Agreement Among Adolescents, Parents, and Teachers on Adolescent Personality

Kaia Laidra
Jüri Allik
Maarike Harro
Liis Merenäkk
Jaanus Harro
University of Tartu
The Estonian Centre of Behavioural and Health Sciences

Agreement between adolescents, mothers, fathers, and teachers on adolescents' personality traits was investigated in a longitudinal study. The targets for personality ratings were the adolescents who participated in the European Youth Heart Study in Estonia. There were 593 participants in the first wave and 480 participants in the follow-up study 3 years later. Adolescents' self-reports as well as father, mother, and teacher ratings were collected using questionnaires to measure the five-factor model of personality. In both waves, interrater agreement was highest between mothers and fathers, was low to moderate for parent-self ratings, and was lowest for ratings between self and teacher, mother and teacher, and father and teacher. Test-retest correlations were moderate for parent and self-ratings but failed to reach statistical significance for three of the five teacher-rated traits, suggesting lower reliability of teacher ratings. Possible explanations for the low agreement between teachers and other judges are discussed.

Keywords: adolescent personality; five-factor model; interrater agreement; self-reports; teacher ratings; parent ratings; personality development

Personality judgments made by close acquaintances like near relatives, spouses, or friends tend to be reasonably accurate. The agreement between two judges who know the target well, or judges and the target, often yields consensus correlation of about .50 or even higher (Funder, 1999). One unexpected result of personality judgments is that it takes surprisingly little time to develop consensus between different judges (Kenny, 1994). Many studies have shown that judges who have been acquainted with strangers for only 5 to 10 minutes can deduce from these brief encounters a sufficient amount of personality information, allowing them to obtain statistically significant agreement with the target's self-description (Ambady, Hallahan, & Rosenthal, 1995; Borkenau & Liebler, 1993; Funder & Colvin, 1988; Watson, 1989). Nevertheless, many controlled experiments have shown that more information leads to more accuracy. Agreement among judges of personality improves with the duration and quality of acquaintanceship; well-acquainted informants agree to a much greater extent with each other and with their targets about the personality traits of the target than do relative or complete strangers (Banai, Weller, & Mikulincer, 1998; Borkenau, Mauer, Riemann, Spinath, & Angleitner, 2004; Colvin & Funder, 1991; Funder, Kolar, & Blackman, 1995). Even less visible personality traits become more readily judged when the judge has the opportunity to know the target longer and more intimately (Paunonen, 1989).

Correspondence concerning this article should be addressed to Kaia Laidra, Department of Psychology, University of Tartu, Tiigi 78, Tartu 50410, Estonia; e-mail: kaia.laidra@ut.ec.

Assessment, Volume 13, No. 2, June 2006 187-196
DOI: 10.1177/1073191106287125

Teacher ratings about children's personality are often used, frequently because they can be collected more easily than responses by other observers. Teachers interact with a large number of different children and thus have a broad frame of reference on which to base their responses. In a recent study, Baker, Victor, Chambers, and Halverson (2004) demonstrated higher validity and reliability of teacher ratings on trait markers of the five-factor personality dimensions compared with adolescents' self-ratings. In particular, this study compared the personality ratings provided by four teachers and 163 students. Convergent correlations between teacher ratings and self-reports for the five dimensions were moderate (from .37 to .57), except for a low correlation of .15 for emotional stability.

Other evidence suggests that ratings made by mothers and fathers are more valid and accurate than teachers' judgments (Marsh & Craven, 1991). Multiple studies have demonstrated that teachers are not particularly accurate in judging their pupils' academic achievements (Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003) and are even worse in judging their pupils' personalities. For example, Miller and Davis (1992) found that in judging cognitive abilities, teachers were as accurate as mothers; they were less successful, however, in rating personality traits. Similarly, Barbaranelli, Caprara, Rabasca, and Pastorelli (2003) analyzed the convergence of teacher ratings, mother ratings, and self-reported ratings using the Big Five Questionnaire for Children. They reported higher agreement between self-reports and mother ratings than between self-reports and teacher ratings. Although there was a high convergence among mothers and teachers for conscientiousness and intellect/openness, the other convergent correlations were lower. In summary, it appears that teachers are not particularly accurate in the estimation of personality traits that are not directly relevant to behavior in the classroom and to academic achievement. For example, one recent study found that, with the exception of "troublesomeness," there was little correspondence between teachers' ratings of pupils and the behavior of those pupils in the classroom (as observed independently by naive observers; ter Laak, DeGoede, & Brugman, 2001).

Another line of evidence suggesting that teachers' opinions of children do not converge with those of other raters comes from numerous studies of interrater agreement on problem behavior or psychopathology (e.g., Achenbach, McConaughy, & Howell, 1987; Stanger & Lewis, 1993; Youngstrom, Loeber, & Stouthamer-Loeber, 2000). Achenbach et al. (1987) conducted a meta-analysis of 119 such studies and found that whereas mean correlations between parents were .60, there was only a moderate correlation between parent and teacher reports (r = .27) and between self-reports and observer ratings (r = .22). Stanger and Lewis (1993) examined the agreement between mothers, fathers, teachers, and 13-year-old children on internalizing and externalizing behavior problems. Results revealed that the highest agreement existed between mothers and fathers, and lowest levels of agreement emerged for rater pairs involving teachers. It has also been repeatedly found that agreement is higher for externalizing versus internalizing behavior problems (Achenbach et al., 1987; Duhig, Renk, Epstein, & Phares, 2000).

One explanation for these discrepancies is that although students and teachers share a common classroom for a considerable period of time, they may still live in "separate worlds." For example, being in a classroom has the potential to alter children's behavior; an aggressive child may be more likely to suppress his or her hostility in a classroom than when observed on a playground. A typical classroom setting also seems to belong to the category of what Mischel (1977) has called a "strong situation," characterized by high situational constraints and only few available behavioral choices, thereby providing little information about less visible personality traits. De Raad (1996) compared teachers' ratings with ratings made by lay persons and found that three of the Big Five personality dimensions may be relevant in an educational context. Whereas extraversion, conscientiousness, and openness were considered to be "educational" traits with relevance to educational environment, neuroticism and agreeableness may have less importance.

The purpose of this study was to compare personality judgments made by adolescents, their mothers and fathers, and their school teachers. Because parents generally share the information they have about their child with one another and have the opportunity to observe the child's behavior in a wide range of situations, whereas teachers' observations of the child are limited to the classroom context, it was hypothesized that parents' judgments would be more similar to each other than they would be to teachers' judgments. It was also proposed that teachers' judgments would be in better agreement with other raters on educational traits that are relevant to classroom behavior than on those personality dimensions that are not as relevant in an academic setting.

Data were collected as a part of the European Youth Heart Study in Estonia, an ongoing longitudinal study designed to examine risk factors for cardiovascular diseases, which also included the assessment of participants' personality traits. This article will present only the results concerning personality ratings given by different raters in two waves, when the participants were approximately 15 and 18 years old.

Schooling in Estonia

Compulsory schooling in Estonia begins at the age of 7. Basic school (grades 1-9) is obligatory for all children and is usually followed by either upper secondary school

(grades 10-12) or vocational secondary school. Classroom size varies considerably depending on the location of the school, ranging from fewer than 10 pupils per class in rural areas to approximately 35 pupils in cities. The same students are together in most of their classes throughout the school years. New classes are formed only in the transition from basic to secondary school (i.e., in the beginning of 10th grade). Because most schools teach both basic and secondary programs, many pupils continue their studies in the same school. From 5th to 12th grade, different academic subjects are taught by different teachers. However, each class has a homeroom teacher who not only teaches a particular subject but is also more connected with students, overlooks their general educational process, and maintains links with families, among other things. Students may have the same homeroom teacher for all their years in school, but usually the homeroom teacher changes when students move to the upper secondary school.

METHOD

Participants

The targets for personality judgment in this study were children who participated in the European Youth Heart Study in Estonia. The study was approved by the ethical committee of the University of Tartu (protocol no. 49/30-1997).

In Wave 1 (1998), 25 schools from Tartu County, Estonia, were sampled and all ninth graders were invited to participate. Written consent was obtained from children and their parents. Of all those invited to the study (n = 770), 77% (i.e., 593 children; 333 girls and 260 boys) agreed to participate. The mean age of the participants was 15.5 ± 0.6 (mean ± *SD*) years, ranging from 14 to 17 years old. Detailed sampling procedures are described by Harro et al. (2001).

Three years later (2001; Wave 2), a follow-up study was performed with 480 adolescents (417, or 70% of those who participated in the first wave, plus 63 adolescents who did not participate in 1998). The mean age of adolescents in the second wave was 18.2 ± 0.7 years, ranging from 16 to 20 years old. Because the researchers did not have access to population registers, it was not possible to locate all those who had participated in the first wave. Of those 176 adolescents who participated in Wave 1 but not Wave 2, 85 could not be contacted because they had transferred schools and/or changed their addresses, 89 refused to participate due to various reasons (e.g., poor health, parents did not give their consent, they were in the army, etc.; however, the majority did not specify reasons), and two were deceased. Because of the high attrition rate, an additional 63 students who had not participated in Wave 1 were recruited from among participants' classmates to participate in Wave 2. Differences between the participants in the two waves are addressed in the Results section.

Personality Assessment

Personality traits were measured by questionnaires based on the five-factor model. In Wave 1, all informants filled out the Estonian Brief Big Five Inventory. In Wave 2, self-reports from the adolescents were collected via the Revised NEO Personality Inventory, whereas all other informants completed the Estonian Brief Big Five Inventory a second time.

Estonian Brief Big Five Inventory (EBBFI). Because a brief measure of personality was needed, a 40-item Estonian Brief Big Five Inventory was constructed especially for this study following the example of the "Common Language" California Child Q-Set (John, Caspi, Robins, Moffitt, & Stouthamer-Loeber, 1994). Each of the five basic personality dimensions (extraversion, neuroticism, openness, agreeableness, and conscientiousness) was measured by eight items on a 5-point Likert-type scale. At face value, all of the items on the EBBFI refer to behavior tendencies that can be observed in the classroom.

Revised NEO Personality Inventory (NEO-PI-R). In the second wave, adolescents provided self-reports via the Estonian version of the Revised NEO Personality Inventory (Costa & McCrae, 1992; Kallasmaa, Allik, Realo, & McCrae, 2000). The NEO-PI-R is a 240-item measure of the five basic personality factors, where each factor is represented by six 8-item facet scales. Items are answered on a 5-point Likert-type scale. The NEO-PI-R was used instead of the EBBFI because of its superior psychometric properties and its ability to measure personality at a more specific level.

The EBBFI has been validated in relation to the NEO-PI-R on a separate adult sample (n = 142, mean age = 29.8 ± 15.4 years) that simultaneously completed the two questionnaires. The convergent correlations of the five scales were .71 for neuroticism, .72 for extraversion, .55 for openness, .52 for agreeableness, and .60 for conscientiousness. We consider the convergence of the two personality measures to be acceptable given that the convergent correlations are in the typical range of intercorrelations among the facets of the NEO-PI-R that measure the same trait (the mean correlation between the subscales measuring the same trait was .38 for the matrix of intercorrelations of the American normative sample; Costa & McCrae, 1992, Appendix F).

Procedure

The personality questionnaires were completed by adolescents as well as their mothers, fathers, and classroom

teachers. Unfortunately, it was not possible to obtain personality data from all four sources for each participant in the Heart Study. Although there was a total of 593 adolescents participating in Wave 1 and 480 in Wave 2, not all of them provided self-reports in addition to being rated by teachers and both parents. Parental questionnaires were distributed in schools, and adolescents were asked to give them to their parents and to return the completed questionnaires in sealed envelopes. Mothers of 482 (81.5%) and fathers of 366 (61.7%) participants completed questionnaires in Wave 1, versus 404 (84.2%) mothers and 288 (60.0%) fathers who provided parental ratings in Wave 2.

Teachers were contacted personally and were asked to judge the personality of all their students participating in the study. There were 68 teachers who provided personality judgments of a total of 489 students (82.5% of all participants) in the first wave and 90 teachers who rated 313 students (65.2%) in the second wave. The number of targets per teacher varied according to the number of students participating from each class, with a mean of 6.8 targets (SD = 4.3) per teacher in Wave 1 and 3.5 targets (SD = 5.8) in Wave 2. Because of the transition from basic to secondary or vocational school, in most cases the teachers who rated students in Wave 2 were not the same teachers who had rated them in Wave 1, with the exception of 58 students. In Wave 2, teachers were also asked about how many years they had taught each student. On the average, they had taught the students for 3.8 years (SD = 2.4 years), ranging from 0.3 to 9 years.

Self-reports were completed in the laboratory where other procedures of the Heart Study were carried out. Self-report data were available for 276 adolescents (46.5%) in Wave 1 and 436 adolescents (90.8%) in Wave 2.

RESULTS

Preliminary Analyses

t tests were conducted to evaluate differences in personality scores between the students who participated both times and students who participated in only Wave 1 or Wave 2. The adolescents who participated only in Wave 1 scored significantly lower on conscientiousness at age 15 according to teachers, mothers, and fathers ($p < .05$) compared with the adolescents who participated in both waves. There were no differences between the two groups on self-rated conscientiousness, nor on other personality traits. The small group of students that did not participate in Wave 1 but was included in Wave 2 was rated as less neurotic, more extraverted, and more conscientious by their teachers ($p < .05$) compared with the students who participated in both waves. There were no differences between the two groups according to the ratings of other informants.

Mean Differences

Table 1 reports basic descriptive statistics for personality ratings given by different raters in the two study waves. Internal consistency of the EBBFI can be considered satisfactory, with a mean coefficient alpha of .76. Only father-rated openness in the follow-up study demonstrated relatively low internal consistency ($\alpha = .45$). Coefficient alphas for the five domains of the NEO-PI-R are comparable with those reported by Kallasmaa et al. (2000) for the Estonian normative adult sample.

There are several statistically significant differences in the mean scores provided by different judges as indicated by pairwise comparison with t test for dependent samples; however, these differences are quite small in magnitude. Because the NEO-PI-R was used for self-reports in 2001, self-ratings and observer ratings in Wave 2 are not directly comparable. In Wave 1, the mean self-ratings are different from teacher ratings on three dimensions: teachers see students as less extraverted, $t(231) = –4.36$, $p = .00$, Cohen's $d = –0.34$, and open, $t(231) = –5.32$, $p = .00$, $d = –0.43$, but more agreeable, $t(231) = 4.13$, $p = .00$, $d = 0.35$, than students themselves. Both mothers and fathers rate their children to be less neurotic, $t(234) = –2.13$, $p = .03$, $d = –0.16$ for mothers and $t(181) = –2.87$, $p = .00$, $d = –0.26$ for fathers, and more agreeable, $t(235) = 4.90$, $p = .00$, $d = 0.37$ for mothers and $t(181) = 2.44$, $p = .02$, $d = 0.23$ for fathers, than the children. The group of raters whose mean scores show more differences from all other raters are teachers. The differences are most pronounced on extraversion and openness, which are consistently rated higher by mothers and fathers than by teachers (with effect sizes ranging from 0.23 to 0.49). There is a good resemblance between mothers' and fathers' mean ratings in Wave 2, whereas in Wave 1, there are marginally significant differences in agreeableness, $t(351) = 2.09$, $p = .04$, $d = 0.09$, and extraversion, $t(347) = 2.59$, $p = .01$, $d = 0.10$.

Paired t tests for dependent samples indicate slight, yet statistically significant, differences among mean ratings provided by the same raters over the 3 years between Wave 1 and Wave 2. Generally, different judges did not agree with each other on whether and how the adolescents changed from 15 to 18 years of age. Mothers saw their children as becoming less extraverted, $t(298) = –2.03$, $p = .04$, $d = –0.12$, teachers saw them as becoming more agreeable, $t(225) = 2.61$, $p = .01$, $d = 0.23$, and fathers saw them as becoming more open, $t(193) = 2.62$, $p = .01$,

TABLE 1
Descriptive Statistics of Personality Scales

	Wave 1			*Wave 2*			
Scale	M	SD	α	M	SD	α	*% of Cases With RCI > |1.96|*
Self-ratings		$n = 276$			$n = 436$		
Neuroticism	22.84[ef]	5.33	.72	89.10	23.13	.91	
Extraversion	26.32[c]	5.33	.70	117.97	23.90	.91	
Openness	26.96[c]	4.70	.57	103.30	18.05	.84	
Agreeableness	27.03[cef]	4.26	.58	106.52	16.43	.82	
Conscientiousness	28.03	5.22	.71	110.24	19.39	.88	
Mother ratings		$n = 482$			$n = 404$		
Neuroticism	22.15[e]	5.16	.73	22.10[b]	5.65	.77	15.00
Extraversion	26.53[abg]	5.30	.68	26.62[ab]	5.35	.68	8.03
Openness	26.78[b]	4.31	.51	27.26[b]	4.33	.52	8.00
Agreeableness	28.69[eg]	4.80	.66	28.77	4.83	.64	11.67
Conscientiousness	28.19[b]	5.92	.78	28.99	5.95	.79	18.33
Father ratings		$n = 366$			$n = 288$		
Neuroticism	21.75[f]	4.90	.71	22.02[d]	5.07	.72	14.36
Extraversion	25.97[dg]	5.40	.71	26.75[d]	4.90	.64	9.05
Openness	26.57[ad]	4.35	.53	27.18[ad]	4.01	.45	5.67
Agreeableness	28.09[afg]	4.63	.63	28.50[a]	4.81	.66	12.81
Conscientiousness	28.24[a]	5.79	.79	29.40[a]	5.80	.79	16.50
Teacher ratings		$n = 489$			$n = 313$		
Neuroticism	21.98[c]	5.14	.76	20.88[bd]	5.73	.80	38.05
Extraversion	24.73[bcd]	6.95	.86	25.29[bd]	6.32	.83	30.40
Openness	24.97[bcd]	4.57	.66	25.32[bd]	4.13	.59	14.16
Agreeableness	28.21[ac]	5.81	.81	29.21[a]	5.96	.82	30.53
Conscientiousness	26.64[b]	6.61	.88	28.48	6.22	.86	39.73

NOTE: The NEO-PI-R was used for self-reports in 2001; all other ratings were obtained by the Estonian Brief Big Five Inventory. RCI = Reliable Change Index; n = number of valid cases.
a. $p < .05$, significant difference between the scores in Wave 1 and Wave 2.
b. $p < .05$, significant difference between teacher and mother ratings in the same wave.
c. $p < .05$, significant difference between teacher and self-ratings in the same wave.
d. $p < .05$, significant difference between teacher and father ratings in the same wave.
e. $p < .05$, significant difference between self- and mother ratings in the same wave.
f. $p < .05$, significant difference between self- and father ratings in the same wave.
g. $p < .05$, significant difference between mother and father ratings in the same wave.

$d = 0.20$, agreeable, $t(202) = 2.24$, $p = .03$, $d = 0.19$, and conscientious, $t(199) = 2.77$, $p = .01$, $d = 0.20$, over time. A Reliable Change Index (RCI) proposed by Jacobson and Truax (1991) was computed for all cases. For these analyses, we used Cronbach reliability coefficients from Wave 1 as the reliability estimates, because short-term test-retest reliability for the EBBFI has not been studied. The last column of Table 1 presents the percentage of cases that have shown reliable change. As the last column indicates, teachers have reported substantially more change for the adolescents in comparison with parents.

Agreement and Stability

Table 2 reports convergent interrater correlations for the five factors. In spite of the use of a different self-report measure in Wave 2, the pattern of correlations is very similar for the two waves. With regard to self-other correlations, parents' ratings agree slightly better with self-ratings than teacher ratings do. The median correlations for self-mother and self-father ratings over the five factors and two waves are .34 and .30, respectively, whereas the median for self-teacher ratings is .19. Further evidence that teachers rate students' personality in a different manner compared with parents is provided by the correlations between two observer raters. There is a high level of consensus between mothers and fathers about their children's personality with a median correlation of .65, but parents and teachers show only a limited agreement as indicated by a median of .19 for mother-teacher correlations and .17 for father-teacher correlations. Comparison of the five factors with each other reveals that extraversion is the trait on which the opinion of different judges converges the most (median $r = .39$), followed by conscientiousness (median $r = .31$), neuroticism (median $r = .22$), openness (median $r = .20$), and agreeableness (median $r = .19$). Teachers also showed the

highest agreement with all other informants on extraversion (median $r = .31$) and conscientiousness (median $r = .23$).

Separate correlational analyses were performed with only those cases for which complete data from all four sources were available ($n = 137$ in Wave 1, and $n = 168$ in Wave 2). The results were very similar to those presented in Table 2; none of the correlation coefficients differed significantly between the data sets. The largest difference was found in correlations between self- and mother-rated agreeableness in Wave 1 ($r = .34$, $n = 234$ vs. $r = .20$, $n = 137$), but it was not statistically significant ($p = .16$, two-sided test). Median correlations were as follows: .26 for self-mother agreement, .31 for self-father agreement, .23 for self-teacher agreement, .69 for mother-father agreement, .18 for mother-teacher agreement, and .16 for father-teacher agreement.

To test the possibility that teachers who have taught the students longer will provide ratings that agree more with parents' ratings and self-ratings, teachers were split into two groups along the median length of teacher-student contact (3 years), and separate correlations were computed for the two groups. As the bottom of Table 2 shows, there are more significant correlations (9 versus 3) and correlations are generally slightly higher for teachers who have taught their targets at least 3 years compared with teachers whose contact with students has lasted less than 3 years. However, the only statistically significant difference between the two groups is in teacher-self agreement on neuroticism ($p = .04$, one-sided test). Although longer contact appears to have a modest effect on increasing the agreement between teachers and other raters, especially between teachers and self-reports, agreement is low even for teachers who have known the students for 3 years or more.

Analysis of 3-year test-retest correlations highlights another finding that implies that teachers may not be very reliable raters. As can be seen in Table 3, self-ratings, mother ratings, and father ratings show moderate test-retest correlations on all personality scales, which is common at that age (see Roberts & DelVecchio, 2000). On the other hand, three of five test-retest correlations for teacher ratings are not statistically significant. With respect to self-ratings, these correlations are probably deflated, because two instruments that are not perfectly equivalent to each other were used in Wave 1 and Wave 2. Thus, estimated test-retest correlations could be higher, especially for openness and agreeableness scales that showed the lowest concordance between the two questionnaires.

It is possible that teachers' test-retest correlations were lower because homeroom teachers had changed for the majority of students during the 3 years and, therefore, ratings were not provided by the same person at both times. To test this hypothesis, test-retest correlations were computed separately for adolescents who were rated by the same teacher in both waves and for adolescents who were rated by different teachers. The two last rows of Table 3 show that test-retest correlations are not consistently higher for adolescents rated by the same teacher.

The previous results raise the question of whether all teachers are poor personality raters or if there are individual differences in their rating abilities. We examined this possibility by computing for each teacher who had judged at least three students the correlations between the ratings that the teacher had given to all his or her targets and their respective self- and parent ratings. These correlations indeed demonstrated substantial variability, ranging from –.75 to .85 (mean $r = .31$, $SD = .30$). There was a tendency for teachers who converged highly with students' self-ratings to show higher agreement with mothers and fathers as well. It is apparent that there is a considerable interindividual variability in how teachers rate their students' personalities, with only the best teacher raters reaching the level of consensus that exists between mothers and fathers.

DISCUSSION

Agreement on Adolescent Personality

Four sources of information were used to assess the personality traits of the adolescents participating in the European Youth Heart Study in Estonia: teachers, mothers, fathers, and the adolescents themselves. Analyses of interrater agreement revealed certain discrepancies in how the four types of informants perceived adolescents' personality characteristics. In concordance with past studies of interparental agreement on both normal personality (e.g., De Fruyt & Vollrath, 2003) and problem behaviors (e.g., Achenbach et al., 1987; Duhig et al., 2000), we found a high consensus between mothers and fathers, with a median correlation of 0.65. Correlation analyses as well as the comparison of mean ratings suggested that fathers and mothers are very similar informants when assessing the personality of their child.

Mothers and fathers were also similar in showing lower agreement (median $r = 0.32$) with adolescents' self-reports than with each other. A potential explanation for why adolescents' descriptions of their own personality do not exactly match the reports given by their parents could be lower validity of adolescents' self-reports, as has been proposed by Baker et al. (2004). However, this explanation does not seem very plausible, considering the growing body of research indicating that a typical 15- or 18-year-old adolescent has all the mental capabilities that are necessary to observe, analyze, and give sound reports of his or her personality traits (see De Fruyt, Mervielde, Hoekstra, & Rolland, 2000; Markey, Markey, Tinsley, & Ericsen,

TABLE 2
Convergent Interrater Correlations for Five Personality Factors

Raters	*Wave*	n	*N*	*E*	*O*	*A*	*C*
Self-other agreement							
Self-mother	1	234	**.35**[b]	**.33**[a]	**.25**	**.34**[b]	**.29**
	2	367	**.45**[b]	**.49**[a]	**.18**	**.28**	**.34**[b]
Self-father	1	181	**.26**[c]	**.40**	**.31**	**.19**	**.29**
	2	259	**.33**	**.42**	**.19**	**.27**	**.33**[c]
Self-teacher	1	230	.04[bc]	**.32**	**.24**	**.14**[b]	**.19**
	2	266	**.17**[b]	**.41**	**.19**	**.18**	**.15**[bc]
Other-other agreement							
Mother-father	1	348	**.65**[de]	**.72**[de]	**.66**[de]	**.65**[de]	**.72**[ade]
	2	270	**.63**[de]	**.67**[de]	**.58**[de]	**.59**[de]	**.60**[ade]
Mother-teacher	1	392	**.17**[d]	**.20**[d]	**.21**[d]	**.15**[d]	**.32**[d]
	2	274	**.12**[d]	**.29**[d]	**.19**[d]	**.17**[d]	**.19**[d]
Father-teacher	1	292	**.17**[e]	**.20**[ae]	**.12**[e]	.02[e]	**.37**[e]
	2	186	.12[e]	**.37**[ae]	.10[e]	**.17**[e]	**.26**[e]
Agreement with teachers who have taught their targets for less than 3 years							
Self-teacher	2	84	–.02[f]	**.40**	.19	.04	.05
Mother-teacher	2	87	–.00	**.27**	.10	.17	.16
Father-teacher	2	55	.09	**.34**	.07	.17	.15
Agreement with teachers who have taught their targets for 3 years or more							
Self-teacher	2	122	**.23**[f]	**.50**	**.22**	**.23**	**.28**
Mother-teacher	2	127	.16	**.35**	.16	.09	**.18**
Father-teacher	2	97	.18	**.40**	.11	.09	**.26**

NOTE: *n* = number of valid cases; N = neuroticism; E = extraversion; O = openness; A = agreeableness; C = conscientiousness. Statistically significant correlations ($p < .05$) are in boldface.
a. $p < .05$, significant difference between Wave 1 and Wave 2 in correlations between the same pair of raters.
b. $p < .05$, significant difference between self-mother and self-teacher correlations of the same wave.
c. $p < .05$, significant difference between self-father and self-teacher correlations of the same wave.
d. $p < .05$, significant difference between mother-father and mother-teacher correlations of the same wave.
e. $p < .05$, significant difference between mother-father and father-teacher correlations of the same wave.
f. $p < .05$, significant difference between two groups of teachers.

TABLE 3
Test-Retest Correlations Over 3 Years

Raters	n	*N*	*E*	*O*	*A*	*C*
Self-reports[a]	191	**.40**	**.43**	**.32**	**.19**	**.37**
Mothers	299	**.46**	**.51**	**.29**	**.44**	**.47**
Fathers	184	**.47**	**.51**	**.46**	**.28**	**.51**
Teachers	224	.07	**.32**	**.29**	.08	.12
Same teacher[b]	58	.19	**.32**	**.37**	.03	.15
Different teacher[c]	166	.05	**.34**	**.28**	.12	.14

NOTE: *n* = number of valid cases; N = neuroticism; E = extraversion; O = openness; A = agreeableness; C = conscientiousness. Statistically significant correlations ($p < .05$) are in boldface.
a. Different instruments were used for self-reports in two study waves.
b. Correlations between ratings provided by teachers who rated student's personality twice, in both waves.
c. Correlations between ratings provided by different teachers in two waves.

2002; McCrae et al., 2002). Another possibility is that parents and adolescents rely on different information when describing personality. Adolescents are already relatively independent and have their own lives separate from those of their family and parents. Because they are more concerned about the opinions of their classmates and friends than of their parents (Harris, 1998), they may base their judgments on experiences that are largely hidden from their parents. Greater agreement between raters who interact with the target in similar situations (e.g., between mothers and fathers) than between raters who interact with the target in different situations (e.g., between parents and children or between parents and teachers) has been a consistent finding in the research on children's and adolescents' behavior problems (Stanger & Lewis, 1993), so it is not surprising that this result extends to the assessment of normal personality as well.

Situational specificity probably accounts, at least in part, for the low agreement between teachers and all other sources of information. Agreement between two groups of informants, one of which was the group of teachers, was always lower than agreement between any other pair of judges. Nevertheless, even if students indeed

think, feel, and behave profoundly differently in school, this does not explain the instability of teachers' ratings over the 3-year span compared with self- and parent ratings. Test-retest correlations indicated that teachers were able to demonstrate sufficient reliability on only two of the five personality dimensions—extraversion and openness. Correlations on the other three dimensions failed to reach statistical significance, even when the same teachers made the ratings in both waves. Based on studies demonstrating that good personality judges are generally more intelligent, conscientious, and dependable people (e.g., Davis & Kraus, 1997; Realo et al., 2003), one might have expected that teachers, who have above average education and are likely to be dependable, would in fact be better informants about others' personalities than any randomly selected group of laypersons. However, this does not seem to be the case when teachers judge the personalities of their students.

There are two factors that might have influenced teachers' ratings. First, unlike parents and adolescents, teachers were asked to estimate many targets at the same time. Although estimation of several targets simultaneously probably provides a more stable frame of reference (Marsh & Craven, 1991), it is also possible that judgments of different persons start to interfere with one another, and instead of leading to greater distinctiveness, the simultaneous judgments become more similar to each other. Second, it is probable that teachers were overwhelmed with their everyday duties and were not motivated enough for this particular task. If this were the case, the lack of validity may have resulted from a high percentage of random responding on the questionnaires. Although this interpretation seems plausible, there is little evidence to support it, as both internal reliabilities of teacher ratings as well as the number of missed items were similar to those of other informants.

Funder and Colvin (1988) have discussed the association between interrater agreement and accuracy of personality judgment. According to them, interjudge agreement is a necessary but not sufficient condition for accuracy: Two judgments that agree may not be accurate, whereas two judgments that disagree cannot both be accurate. A different point of view about low interrater agreement is presented by Meyer (1996; Meyer et al., 2001), who argues that it is impossible to say that one source is more true than any other without having good criteria available (Meyer et al., 2001). Instead of questioning the accuracy of disagreeing ratings, Meyer emphasizes the value of unique information provided by different sources and distinct assessment methods. Cross-method disagreement, in Meyer's (1996) view, is "a phenomena [*sic*] that can lead to a more refined identification of people and more accurate behavioral predictions" (p. 575). By demonstrating that teachers' judgments of their students' personalities not only diverge from all other judges but also show less stability over time than ratings given by other informants, we have more reason to doubt the accuracy of teachers' personality ratings. However, the possibility that teachers' ratings might present a perspective that is superior to ratings given by other sources in predicting certain outcomes remains to be proved.

Adolescent Personality Development

The results of this study have implications for research on personality development in adolescence. Our previous study demonstrated that the mean scores of personality traits of Estonian adolescents were quite similar to the respective scores of Estonian adults, and cross-sectional differences in the mean scores from ages 12 to 18 were very modest (Allik, Laidra, Realo, & Pullmann, 2004; see also McCrae et al., 2002). This study, using a different set of participants, expanded these findings in two important ways. First, results suggested that the mean level of personality traits remained more or less on the same level from 15 to 18 years of age also when reported by mothers, fathers, and teachers. Thus, the stability of the mean level of personality traits does not appear to be an artifact caused, for example, by a self-serving or any other form of bias characteristic of self-report data. Second, previous cross-sectional data were supplemented by the longitudinal data of this study. It is well-known that the stability of cross-sectional data does not necessarily mean that the personality traits of any given individual are stable during the observed time-span but may also result from the fact that unsystematic changes in opposite directions cancel each other out. Results of this study demonstrated that in the eyes of their mothers and fathers, personality traits of the majority of adolescents changed very little over the 3-year period. Only approximately 10% to 15% of parents reported that the mean level of personality traits of their child had changed significantly. Extraversion and openness were perceived by parents as the most established personality characteristics, which significantly increased or decreased only in 6% to 9% of cases. Thus, this study provided additional evidence to support the conclusion that at both a cross-sectional and individual level of analysis, mean levels of personality traits change very little during adolescence (cf. Costa & McCrae, 2002).

Limitations of the Study

There are two main limitations to this study. First, there was a substantial attrition rate between the two waves; only 68% of the participants in the first study wave also

completed the second wave of the study. As such, the sample of Wave 2 may have been biased toward stronger orientation to educational achievements, compared with those adolescents who were not reassessed in the follow-up study. This bias was also indicated by higher ratings on conscientiousness—the trait that has consistently been found to predict academic achievement (e.g., Barbaranelli et al., 2003; Busato, Prins, Elshout, & Hamaker, 2000; Gray & Watson, 2002; John et al., 1994; Mervielde, Buyst, & De Fruyt, 1995; Musgrave-Marquart, Bromley, & Dalley, 1997; Wolfe & Johnson, 1995)—given by other informants to the adolescents who participated twice. As a consequence, it is possible that the parent-teacher agreement in Wave 2 and test-retest correlations are somewhat lower for conscientiousness due to the restricted variation in this trait.

Second, a different personality measure was used for self-reports in Wave 2 than in Wave 1, making it impossible to compare the means of self-ratings with ratings provided by other judges. Although the change of instruments certainly limits our findings, we believe that it does not invalidate them because (a) there is a moderate-to-high convergence between the two personality inventories, as was shown in an adult sample; (b) the pattern of correlations between self- and other ratings was obviously similar for the two waves, which would not have been possible if the two inventories had not been very similar in content; and (c) the means provided by observer raters who were administered the same inventory twice were still able to be compared. Previous research has reported better agreement across informants when parallel scales were used (Epkins, 1996; Epkins & Meyers, 1994). In our study, however, the agreement with self-reports in Wave 1, when all the sources completed the EBBFI, was not higher than in Wave 2, when the EBBFI was used for observer ratings and the NEO-PI-R for self-ratings.

CONCLUSION

In line with several previous investigations (Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003; Marsh & Craven, 1991; Miller & Davis, 1992; Stanger & Lewis, 1993; ter Laak et al., 2001), the findings of this study point to certain limitations in how teachers judge their students' personalities. Despite the reason cited to explain teachers' disagreement in their perceptions of their students, this is a serious matter because, beside teaching, educators continually judge students' achievement, motivation, and character. These judgments may, and probably do, affect students' lives.

REFERENCES

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213-232.

Allik, J., Laidra, K., Realo, A., & Pullmann, H. (2004). Personality development from 12 to 18 years of age: Changes in mean levels and structure of traits. *European Journal of Personality*, *18*, 445-462.

Ambady, N., Hallahan, M., & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology*, *69*, 518-529.

Baker, S. R., Victor, J. B., Chambers, A. L., & Halverson, C. F., Jr. (2004). Adolescent personality: A five-factor model construct validation. *Assessment*, *11*, 303-315.

Banai, E., Weller, A., & Mikulincer, M. (1998). Inter-judge agreement in evaluation of adult attachment style: The impact of acquaintanceship. *British Journal of Social Psychology*, *37*, 95-109.

Barbaranelli, C., Caprara, G. V., Rabasca, A., & Pastorelli, C. (2003). A questionnaire for measuring the Big Five in late childhood. *Personality and Individual Differences*, *34*, 645-664.

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology*, *21*, 177-187.

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology*, *65*, 546-553.

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, *86*, 599-614.

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Differences*, *29*, 1057-1068.

Colvin, C. R., & Funder, D. C. (1991). Predicting personality and behavior: A boundary on the acquaintanceship effect. *Journal of Personality and Social Psychology*, *60*, 884-894.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and the NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (2002). Looking backward: Changes in the mean levels of personality traits from 80 to 12. In D. Cervone & W. Mischel (Eds.), *Advances in personality science* (pp. 219-237). New York: Guilford.

Davis, M. H., & Kraus, L. A. (1997). Personality and empathic accuracy. In W. Ickes (Ed.), *Empathic accuracy* (pp. 144-168). New York: Guilford.

De Fruyt, F., Mervielde, I., Hoekstra, H. A., & Rolland, J-P. (2000). Assessing adolescents' personality with the NEO-PI-R. *Assessment*, *7*, 329–345.

De Fruyt, F., & Vollrath, M. (2003). Inter-parent agreement on higher and lower level traits in two countries: Effects of parent and child gender. *Personality and Individual Differences*, *35*, 189-301.

De Raad, B. (1996). Personality traits in learning and education. *European Journal of Personality*, *10*, 185-200.

Duhig, A. M., Renk, K., Epstein, M. K., & Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: A meta-analysis. *Clinical Psychology: Science & Practice*, *7*, 435-453.

Epkins, C. C. (1996). Parent ratings of children's depression, anxiety, and aggression: A cross-sample analysis of agreement and differences with child and teacher ratings. *Journal of Clinical Psychology*, *52*, 599-608.

Epkins, C. C., & Meyers, A. W. (1994). Assessment of childhood depression, anxiety, and aggression: Convergent and discriminant validity of self-, parent-, teacher-, and peer-report measures. *Journal of Personality Assessment*, *62*, 364-381.

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgements in predicting oral reading fluency. *School Psychology Quarterly*, *18*, 52-65.

Funder, D. C. (1999). *Personality judgement: A realistic approach to person perception*. San Diego: Academic Press.

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology*, *55*, 149-158.

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology*, *69*, 656-672.

Gray, E. K., & Watson, D. (2002). General and specific-traits of personality and their relation to sleep and academic performance. *Journal of Personality*, *70*, 177-206.

Harris, J. R. (1998). *The nurture assumption: Why children turn out the way they do*. New York: Free Press.

Harro, M., Eensoo, D., Kiive, E., Merenäkk, L., Alep, J., Oreland, L., & Harro, J. (2001). Platelet monoamine oxidase in healthy 9- and 15-years old children: The effect of gender, smoking and puberty. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *25*, 1497-1511.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12-19.

John, O. P., Caspi, A., Robins, R. W., Moffitt, T. E., & Stouthamer-Loeber, M. (1994). The "little five": Exploring the nomological network of the five-factor model of personality in adolescent boys. *Child Development*, *65*, 160-178.

Kallasmaa, T., Allik, J., Realo, A., & McCrae, R. R. (2000). The Estonian version of the NEO-PI-R: An examination of universal and culture-specific aspects of the five-factor model. *European Journal of Personality*, *14*, 265-278.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford.

Markey, P. M., Markey, C. N., Tinsley, B. J., & Ericsen, A. J. (2002). A preliminary validation of preadolescents' self-reports using the five-factor model of personality. *Journal of Research in Personality*, *36*, 173–181.

Marsh, H. W., & Craven, R. G. (1991). Self-other agreement on multiple dimensions of preadolescent self-concept: Inferences by teachers, mothers, and fathers. *Journal of Educational Psychology*, *83*, 393-404.

McCrae, R. R., Costa, P. T., Jr., Terracciano, A., Parker, W. D., Mills, C. J., De Fruyt, F., & Mervielde, I. (2002). Personality trait development from age 12 to age 18: Longitudinal, cross-sectional, and cross-cultural analyses. *Journal of Personality and Social Psychology*, *83*, 1456-1468.

Mervielde, I., Buyst, V., & De Fruyt, F. (1995). The validity of the Big Five as a model for teachers' ratings of individual differences among children aged 4-12 years. *Personality and Individual Differences*, *18*, 525-534.

Meyer, G. J. (1996). The Rorschach and MMPI: Toward a more scientifically differentiated understanding of cross-method assessment. *Journal of Personality Assessment*, *67*, 558-578.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128-165.

Miller, S. A., & Davis, T. L. (1992). Beliefs about children: A comparative study of mothers, teachers, peers and self. *Child Development*, *63*, 1251-1265.

Mischel, W. (1977). The interaction of person and situation. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333-352). Hillsdale, NJ: Lawrence Erlbaum.

Musgrave-Marquart, D., Bromley, S. P., & Dalley, M. B. (1997). Personality, academic attribution, and substance use as predictors of academic achievement in college students. *Journal of Social Behavior and Personality*, *12*, 501-511.

Paunonen, S. V. (1989). Consensus in personality judgments: Moderating effects of target-rater acquaintanceship and behavior observability. *Journal of Personality and Social Psychology*, *56*, 823-833.

Realo, A., Allik, J., Nõlvak, A., Valk, R., Ruus, T., Schmidt, M., & Eilola, T. (2003). Mind-reading ability: Beliefs and performance. *Journal of Research in Personality*, *37*, 420-445.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*, 3-25.

Stanger, C., & Lewis, M. (1993). Agreement among parents, teachers, and children on internalizing and externalizing behavior problems. *Journal of Clinical Child Psychology*, *22*, 107-115.

ter Laak, J.J.F., DeGoede, M.P.M., & Brugman, G. M. (2001). Teacher's judgements of pupils: Agreement and accuracy. *Social Behavior and Personality*, *29*, 257-270.

Watson, D. (1989). Stranger's ratings of the five robust personality factors: Evidence of a surprising convergence with self-report. *Journal of Personality and Social Psychology*, *57*, 120-128.

Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational & Psychological Measurement*, *55*, 177-185.

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology*, *68*, 1038-1050.

Kaia Laidra, MSc (psychology), is a graduate student in the Department of Psychology at the University of Tartu. Her research focuses on personality development.

Jüri Allik received his PhD from Moscow University in 1976 and also from Tampere University, Finland, in 1991. He is a professor of experimental psychology at the University of Tartu. He is a foreign member of the Finnish Academy of Science and Letters (1997). His primary field of research is visual psychophysics, especially perception of visual motion. His recent research, however, is more concentrated on personality, emotions, intelligence, and cross-cultural comparison. With Robert R. McCrae, he edited The Five-Factor Model of Personality Across Cultures (Kluwer Academic, 2002).

Maarike Harro, MD, PhD, is the director of the National Institute for Health Development and a visiting professor at the Department of Public Health, University of Tartu. Her research focuses on determinants of health-related behavior in children and adolescents and their long-term health consequences.

Liis Merenäkk, BSc (biology), MSc (public health), is a PhD student in neuroscience at the Faculty of Medicine at the University of Tartu, Estonia. Her main research interests are psychological and biological determinants of substance use.

Jaanus Harro, MD, PhD, is a professor of psychophysiology at the Department of Psychology, University of Tartu, and the director of the Estonian Centre of Behavioural and Health Sciences. His research is on neurobiological regulation of affect, which includes attempts to define the biological basis for the structure of personality.