. https://doi.org/10.1037/0012-1649.25.5.729

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS). *Zeitschrift Für Erziehungswissenschaft: Sonderheft*, Vol. *14*, Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation*, *17*, 141–159. https://doi.org/10.1080/13803611.2011.597112

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement*, *30*, 123–142. https://doi.org/10.1111/j.1745-3984.1993.tb01070.x

Brookhart, S. M. (2004). *Grading*. Upper Saddle River, NJ: Pearson.

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*(1), 78–96. https://doi.org/10.1348/2044-8279.002004

Cobley, S., McKenna, J., Baker, J., & Wattie, N. (2009). How pervasive are relative age effects in secondary school education? *Journal of Educational Psychology*, *101*, 520–528. https://doi.org/10.1037/a0013845

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the big five personality traits. *Journal of Personality and Social Psychology*, *102*, 874–888. https://doi.org/10.1037/a0027403

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*, 309–325. https://doi.org/10.1007/s10802-008-9290-0

Dewaele, J.-M., & Furnham, A. (1999). Extraversion: The unlovedvariable in applied linguistic research. *Language Learning*, *49*, 509–544. https://doi.org/10.1111/0023-8333.00098

Dewaele, J.-M., & Furnham, A. (2000). Personality and speech production: A pilot study of second language learners. *Personality and Individual Differences*, *28*, 355–365. https://doi.org/10.1016/S0191-8869(99)00106-3

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*, 352–370. https://doi.org/10.1037/1082-989X.6.4.352

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York, NY: Psychology Press.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*, 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Ganzeboom, H. B. G. (2010). *A new international socio-economic index (ISEI) of occupational status for the international standard classification of occupation 2008 (ISCO-08) constructed with data from the ISSP 2002-2007; With an analysis of quality of occupational measurement in ISSP*. Paper presented at the Annual Conference of International Social Survey Programme (ISSP), Lisbon, Portugal. Retrieved from http://www.harryganzeboom.nl/pdf/2010-ganzeboom-isei08-issp-lisbon-(paper).pdf

Gao, Z., Lochbaum, M., & Podlog, L. (2011). Self-efficacy as a mediator of children's achievement motivation and in-class physical activity. *Perceptual and Motor Skills*, *113*, 969–981. https://doi.org/10.2466/06.11.25.PMS.113.6.969-981

Garcia, S. M., Hallahan, M., & Rosenthal, R. (2007). Poor expression: Concealing social class stigma. *Basic and Applied Social Psychology*, *29*, 99–107. https://doi.org/10.1080/01973530701330835

Girardelli, D., & Patel, V. (2016). The theory of planned behavior and Chinese ESL students' in-class participation. *Journal of Language Teaching and Research*, *7*(1), 31–41. https://doi.org/10.17507/jltr.0701.0

Girardelli, D., Patel, V. K., & Martins-Shannon, J. (2017). "Crossing the Rubicon": Understanding Chinese EFL students' volitional process underlying in-class participation with the theory of planned behaviour. *Educational Research and Evaluation*, *23*(3–4), 119–137. https://doi.org/10.1080/13803611.2017.1398668

Gosling, S. D., Rentfrow, P. J., & Swann, Jr., W. B. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality*, *37*, 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*, 348–362. https://doi.org/10.1037/0022-3514.85.2.348

Hall, J. A., Rosip, J. C., LeBeau, L. S., Horgan, T. G., & Carter, J. D. (2006). Attributing the sources of accuracy in unequal-power dyadic communication: Who is better and why? *Journal of Experimental Social Psychology*, *42*(1), 18–27. https://doi.org/10.1016/j.jesp.2005.01.005

Hallinan, M. T. (1992). The organization of students for instruction in the middle school. *Sociology of Education*, *65*, 114–127. https://doi.org/10.2307/2112678

Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology*, *106*(1), 289–300. https://doi.org/10.1037/a0033829

Human, L. J., & Biesanz, J. C. (2013). Targeting the good target: An integrative review of the characteristics and consequences of being accurately perceived. *Personality and Social Psychology Review*, *17*, 248–272. https://doi.org/10.1177/1088868313495593

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.

Kappe, R., & van der Flier, H. (2010). Using multiple and specific criteria to assess the predictive validity of the Big Five personality factors on academic performance. *Journal of Research in Personality*, *44*(1), 142–145. https://doi.org/10.1016/j.jrp.2009.11.002

Krejtz, I., & Nezlek, J. B. (2016). It's Greek to me: Domain specific relationships between intellectual helplessness and academic performance. *The Journal of Social Psychology*, *156*(6), 664–668. https://doi.org/10.1080/00224545.2016.1152219

Lang, J. W., & Lang, J. (2010). Priming competence diminishes the link between cognitive test anxiety and test performance: Implications for the interpretation of test scores. *Psychological Science*, *21*, 811–819. https://doi.org/10.1177/0956797610369492

Linn, R., & Miller, M. (2005). *Measurement and assessment in teaching*. Upper Saddle River, NJ: Pearson Prentice Hall.

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences*, *19*, 451–458. https://doi.org/10.1016/j.lindif.2009.03.007

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*(1), 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment*, *14*, 78–102. https://doi.org/10.1080/10627190903039429

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20–32. https://doi.org/10.1111/j.1745-3992.2001.tb00055.x

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, *90*, 862–877. https://doi.org/10.1037/0022-3514.90.5.862

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nachtigall, C. (2015). Landesbericht. Thüringer Kompetenztests 2015 [State report. Written comparison tests in Thuringia]. Retrieved from https://www.kompetenztest.de/downloads/kompetenztests/archive

Nachtigall, C. (2018). Landesbericht. Thüringer Kompetenztests 2018 [State report. Written comparison tests in Thuringia]. Retrieved from https://www.kompetenztest.de/downloads/kompetenztests/archive

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online/Journal Für Bildungsforschung Online*, *5*, 80–109.

O'Neil, H. F., & Herl, H. E. (1998). *Reliability and validity of a trait measure of self-regulation*. Presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Paulhus, D. L., & Morgan, K. L. (1997). Perceptions of intelligence in leaderless groups: The dynamic effects of shyness and acquaintance. *Journal of Personality and Social Psychology*, *72*, 581–591. https://doi.org/10.1037//0022-3514.72.3.581

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., . . . Schiefele, U. (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente [PISA 2003: Documentation of the measures]*. Münster, Germany: Waxmann.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41* (1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Randall, J., & Engelhard, Jr., G. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement*, *46*(1), 1–18. https://doi.org/10.1111/j.1745-3984.2009.01066.x

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education*, *26*, 1372–1380. https://doi.org/10.1016/j.tate.2010.03.008

Sánchez-Rosas, J., Takaya, P. B., & Molinari, A. V. (2016). The role of teacher behavior, motivation and emotion in predicting academic social participation in class. *Pensando Psicología*, *12*, 39–53. https://doi.org/10.16925/pe.v12i19.1327

Schnittjer, I., & Gerken, A.-L. (2017). *NEPS technical report for mathematics: Scaling results of starting cohort 3 in grade 7* (NEPS Survey Paper No. 16). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Spengler, M., Lüdtke, O., Martin, R., & Brunner, M. (2013). Personality is related to educational outcomes in late adolescence: Evidence from two large-scale achievement studies. *Journal of Research in Personality*, *47*, 613–625. https://doi.org/10.1016/j.jrp.2013.05.008

Tetzner, J., Becker, M., & Brandt, N. D. (2020). Personality-achievement associations in adolescence – Examining associations across grade levels and learning environments. *Journal of Personality*, *88*, 356–372. https://doi.org/10.1111/jopy.12495

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Westphal, A., Becker, M., Vock, M., Maaz, K., Neumann, M., & McElvany, N. (2016). The link between teacher-assigned grades and classroom socioeconomic composition: The role of classroom behavior, motivation, and teacher characteristics. *Contemporary Educational Psychology*, *46*, 218–227. https://doi.org/10.1016/j.cedpsych.2016.06.004

Westphal, A., Vock, M., & Lazarides, R. (2020). Are more conscientious seventh-and ninth-graders less likely to be retained? Effects of Big Five personality traits on grade retention in two different age cohorts. *Journal of Applied Developmental Psychology*, *66*, 101088. https://doi.org/10.1016/j.appdev.2019.101088

Received 15 June 2020; revised version received 5 November 2020

Appendix :

Factor Loadings and Factor Correlations of BFI-10 Scales

I see myself as someone who . . .	O	C	E	A	ES
Factor loadings					
has an active imagination	**0.38**	−0.07	0.15	0.02	0.00
has few artistic interests	−**0.64**	−0.05	0.03	0.01	−0.01
does a thorough job	0.06	**0.52**	0.01	0.06	0.01
tends to be lazy	0.01	−**0.75**	−0.01	0.01	0.01
is outgoing, sociable	−0.01	0.02	**0.72**	0.01	0.01
is reserved	0.01	0.01	−**0.33**	0.19	−0.25
is generally trusting	0.07	0.00	0.18	**0.31**	−0.14
tends to find fault with others	0.04	−0.17	0.04	−**0.46**	−0.04
is relaxed, handles stress well	0.02	−0.09	0.01	0.20	**0.38**
gets nervous easily	0.00	−0.05	−0.01	0.00	−**0.66**
Factor correlations					
C	.17				
E	.18	.03			
A	.23	.31	−.10		
ES	−.12	.05	.28	−.06	

Note. Standardized loadings of the exploratory structure equation model controlling for acquiescence. Target loadings are printed bold. O = Openness; C = Conscientiousness; E = Extraversion; A = Agreeableness; ES = Emotional Stability.