Thin Slices of Behavior as Cues of Personality and Intelligence

Peter Borkenau and Nadine Mauer
Martin-Luther Universität

Rainer Riemann
Friedrich-Schiller Universität

Frank M. Spinath and Alois Angleitner
Universität Bielefeld

Self-reports, peer reports, intelligence tests, and ratings of personality and intelligence from 15 videotaped episodes were collected for 600 participants. The average cross-situational consistency of trait impressions across the 15 episodes was .43. Shared stereotypes related to gender and age were mostly accurate and contributed little to agreement among judges. Agreement was limited mainly by nonshared meaning systems and by nonoverlapping information. Personality inferences from thin slices of behavior were significantly associated with reports by knowledgeable informants. This association became stronger when more episodes were included, but gains in prediction were low beyond 6 episodes. Inferences of intelligence from thin slices of behavior strongly predicted intelligence test scores. A particularly strong single predictor was how persons read short sentences.

Consensus[1] and accuracy of personality judgments are important both for applied and theoretical reasons (Funder & West, 1993). They are important for applied reasons because personality judgments have effects on interpersonal attraction as well as on personal and organizational decisions. To the extent that personality judgments are accurate, they are likely to improve the quality of such decisions, whereas if judgments of personality were unrelated to actual individual differences, they would impair the quality of such decisions.

For theoretical reasons, consensus and accuracy are important because judgmental data are frequently used in personality research and because there are elaborate testable models on factors that influence consensus and accuracy of personality judgments. Most notable are Funder's (1995) realistic accuracy model and Kenny's (1991) weighted average model (WAM). We focus on WAM here because it allows for more specific predictions in the context of the present study.

Accuracy of Personality Judgments

There are two kinds of studies on consensus and accuracy of personality judgments. The first uses descriptions by close acquaintances or spouses of the target persons, because these persons know the target well and are therefore able to provide quite accurate judgments of his or her personality. This research (e.g., McCrae, 1982; McCrae & Costa, 1987) has typically yielded consensus correlations of about .50, but the sources of the achieved consensus (and lack of consensus) are not clear: The observers may have observed the targets in particular situations (like parties or the workplace) only, and to what extent consensus reflects similar inferences from the targets' observed behavior or communication between judges on the targets' personality cannot be controlled.

Second, there are studies that use personality ratings by strangers who are exposed to controlled samples of the targets' appearance (Harker & Keltner, 2001; Borkenau & Liebler, 1992b), behavior (Funder & Sneed, 1993; Kenny, Horner, Kashy, & Chu, 1992), or behavioral residues (Gosling, Ko, Mannarelli, & Morris, 2002; Rentfrow & Gosling, 2003). Although consensus tends to be lower in this kind of research than in studies that relate self-reports to peer reports, it has become clear that very short observations by

Peter Borkenau and Nadine Mauer, Department of Psychology, Martin-Luther Universität, Halle, Germany; Rainer Riemann, Department of Psychology, Friedrich-Schiller Universität, Jena, Germany; Frank M. Spinath and Alois Angleitner, Department of Psychology, Universität Bielefeld, Bielefeld, Germany.

The research reported in this article was supported by a grant from the Deutsche Forschungsgemeinschaft to Alois Angleitner, Peter Borkenau, and Rainer Riemann. We are indebted to the twins and the judges for their participation; to the experimenters Susanne Hempel, Veronika Koch, Holger Lorenz, Conny Post, Beatrice Rammstedt, Birgit Schlangen, and Robert Weiss for collecting the data; and to Holger Lorenz and Wolfgang Thiel for their help in the data analysis.

Correspondence concerning this article should be addressed to Peter Borkenau, Department of Psychology, Martin-Luther Universität, D-06099 Halle, Germany. E-mail: p.borkenau@psych.uni-halle.de

[1] In this article, we do not systematically distinguish between *consensus* and *agreement* but use both terms interchangeably.

Journal of Personality and Social Psychology, 2004, Vol. 86, No. 4, 599–614
 0022-3514/04/$12.00 DOI: 10.1037/0022-3514.86.4.599

strangers may be sufficient to obtain statistically significant self–stranger agreement for judgments of personality (Albright, Kenny & Malloy, 1988; Borkenau & Liebler, 1992b; Funder, 1999; Funder & Colvin, 1988; Norman & Goldberg, 1966). Moreover, such studies allow separation of several sources of consensus and accuracy in judgments of personality, such as acquaintance or extent of information, information overlap, similarity of meaning systems, cross-situational consistency of target behavior, agreement on stereotypes, and a kernel of truth in stereotypes (Kenny, 1991, 1994).

Acquaintance

Theoretically, WAM predicts that accuracy increases with acquaintance and that consensus increases with acquaintance if information overlap between observers is less than perfect. Moreover, it predicts diminishing returns of additional information, that is, that subsequent information contributes less to consensus and accuracy than previous information. Consensus and accuracy of personality judgments should therefore approach asymptotic values after presentation of a moderate number of target behaviors, whereby the extent of shared meaning systems sets an upper limit to consensus (Kenny, 1994).

Empirically, however, the evidence on the acquaintance–accuracy relation is mixed: Ambady and Rosenthal (1992) published a meta-analysis on the accuracy of predictions of various objective outcomes in the areas of clinical and social psychology from under-5-min observations of expressive behavior. They found that predictions based on observations under 0.5 min in length did not differ significantly from predictions based on 5-min observations. Ambady, Bernieri, and Richeson (2000) published similar findings, and they explained the lack of relation between acquaintance and accuracy by (a) the importance of stylistic variables that can be observed even in extremely thin slices of behavior and (b) the fact that impressions of personality are formed immediately and are hardly modified when additional information becomes available. Similarly, Colvin and Funder (1991) and Funder and Sneed (1993) found that judgments by acquaintances and by strangers did equally well at predicting behavior, but this may reflect that in these studies, the strangers observed exactly those behavioral sequences that were used as criteria for accuracy. Thus, lower acquaintance was confounded with higher overlap, and this may have masked the effects of acquaintance on accuracy.

Other studies, however, have shown that self–other agreement increases with acquaintance (Blackman & Funder; 1998; Funder & Colvin, 1988; Norman & Goldberg, 1966; Paulhus & Bruce, 1992; Paunonen, 1989) and that self–stranger agreement is stronger if strangers view a sound film than if they view a silent film or a still picture of targets (Borkenau & Liebler, 1992b). Because WAM predicts a decreasing slope of the curve that describes consensus and accuracy as a function of acquaintance, however, discrepant findings may reflect different stages in the acquaintance process. This makes it desirable to study the effects of available information on consensus and accuracy more systematically than has been done previously. We expect that in a systematic study on the effects of acquaintance on consensus and accuracy, the curvilinear pattern predicted by WAM will be obtained.

Overlap

Another factor that is likely to contribute to interjudge agreement is *overlap,* the extent that judges are exposed to overlapping information on the targets' behavior (Kenny, 1991). Support for this assumption stems from at least two sources. First, Kenny et al. (1992) followed the consensus among observers across the acquaintance process and found that the level of consensus declined when initially unacquainted targets and judges interacted in dyads but that consensus increased when targets and judges interacted in a group. Because observers are exposed to different behaviors of the targets when they interact in dyads but to the same behaviors when they interact in a group, this finding suggests that information overlap contributes to consensus. Second, Borkenau and Liebler (1992a) videotaped target persons and derived sound films, silent films, still pictures, and audiotapes from these videotapes. This information was then presented to independent groups of judges, and the consensus between the judges in different conditions was studied. If judgments were based on overlapping information (e.g., silent film and still picture), consensus was higher than if judgments were based on nonoverlapping information (silent film and audiotape; still picture and audiotape).

We therefore predict higher interjudge agreement if judgments rely on overlapping behavioral episodes than if they rely on nonoverlapping behavioral episodes. Moreover, we predict that personality descriptions by strangers are as accurate as, or even more accurate than, self-reports or descriptions by close acquaintances if the strangers' personality impressions rely on behavioral episodes that constitute the criteria of accuracy.

Shared Meaning Systems

Shared meaning systems refer to the similarity of personality impressions that different judges infer from the same observed behavior. A proxy for the extent of shared meaning systems is the consensus between judges who are exposed to the same behavioral episodes of the targets. That this is only a proxy reflects that not only shared meaning systems but also shared stereotypes (e.g., "young people are more intelligent than old people") may contribute to consensus (Kenny, 1994). One might therefore argue that estimates of consensus should be adjusted for the potential effects of, for example, gender and age on personality impressions to obtain more precise estimates of the extent of shared meaning systems. This, however, is not without problems because there are actual gender and age differences in personality (Srivastava, John, Gosling, & Potter, 2003).

Consistency

According to WAM, consistency is the extent to which the same targets convey similar impressions of their personality via different acts. We study the cross-situational consistency of personality impressions systematically here because estimates of cross-situational consistency in personality vary widely: Mischel and Peake (1982) reported an average correlation of .13 among 19 reliable behavioral indicators of conscientiousness in their Carleton Behavior Study, whereas Hartshorne and May (1928) reported an average correlation of .23 among various observational tests of honesty in their Character Education Inquiry. Higher cross-

situational consistency coefficients of about .28 were reported by Funder and Colvin (1991) from the Riverside Accuracy Project if the targets' behavior in three settings was described on 62 behavioral Q-sort items. Moreover, when the 62 behavioral Q-sort items were combined into four common factors, the correlations among behavior descriptions by independent observers of different behavioral episodes rose to an average of .46.

Several explanations of these discrepant findings are reasonable. First, the difference between the studies by Mischel and Peake (1982) and by Hartshorne and May (1928) may reflect differences between behavioral domains; the narrow category of honest behaviors may be more consistent than the broader category of conscientious behaviors. Second, the higher estimates of cross-situational consistency reported by Funder and Colvin (1991) may reflect that their behavioral Q-sort largely referred to behavioral style variables (e.g., "speaks in a loud voice") instead of operant behaviors like ways of cheating (Hartshorne & May, 1928) or class attendance (Mischel & Peake, 1982). Indeed, speaking in a loud voice was the most cross-situationally consistent behavior in Funder and Colvin's study. This suggests the highly interesting hypothesis that persons may be more consistent in how they are performing tasks (behavioral style) than in their operant behaviors. Funder and Colvin's study provides no clear evidence in this respect, however, because their relatively high consistency estimates may reflect that their participants were observed in quite similar settings, two of them being "get-acquainted" settings. The fact that in this study the mean coefficient of cross-situational consistency across the two get-acquainted settings (.58 at the level of the four common factors) was higher than across either of these two settings and the third "having a debate" setting (.45 and .37 at the level of the four common factors) speaks to the importance of situational similarity. Consequently, it is desirable to investigate the cross-situational consistency of behavioral style in larger and more heterogeneous samples of settings.

Moreover, even the relatively high cross-act/cross-judge correlations reported by Funder and Colvin (1991) underestimate the cross-situational consistency of personality impressions. This is because two factors attenuate cross-act/cross-judge agreement: (a) lack of cross-situational consistency of the targets' behavior and (b) nonshared meaning systems and nonshared stereotypes. To obtain unbiased estimates of the cross-situational consistency of personality impressions, cross-act/cross-judge correlations should be corrected for lack of consensus (that is, for less-than-perfect same-act/cross-judge correlations) using the conventional correction-for-attenuation formula. Obviously, more than one judge has to observe the same behavioral episode of each target to obtain estimates of consensus, and different panels of judges have to observe different behavioral episodes of the targets to obtain independent estimates of the personality impressions that the targets convey in different settings. The present study was designed accordingly.

Stereotypes

Consensus and accuracy of personality impressions may reflect not only behavioral information but also shared stereotypes, or a kernel of truth in stereotypes, respectively (Kenny, 1994). There is much evidence showing that variables like gender and age influence first impressions of personality, although they seem to decrease in importance when more behavioral information becomes available (Krueger & Rothbart, 1988). Insofar as such stereotypes are shared by judges, they may contribute to interjudge agreement. For example, if there is a shared stereotype that women are less aggressive than men, and if a mixed-sex sample of targets is studied, that stereotype may contribute to agreement on individual differences in aggressiveness, particularly if the judges are strangers.

The fact that shared stereotypes influence personality impressions does not imply, however, that they inflate same-act/cross-judge correlations under any circumstances. This is because inferences from behavioral observations and from stereotypes contribute to the covariance between personality impressions by different judges as well as to the variance of personality impressions within judges. Because the consensus correlation is the ratio of the covariance divided by the geometric mean of the variances, shared stereotypes inflate same-act/cross-judge correlations only if they contribute more to the covariance between judges than to the variance of the ratings of individual judges (Kenny, 1994). Thus, whether control for group membership actually reduces the consensus between observers of the same target behaviors is an empirical issue.

Because the targets' group membership remains constant across behavioral episodes, shared stereotypes also contribute to cross-act/cross-judge correlations. Moreover, it is likely that control of group membership diminishes cross-act/cross-judge correlations more than same-act/cross-judge correlations. This is because covariance between judges that reflects shared stereotypes should be unaffected by whether the same or different acts of the targets are observed, whereas covariance that reflects inferences from behavior should be larger if the same rather than different acts are observed. Thus, we predict that control of group membership attenuates cross-act/cross-judge correlations more than same-act/cross-judge correlations. Consequently, lower estimates of cross-situational consistency of personality impressions are expected if group membership is controlled.

The kernel-of-truth hypothesis becomes important if one studies accuracy of personality impressions instead of consensus. Let us assume that young people are consensually perceived as more intelligent than old people, and that younger people indeed perform better on performance measures of intelligence. In that case, there would be a kernel of truth in the stereotype that young people are more intelligent, and this stereotype may contribute to the accuracy of perceptions of intelligence in age-heterogeneous samples. We predict that there is a kernel of truth in stereotypes concerning relations of gender and age to intelligence. This implies that correlations of these variables with judgments of intelligence resemble their correlations with performance measures of intelligence. However, control for group membership will reduce the accuracy of perceptions of intelligence only if group membership contributes more to the covariance between judgments and performance than to the variances within judgments and performance. This is an empirical issue that is addressed in the present study.

Predictions

To summarize, we expected that various predictions of WAM would be confirmed: (a) Interjudge agreement and accuracy increases with the extent of available information and shows the

curvilinear pattern predicted by WAM; (b) consensus is higher if there is behavioral overlap than if there is no behavioral overlap; and (c) consistent with the kernel-of-truth hypothesis, correlations of gender and age with judgments of personality and intelligence by observers resemble their correlations with self-reports of personality and with performance measures of intelligence. Another prediction that does not follow from WAM is that the cross-situational consistency of personality impressions is higher than the cross-situational consistency of behavior.

The Present Study

The analyses reported in this article are based on data that were collected in the German Observational Study of Adult Twins (GOSAT), a multimethod behavior–genetic study with 600 participants. The available data include (a) self-reports; (b) independent descriptions by two acquaintances per participant; (c) ratings of the participants' personality by four independent judges, based on observations of 1 out of 15 videotaped behavior sequences; (d) descriptions of the participants' personality by experimenters who guided them through the study; (e) descriptions of the participants' personality by experimental confederates who interacted with them in 6 of the 15 videotaped behavior sequences; and (f) performance of the participants on two psychometric tests of intelligence. Inclusion of performance tests allowed the study of accuracy in addition to agreement. A comprehensive description of the data that were collected in GOSAT is available in Spinath et al. (1999), and the major behavior–genetic analyses from GOSAT were reported by Neubauer, Spinath, Riemann, Borkenau, and Angleitner (2000) and by Borkenau, Riemann, Angleitner, and Spinath (2001).

In this article, we do not compare the similarity of monozygotic and dizygotic twins but instead analyze the data at the level of individual participants. Because the participants were twins, however, and the observations were therefore not independent, the sample of 600 participants was split into two subsamples consisting of 300 participants each, that is, one co-twin from each pair. Independent analyses were run for the subsamples, and the findings from these analyses were combined.

Method

Participants

Three hundred pairs of same-sex adult twins (168 monozygotic and 132 dizygotic) who had been recruited from all over Germany by reports in German media participated in GOSAT. They were invited for a 1-day testing session that took place at the University of Bielefeld. The participants were reimbursed for their travel expenses and received a flat rate of the equivalent of approximately $15 US per person for catering. Women (234 pairs) participated more frequently than men (66 pairs). The participants' ages varied between 18 and 70 years, with a mean of 34.28 (SD = 12.99) and a median of 30.50 years. Gender and age were independent (r = .06, p = .16). The ethnicity of all participants was Caucasian.

Measures

Self-reports and acquaintance reports. Most of the GOSAT twin pairs had previously participated in a peer rating study (Riemann, Angleitner, & Strelau, 1997) where they had been administered, among others, the German version of Costa and McCrae's (1992) NEO-Five Factor Inventory (NEO-FFI; German version by Borkenau & Ostendorf, 1993). The NEO-FFI measures the personality domains Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness with 12 items each. Moreover, each participant had independently been described by two acquaintances that differed between co-twins, using Form R of the NEO-FFI in which the items are worded in the third person instead of the first-person singular. A few twin pairs who participated in GOSAT had not participated in the peer rating study, but for most of these participants, self-reports and reports by acquaintances were collected.

Measures of intelligence. Horn's (1962) Leistungspruefsystem [Performance Test System] (LPS) and Raven's (1958) Advanced Progressive Matrices (APM) were administered. The LPS is a frequently used multifactorial German intelligence test. The complete form consists of 14 subtests with a total number of 560 items. Because of time limitations, we used the 7-subtest short form suggested by Sturm and Willmes (1983), which allows the separate assessment of verbal and nonverbal aspects of intelligence as well as the calculation of a general intelligence score. Moreover, we administered Set II of Raven's APM and allowed the participants 20 min to solve the 36 items.

Videotaped behavior sequences. To obtain reliable and valid observational measures of the participants' personalities, they were individually videotaped in 15 settings in which they had to complete assigned tasks. These videotapes were later presented to independent judges who never met the participants in person.

Because it was desirable to collect personality descriptions by strangers that would be informative of the targets' actual personality, we sought tasks to be completed by the participants in which personality differences were likely to be observable. Assessment centers face similar problems because their task is to predict the aptitude of testees for positions in organizations from short-term behavior observations in structured settings. Thus, we made use of the literature on assessment centers when the tasks were designed. Specifically, our tasks included the following (average duration in parentheses):

1. Introducing oneself (1.25 min).
2. Arranging three photographs in a meaningful order and telling an interesting story that the three pictures might illustrate (4.50 min).
3. Telling dramatic stories based on three cards from Murray's (1943) Thematic Apperception Test (6 min).
4. Telling a joke to an experimental confederate (1.50 min).
5. Persuading an "obstinate neighbor" (actually a confederate) on the phone to reduce the volume of her stereo after 11 p.m. (2.25 min).
6. Refusing a request for help by "a friend" (actually a confederate) who says that she has just had a minor car accident (2 min).
7. Introducing oneself to a stranger (an experimental confederate) and telling her about one's hobbies (after the confederate has introduced herself) (12 min).
8. Recalling objects just seen in a waiting room (3 min).
9. Solving a complex logical problem as fast as possible. Another "participant" (actually a confederate) received the same problem and ostensibly "solved" it at an enormous speed (4.50 min).
10. Introducing the stranger from Setting 4 to the experimenter (2.50 min).

Table 1
Correlations Between Self-Reports, Reports by a Single Acquaintance, Experimenter Reports, and Reports by Experimental Confederates

	Self ×			Acquaintance ×			
Personality domain	Acquaintance	Experimenter	Confederate	Acquaintance	Experimenter	Confederate	Experimenter × Confederate
Neuroticism	.45 (.43)	.20 (.18)	.15 (.10)	.42 (.41)	.27 (.26)	.21 (.17)	.32 (.31)
Extraversion	.54 (.53)	.38 (.38)	.31 (.29)	.47 (.47)	.29 (.32)	.27 (.27)	.45 (.46)
Openness to Experience	.46 (.46)	.39 (.38)	.31 (.31)	.42 (.41)	.30 (.30)	.31 (.31)	.42 (.43)
Agreeableness	.41 (.41)	.17 (.16)	.17 (.17)	.33 (.33)	.18 (.18)	.19 (.19)	.25 (.24)
Conscientiousness	.39 (.37)	.20 (.18)	.14 (.12)	.38 (.37)	.21 (.20)	.13 (.11)	.37 (.34)
Column mean	.45 (.44)	.27 (.26)	.22 (.20)	.41 (.40)	.25 (.25)	.22 (.21)	.36 (.36)

Note. Coefficients in parentheses are second-order correlations that control for gender and age.

11. Inventing a "definition" for a neologism and providing arguments for why that definition would be appropriate (6.25 min).

12. Rigging up a high and stable paper tower within 5 min, using only scissors, paper, and glue (5.25 min).

13. Reading 14 newspaper headlines and their subtitles aloud (3 min).

14. Describing multiple uses of a brick, using pantomime only (2.75 min).

15. Singing a song of one's choice (1 min).

Approximately 60 min of video material per participant, that is, a total of 600 hr of footage for all participants, was collected in this way.

Video-based personality ratings. Altogether, 120 judges provided trait ratings of the participants that were based on 1 of the 15 videotaped episodes only. Moreover, each participant was observed in each setting by four independent judges. The behavior in different settings was rated by different judges to secure independence of ratings for different settings. Thus, behavioral overlap between observers of the same setting was perfect, whereas there was no behavioral overlap between observers of different settings. Finally, different panels of judges were used for co-twins. This resulted in 4 (parallel judgments) × 15 (number of settings) × 2 (co-twins) = 120 judges used. All judges were students either of the University of Bielefeld or the University of Halle and were paid for their participation.

The judges provided their ratings via a computer on 35 bipolar 5-point rating scales. Each of the domains of the five-factor model of personality was represented by four adjective scales, and four additional scales were included to measure intellect (Goldberg, 1990), among them a 5-point rating scale labeled *Unintelligent–Intelligent.* The selection of these 24 scales relied on a large trait-taxonomic study of the German personality-descriptive language by Angleitner, Ostendorf, and John (1990). Two of the 4 scales that measured the same factor were reverse scored to control for acquiescence response set.

Two additional scales asked for ratings of the targets' attractiveness and likability, and the remaining nine scales differed between the 15 settings and were included to account for setting-specific behavior. In this article, however, we focus on analyses for the markers of Neuroticism, Extraversion, Openness to Experience, Agreeableness, Conscientiousness, and on the ratings of intelligence.

The judges' workstations were equipped with a video cassette recorder, a video monitor, and a personal computer. The instructions were to (a) watch a video sequence showing one target person, (b) provide the 35 trait ratings for that target using the computer keyboard, and (c) restart the video recorder to watch another target until they had provided ratings of 300 targets. The software had been programmed to present the 35 adjective scales in a random order that differed between video sequences and to store the judges' responses. Before the judges provided their ratings of the targets, their understanding of the instruction was checked in practice trials, using an extra videotape displaying target persons not included in the actual set for the empirical study. Altogether, 1.26 million video-based ratings were collected in this way, taking more than 4,100 hr for observation and ratings.

Ratings by experimenters and confederates. The participants were also described by the experimenter and a confederate, using the observer-rating form of the NEO-FFI. The (always female) confederate was involved in six observational settings (Setting 4 to Setting 9), and she provided her descriptions when she had interacted with the target for about 1 hr. The experimenter described the target at the end of the observation day after about 6 hr of interaction and observation. In the period between meeting and describing the targets, experimenters and confederates had no opportunity for communication (Spinath, 1999).

Results

All analyses were run separately for the two subsamples (that included 1 co-twin from each pair), and the correlations were averaged, using Fisher's *Z* transformation for correlations. Moreover, the correlations with the descriptions by the targets' two acquaintances were averaged. Thus, the coefficients of rater agreement were averages of two or four correlations, each based on a sample of approximately 300 participants.[2] Hence, even small differences were reliable.

Agreement Between Knowledgeable Informants

Agreement between self-reports, reports by acquaintances, experimenter ratings, and confederate ratings on the NEO-FFI scales is reported in Table 1. It differed between types of judges: Self × Acquaintance, Acquaintance × Acquaintance, and Experimenter × Confederate agreement was relatively high, whereas Self × Experimenter, Self × Confederate, Acquaintance × Experimenter, and Acquaintance × Confederate agreement was relatively low. This pattern cannot be explained by differences between judges in the extent of available information because the

[2] Because of missing data, particularly for self-reports and reports by acquaintances, the sample size for the different analyses varied from 279 to 300 per subsample.

Table 2
Correlations of Descriptions of Personality by Targets, Acquaintances, Experimenters, and Experimental Confederates With Targets' Gender and Age

	Self-reports		Acquaintance reports		Experimenter reports		Confederate reports	
Personality domain	Gender	Age	Gender	Age	Gender	Age	Gender	Age
Neuroticism	.16	−.14	.19	−.03	.14	.02	.22	−.03
Extraversion	.05	−.08	.07	−.03	.00	−.05	−.04	−.09
Openness to Experience	−.08	−.08	−.05	−.05	−.06	−.15	−.04	.05
Agreeableness	.12	.02	.07	.03	.08	−.05	.09	−.05
Conscientiousness	−.03	.18	−.01	.17	−.11	.08	−.13	.10

Note. Positive correlations with gender indicate higher scores for women than for men.

targets and their acquaintances knew more about the targets' personality than the experimenters and confederates, but the latter agreed more with each other than with the self- and peer descriptions. Rather, agreement was relatively high if both judges relied on observations outside the study or if both judges relied on observations inside the study, whereas agreement was lower if one judge relied on observations outside and the other relied on observations inside the study. This shows that behavioral overlap is important for consensus.

Moreover, agreement differed systematically among the five trait domains. For most pairs of judges, it was highest for Extraversion and for Openness to Experience, lowest for Agreeableness and Conscientiousness, and intermediate for Neuroticism.

Effects of Gender and Age

The correlations of gender and age with self-descriptions and descriptions by acquaintances, experimenters, and experimental confederates on the NEO-FFI are reported in Table 2. Correlations of personality descriptions with the targets' gender and age were low and somewhat more consistent for gender than for age. When looking at the correlations with self-reports and reports by acquaintances only, Neuroticism, Extraversion, and Openness to Experience decreased with age whereas Agreeableness and Conscientiousness increased. However, if descriptions by experimenters and confederates were also considered, the only consistent trend was that Conscientiousness was higher in older participants. These correlations may reflect effects of age or cohort. Concerning gender, mean levels of Neuroticism and of Agreeableness were consistently higher for women than for men, mean levels of Openness to Experience and of Conscientiousness were consistently higher for men than for women, and the findings for Extraversion were inconsistent.

To control for agreement among judges that reflected shared stereotypes, second-order correlations were calculated that controlled for the targets' gender and age. They are reported in Table 1 in parentheses. Many of them were marginally lower but some were even higher than the zero-order correlations. This suggests that shared stereotypes associated with gender and age contributed negligibly to the agreement among judges.

Cross-Situational Consistency of Trait Impressions

According to WAM (Kenny, 1994), the same-act/cross-judge correlation is attenuated by nonshared meaning systems and by nonshared stereotypes (and error of measurement) only, whereas the cross-act/cross-judge correlation is also attenuated by lack of cross-situational consistency of the targets' behavior. We therefore estimated the cross-situational consistency of trait impressions by adjusting the cross-act/cross-judge correlations for lack of agreement between observers of the same behavioral episode. For the 20 marker variables (4 per domain) of Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness, and for the adjective scale Unintelligent–Intelligent, Table 3 reports (a) the average agreement between individual judges who observed the same target behavior, (b) the average reliability of the composite score of four judges who were exposed to the same behavioral episode (i.e., how strongly their mean rating is expected to correlate with the mean rating by four other judges who are exposed to the same behavioral information),[3] (c) the average cross-task/cross-judge correlation between the composite scores of two panels of four judges, and (d) the estimate of cross-situational consistency (i.e., the average cross-task/cross-judge correlation [column 3] divided by the reliabilities of the composite scores [column 2]).

For the single-adjective scales, agreement between two individual judges of the same behavioral episode ranged from .19 to .42 with a mean of .28, the reliability of the same-task ratings ranged from .48 to .74 with a mean of .60, and the cross-situational consistency estimates of the personality impressions ranged from .29 to .61 with a mean of .43. To check whether estimates of cross-situational consistency increased if the analyses were run at the level of personality domains instead of single-adjective scales, the adjectives were combined into four-item measures of Neuroticism,[4] Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. The findings for these domain scores are also reported in Table 3: Consensus was higher at the domain level than at the level of single adjectives, but estimates of cross-situational

[3] The coefficients in column 2 are connected to those in column 1 via the Spearman-Brown formula: $R_4 = 4 * R_1/(1 + 3 * R_1)$. There are minor discrepancies, however, because the reported coefficients are means of several coefficients.

[4] Because one of the intended marker variables (Unemotional–Emotional) of Neuroticism did not correlate with the other three markers, this adjective scale was not included in the Neuroticism domain score. Therefore, the Neuroticism domain score is the composite of three adjective scales.

Table 3
Agreement Among Judges of Videotaped Behavior Sequences Who Observed Targets Performing the Same or a Different Task

Adjective scale	Same-task/cross-judge (1 judge)	Same-task/cross-judge (4 judges)	Cross-task/cross-judge (2 × 4 judges)	Cross-situational consistency
Neuroticism				
Assertive–Unassertive	.30 (.30)	.63 (.63)	.29 (.29)	.46 (.46)
Composed–Nervous	.22 (.21)	.52 (.51)	.24 (.22)	.46 (.43)
Self-Confident–Helpless	.31 (.31)	.64 (.64)	.29 (.28)	.45 (.44)
Unemotional–Emotional	.22 (.21)	.53 (.51)	.24 (.20)	.45 (.39)
M	.34 (.34)	.67 (.67)	.31 (.30)	.46 (.45)
Extraversion				
Restrained–Candid	.34 (.35)	.66 (.67)	.34 (.34)	.52 (.51)
Passive–Active	.32 (.33)	.65 (.66)	.29 (.30)	.45 (.45)
Silent–Talkative	.42 (.43)	.74 (.75)	.35 (.34)	.47 (.45)
Aloof–Gregarious	.38 (.40)	.71 (.72)	.38 (.38)	.54 (.53)
M	.49 (.50)	.79 (.80)	.40 (.40)	.51 (.50)
Openness to Experience				
Unwitty–Witty	.33 (.35)	.66 (.67)	.19 (.19)	.29 (.28)
Unimaginative–Imaginative	.31 (.32)	.64 (.65)	.20 (.20)	.31 (.31)
Conventional–Original	.30 (.30)	.63 (.62)	.28 (.26)	.44 (.42)
Uncreative–Creative	.30 (.31)	.63 (.64)	.20 (.20)	.32 (.31)
M	.40 (.41)	.72 (.73)	.25 (.24)	.35 (.33)
Agreeableness				
Unfriendly–Friendly	.27 (.27)	.60 (.59)	.28 (.27)	.47 (.46)
Rude–Polite	.22 (.23)	.53 (.54)	.26 (.24)	.49 (.44)
Disagreeable–Agreeable	.20 (.21)	.50 (.50)	.24 (.21)	.48 (.42)
Unkind–Kind	.19 (.20)	.48 (.48)	.17 (.15)	.35 (.31)
M	.31 (.33)	.64 (.66)	.29 (.27)	.45 (.40)
Conscientiousness				
Negligent–Careful	.23 (.24)	.55 (.55)	.19 (.18)	.35 (.33)
Disorderly–Orderly	.21 (.21)	.51 (.51)	.19 (.18)	.37 (.35)
Careless–Conscientious	.23 (.24)	.55 (.55)	.21 (.21)	.38 (.38)
Unsystematic–Systematic	.21 (.22)	.52 (.51)	.16 (.16)	.31 (.31)
M	.31 (.32)	.64 (.64)	.23 (.22)	.36 (.34)
Unintelligent–Intelligent	.30 (.29)	.62 (.61)	.38 (.35)	.61 (.57)
Mean of single adjectives	.28 (.28)	.60 (.60)	.26 (.25)	.43 (.41)
Mean of domain composites	.37 (.38)	.69 (.70)	.30 (.29)	.43 (.40)

Note. Correlations in parentheses are adjusted for the effects of gender and age.

consistency were unaffected by aggregation across items.[5] Moreover, Table 3 shows that same-task/cross-judge correlations were highest for Extraversion and Openness to Experience, whereas estimates of cross-situational consistency were highest for Extraversion but lowest for Openness to Experience.

Effects of Gender and Age

The extent to which the video-based ratings were related to the targets' gender and age was checked by computing correlations of gender and age with (a) task-specific domain scores, aggregated across the ratings of those 4 judges who had observed the same behavioral episode, and (b) composite domain scores, aggregated across the ratings of all 60 judges who had observed a target (whereby 4 independent judges were nested within each of the 15 tasks). The average reliability (Cronbach's alpha) of the task-specific domain scores was .69, and the average reliability of the composite domain scores was .94. The correlations of the composite and of the task-specific domain scores with the targets' gender and age are reported in Table 4.

By and large, the correlations were low. They were higher for the composite domain scores than for the task-specific domain scores, and correlations with age were higher than correlations with gender. For gender, the correlations with the video-based ratings resembled the correlations with self-reports and reports by acquaintances: Ratings of Neuroticism and of Agreeableness were higher for women than for men, and ratings of Conscientiousness were higher for men than for women. This suggests that the correlations of the video-based ratings with gender reflected actual gender differences in personality. Concerning age, the correlations with the video-based ratings were stronger than those with descriptions by knowledgeable informants. Moreover, video-based ratings of Agreeableness were negatively related to age, whereas Agreeableness as reported by knowledgeable informants was unrelated to age.

[5] The estimates of consensus are consistency coefficients that adjust for systematic differences between judges in means and standard deviations. If the variance in the judgments is decomposed into proportions because of targets, judges, and Target × Judge interactions, the average percentages for the domain scores are .31, .10, and .59 for targets, judges, and Target × Judge interactions, respectively. The target main effect is the intraclass correlation (not adjusted for systematic differences between judges) between single judges of the same targets.

Table 4
Correlations of Video-Based Ratings of Personality With the Targets' Gender and Age

Personality domain	Gender		Age	
	Composite domain score[a]	Task-specific domain score[b]	Composite domain score[a]	Task-specific domain score[b]
Neuroticism	.19	.11	−.21	−.13
Extraversion	.03	.02	.09	.06
Openness to Experience	.02	.01	−.23	−.13
Agreeableness	.06	.05	−.26	−.15
Conscientiousness	−.17	−.09	.20	.12
Intelligence	−.21	−.15	−.27	−.16

Note. Positive correlations with gender indicate higher scores for women than for men.
[a] Correlation with the mean of the ratings by 60 judges. [b] Correlation with the mean of the ratings by 4 judges.

Whether gender and age inflated estimates of consensus and cross-situational consistency was checked by calculating second-order correlations that controlled for gender and age. These second-order correlations are reported in Table 3 in parentheses. The same-act/cross-judge correlations did not decrease systematically when gender and age were controlled, whereas the mean estimate of cross-situational consistency dropped from .43 to .41 for the single-adjective scales and from .43 to .40 for the domain scores. Thus, shared stereotypes for gender and age did not contribute to consensus between observers of the same target behaviors, but they contributed (although minimally) to cross-situationally consistent impressions of personality.

Correlations of Video-Based Ratings With Descriptions by Knowledgeable Informants

How strongly were ratings from thin slices of behavior related to self-reports and reports by acquaintances? Moreover, how strongly were they related to descriptions by experimenters and confederates who knew less about the targets' personality than the targets and their acquaintances? Table 5 reports, for the domains of the five-factor model of personality, the correlations of the video-based personality ratings with self-reports, acquaintance reports, experimenter reports, and confederate reports. These coefficients of agreement are reported separately as (a) average correlations with task-specific domain scores and (b) correlations with the composite domain score. The column means are reported in the two bottom rows.

Several findings are noteworthy. First, all correlations were positive. Second, although experimenters and confederates were less acquainted with the target persons, their judgments were stronger than the judgments by targets and acquaintances related to the video-based ratings. This points to the importance of behavioral overlap. Third, interjudge agreement increased with the extent of information; personality descriptions on the NEO-FFI were consistently related more strongly to the composite video-based score than to the task-specific video-based scores.

Because 4 judges contributed to a task-specific domain score whereas 60 judges contributed to a composite domain score, a higher correlation with a composite domain score might reflect its higher reliability in terms of generalizability of trait impressions across judges. Alternatively, a composite domain score might reflect a more accurate description of the targets' personality. This was checked by adjusting all correlations for lack of reliability of the video-based judgments. The adjusted coefficients are reported

Table 5
Correlations of Video-Based Ratings with Descriptions by Knowledgeable Informants on the NEO Five-Factor Inventory

Personality domain	Level of aggregation (no. of tasks)	Report by			
		Self	Acquaintance	Experimenter	Confederate
Neuroticism	1	.11 (.15)	.13 (.18)	.26 (.35)	.24 (.33)
	15	.18 (.19)	.22 (.23)	.43 (.44)	.40 (.41)
Extraversion	1	.21 (.24)	.21 (.24)	.37 (.42)	.35 (.40)
	15	.30 (.31)	.31 (.32)	.55 (.56)	.53 (.54)
Openness to Experience	1	.17 (.20)	.15 (.18)	.28 (.33)	.23 (.27)
	15	.31 (.32)	.28 (.29)	.52 (.54)	.41 (.43)
Agreeableness	1	.13 (.17)	.12 (.15)	.16 (.21)	.21 (.27)
	15	.22 (.23)	.20 (.21)	.28 (.29)	.36 (.37)
Conscientiousness	1	.05 (.06)	.09 (.12)	.25 (.32)	.26 (.33)
	15	.08 (.08)	.16 (.17)	.47 (.49)	.47 (.49)
M	1	.13 (.16)	.14 (.17)	.27 (.33)	.26 (.32)
	15	.22 (.23)	.24 (.25)	.45 (.47)	.47 (.48)

Note. Coefficients in parentheses are corrected for lack of reliability of video-based judgments.

in parentheses and show that differences in reliability are not sufficient to explain the higher correlations with the composite scores. Rather, a broader sampling of behaviors resulted in more accurate measures of personality.

Differences Between Tasks

Were some of the 15 tasks systematically more diagnostic of the targets' personality than others? This is interesting because it might reveal which tasks yield particularly accurate judgments of specific personality traits. However, differences between tasks in the agreement of video-based judgments with measures of personality and intelligence might also reflect random fluctuations. This was checked by comparing, across tasks and between the two subsamples of participants, the rank order of the correlations of the video-based judgments with the personality descriptions by the targets and their acquaintances. These correlations between vectors of 15 correlations were −.05 and .32 (Neuroticism), −.07 and .03 (Extraversion), .51 and .69 (Openness to Experience), −.14 and −.40 (Agreeableness), and .08 and .22 (Conscientiousness) for self-reports and reports by acquaintances, respectively. Thus, reliable differences in the diagnosticity of the different tasks were obtained for Openness to Experience only. Correlations of self-reports and reports by acquaintances with the task-specific video-based Openness to Experience domain scores are reported in columns 1 and 2 of Table 6. Inferences from describing multiple uses of a brick by pantomime were most strongly related whereas inferences from role-playing in a dyadic interaction were least related to self-reports and acquaintance reports of Openness to Experience. Controlling for gender and age had no systematic effects on the size of these correlations.

Effects of Acquaintance on Accuracy

To study the effects of acquaintance and consistency on accuracy more systematically, we checked how correlations of personality descriptions by knowledgeable informants with video-based ratings from thin slices of behavior increased with the number of tasks that were included in the video-based score. To this end, we (a) combined video-based ratings for different tasks into parcels that included from 1 to 15 tasks; (b) calculated the composite score of those ratings that were included in the same parcel; and (c) correlated the composite video-based score with reports by self, acquaintance, experimenter, and confederate. The formula for the total number of different combinations of k out of n tasks is $n!/[k!(n - k)!]$, with $n = 15$ and k varying from 1 to 15 in the present study. For example, there were 15 different 1-task parcels, 105 possible 2-task parcels, 6,435 possible 7-task parcels, and one 15-task parcel. To reduce this association of parcel size with number of different parcels, we did not calculate all possible parcels but limited the number of different parcels of the same size to 105. If more than 105 different parcels of a fixed size were possible, a random-number generator was used to select those 105 parcels that were actually included. Separate analyses were then run for the two participant subsamples, resulting in 210 correlations for parcels of each size, except for 1-task parcels (30 correlations), 14-task parcels (30 correlations), and the 15-task parcel (2 correlations). Because these analyses were repeated for the domains of the five-factor model of personality and for comparisons with self-reports and reports by acquaintances, experimenters, and confederates, a total of 51,640 correlations were calculated. Correlations of the same type (same parcel size, same personality domain, same type of judge) were averaged, using Fisher's Z transformation for correlations. The findings are reported in Figures 1, 2, 3, and 4, which refer to correlations with reports by targets, acquaintances, experimenters, and confederates, respectively.

Obviously, the effect of parcel size on agreement was not linear; rather, the second derivative of all curves was negative, suggesting a rule of diminishing returns. Moreover, the figures suggest that correlations of descriptions by knowledgeable informants with

Table 6
Correlations of Task-Specific Video-Based Ratings of Openness to Experience and of Intelligence With Descriptions for Openness to Experience and With Psychometric Intelligence

	Openness to Experience		Intelligence tests	
Task	Self-report	Acquaintance report	APM	LPS–PCS
1. Introduce oneself	.16 (.19)	.19 (.21)	.28 (.21)	.36 (.28)
2. Picture arrangement	.20 (.17)	.18 (.17)	.26 (.14)	.33 (.28)
3. TAT stories	.17 (.15)	.18 (.18)	.24 (.11)	.33 (.24)
4. Tell a joke	.20 (.17)	.16 (.14)	.18 (.10)	.28 (.23)
5. Role-playing—noise	.12 (.10)	.05 (.05)	.19 (.14)	.21 (.21)
6. Role-playing—accident	.07 (.06)	.05 (.04)	.17 (.10)	.22 (.13)
7. Conversation on hobbies	.24 (.25)	.22 (.23)	.21 (.13)	.33 (.25)
8. Object recall	.22 (.23)	.23 (.25)	.27 (.16)	.39 (.31)
9. Frustrating problem	.12 (.10)	.09 (.09)	.23 (.07)	.30 (.19)
10. Introduce a stranger	.15 (.12)	.14 (.13)	.35 (.19)	.34 (.24)
11. Invent a neologism	.22 (.20)	.17 (.16)	.36 (.20)	.40 (.36)
12. Paper tower	.20 (.19)	.11 (.11)	.30 (.22)	.31 (.28)
13. Headlines	.15 (.13)	.15 (.15)	.35 (.29)	.53 (.51)
14. Pantomime	.29 (.29)	.27 (.26)	.23 (.19)	.31 (.29)
15. Sing a song	.06 (.07)	.09 (.10)	.24 (.07)	.27 (.18)

Note. Coefficients in parentheses are second-order correlations that control for gender and age. APM = Advanced Progressive Matrices; LPS-PCS = Leistungspruefsystem [Performance Test System]–principal-component score.

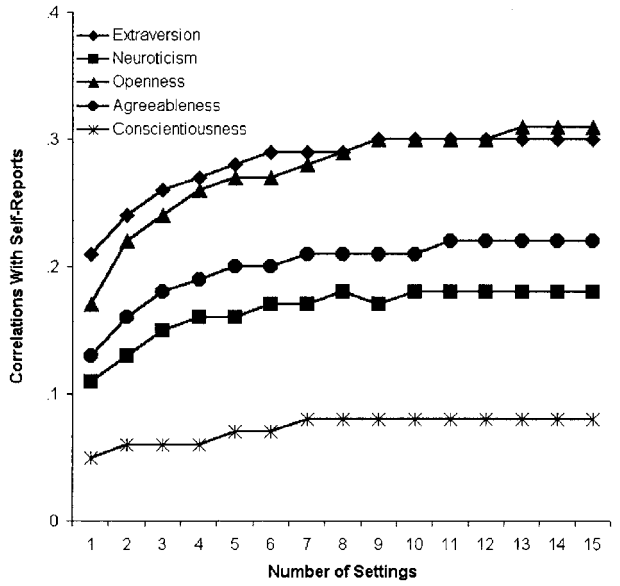

Figure 1. Correlations of self-reports of personality with video-based personality inferences as a function of number of settings included in the video-based score.

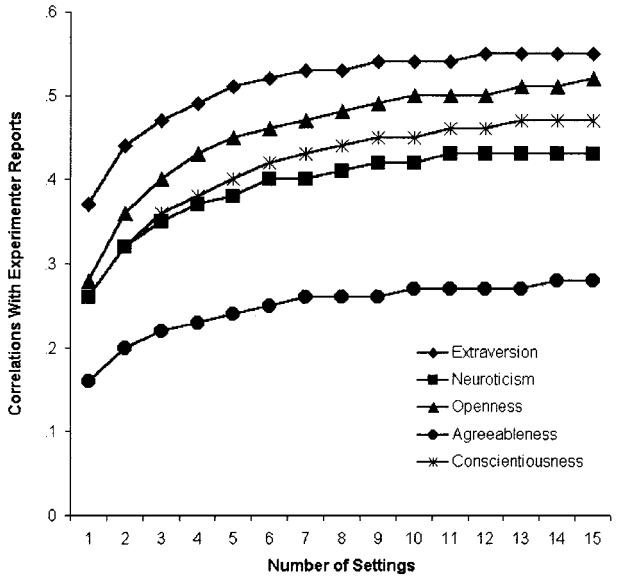

Figure 3. Correlations of personality descriptions by the experimenter with video-based personality inferences as a function of number of settings included in the video-based score.

video-based ratings of Extraversion approached their asymptotic value when about six behavioral episodes were included, whereas the validity of video-based ratings of Openness to Experience increased more strongly with the inclusion of additional tasks. This is consistent with predictions from WAM (Kenny, 1991), because impressions of Extraversion were most consistent and impressions of Openness to Experience were least consistent across tasks (see Table 3). To check whether our data justified an interpretation of differences in shape between the curves in Figures 1–4, the *Z*-transformed correlations were submitted to analyses of variance with parcel size as a within-subject factor and personality domain as a between-subjects factor. Independent analyses were run for correlations with self-reports, reports by acquaintances, experimenter reports, and confederate reports. In each of these analyses, the effects of parcel size were highly significant, all Fs(14, 14630) > 550, ps $< .001$, as were the effects of personality

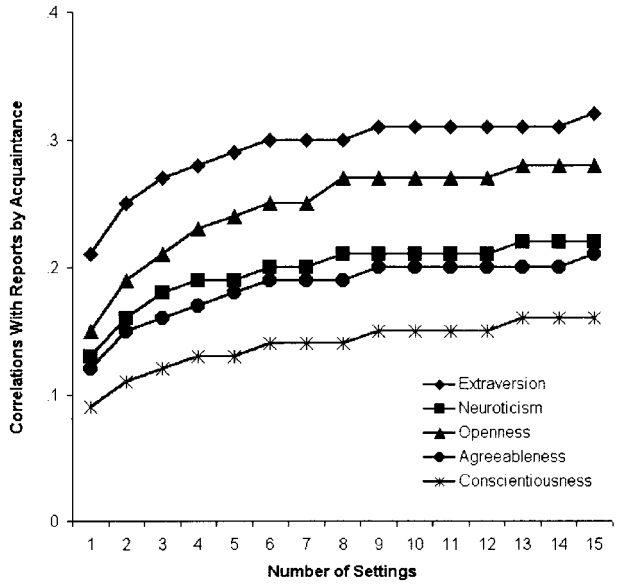

Figure 2. Correlations of personality descriptions by one acquaintance with video-based personality inferences as a function of number of settings included in the video-based score.

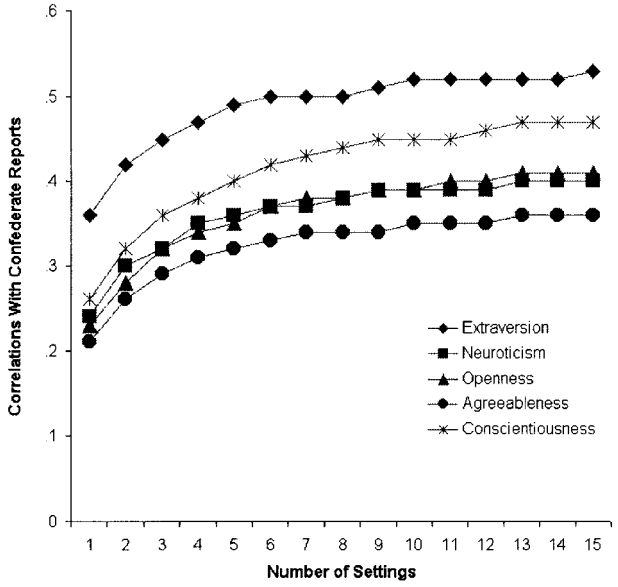

Figure 4. Correlations of personality descriptions by the confederate with video-based personality inferences as a function of number of settings included in the video-based score.

domain, all Fs(4, 1045) > 3020, ps < .001, and the Parcel Size × Domain interactions, all Fs(56, 14630) > 10.73, ps < .001. The significant interactions show that the differences in the shape of the curves do not merely reflect sampling error.

Findings for Intelligence

The subtests of the LPS were submitted to a principal-components (PC) analysis. The first unrotated component accounted for 51% of the total variance, and the LPS–PC score was used in the further analyses. The correlation between APM and the LPS–PC was .65. Concerning the video-based judgments of intelligence, Table 3 (third row from the bottom) shows that the average consensus between two individual judges (based on observations in the same setting) was .30, the reliability of the mean rating of four judges was .62, the average correlation between the mean ratings from observations in two different settings was .38, and the estimate of cross-situational consistency was .61—that is, higher than for the other traits.

The mean correlations of the intelligence test scores with the video-based inferences of intelligence from single tasks were .26 and .33, and the correlations with the composite score, based on the ratings of intelligence in all 15 tasks, were .41 and .53 for the APM and the LPS–PC, respectively. Because the LPS includes verbal and nonverbal tests, whereas the APM measures nonverbal intelligence exclusively, higher correlations of the video-based ratings with the LPS–PC than with the APM might reflect that observer-rated intelligence is more strongly related to crystallized than to fluid intelligence.

Effects of gender and age. The correlations of gender and age with the ratings of intelligence are reported in the bottom row of Table 4: In our sample, males were perceived as more intelligent than females, and younger participants were perceived as more intelligent than older participants. The correlations with the tests of intelligence were even stronger: The correlations with gender were −.20 and −.21, and those with age were −.27 and −.57 for the LPS–PC and the APM, respectively. Thus, insofar as ratings of intelligence were influenced by gender and age stereotypes, these stereotypes reflected a substantial kernel of truth.

To control for effects of stereotypes on consensus and accuracy in ratings of intelligence, second-order consensus and accuracy correlations were calculated that controlled for gender and age. Concerning estimates of consensus and cross-situational consistency, this had only small effects (Table 3, third row from the bottom), suggesting that stereotypes were unimportant for consensus and consistency in impressions of intelligence. However, controlling for gender and age had substantial effects on the correlation between the APM and the LPS–PC, which dropped from .65 to .46, and on the correlations of the aggregated rating of intelligence with the intelligence tests, which dropped from .41 to .29 for the APM and from .53 to .46 for the LPS–PC. Thus, gender and age stereotypes contributed to accuracy in ratings of intelligence.

Differences between tasks. To check whether differences between tasks in the accuracy of the inferences of intelligence were reliable, the rank order of the correlations of ratings with psychometric tests of intelligence was compared between subsamples across tasks. This resulted in correlations, across vectors of 15 correlations, of .50 for the APM and .49 for the LPS–PC. Thus, inferences of intelligence from some tasks were systematically more accurate than inferences from other tasks. Correlations of the task-specific video-based ratings of intelligence with the two intelligence tests are reported in the last two columns of Table 6. Inferences of intelligence from Task 13 (reading 14 newspaper headlines and subtitles aloud) were more strongly related to the LPS–PC in both participant subsamples, and to the APM score in one subsample, than were inferences of intelligence from any of the 14 other tasks. With averaged correlations of .53 with the LPS–PC and of .35 with the APM, performance on the headlines task yielded particularly accurate impressions of the targets' intelligence. As before, correlations involving intelligence tests declined substantially if target gender and age were controlled, but the correlations with ratings of intelligence based on performance in the headlines task were affected only slightly.

Correlations of the APM and the LPS–PC with the video-based ratings of intelligence, as a function of the number of settings included in the video-based score, are reported in Figure 5. Once again, the correlations approached their asymptotic value when about six behavioral episodes were included.

Discussion

The present study yielded four main findings: First, various predictions of WAM (Kenny, 1991, 1994) were confirmed. Second, the cross-situational consistency of personality impressions was substantially higher than the cross-situational consistency of operant behaviors. Third, stereotypes related to gender and age were widely shared, and they had a kernel of truth. Moreover, consensus and cross-situational consistency of personality impressions were hardly reduced if these stereotypes were controlled. Finally, we identified tasks that are particularly diagnostic of Openness to Experience and of intelligence.

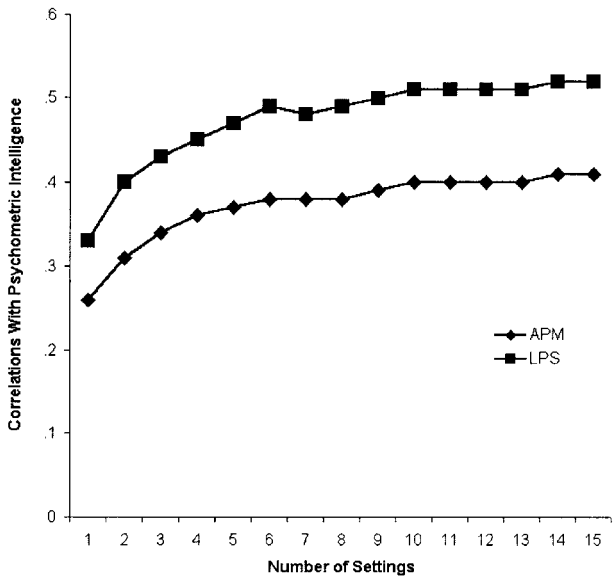

Figure 5. Correlations between measured intelligence and video-based inferences of intelligence as a function of number of settings included in the video-based score. APM = Advanced Progressive Matrices; LPS = Leistungspruefsystem [Performance Test System].

WAM

The results of the present study are relevant for the WAM parameters acquaintance, similarity of meaning systems, stereotypes, behavioral overlap, and consistency.

Acquaintance. As predicted by WAM, subsequent behavioral information contributes less to accuracy than preceding information; the curves plotted in Figures 1–5 have a negative second derivative. The fact that the present study confirms this prediction from WAM is important but not surprising, given that the rating composites were calculated as averages of independent judgments for different behavioral sequences. Consistent with WAM, the weight of the behavioral information was independent of any sequence in which it was presented, and different findings might be obtained if the same judges successively observed new behaviors. This would be more representative of everyday acquaintance processes where perceivers form a first impression that is modified when additional information becomes available. In that case, primacy effects may occur; that is, subsequent information may receive less weight than preceding information (Jones, Rock, Shaver, Goethals, & Ward, 1968). Under such circumstances, weaker effects of acquaintance on accuracy than predicted by WAM may be found.

Accuracy approached its asymptotic value after inclusion of about six behavioral episodes, but whether this is a general rule should be clarified in future research. It might well be that accuracy approaches its maximum later if "thinner" slices of behavior are available. After all, the behavioral sequences that were used in the present study spanned one or several minutes, leaving room for studies using much shorter sequences.

Most analyses in the present study relate judgments based on thin slices of behavior to descriptions by knowledgeable informants, but the inclusion of intelligence tests provides additional "hard" criteria of the accuracy of judgments from thin slices of behavior. The findings are striking: Correlations of video-based ratings with performance measures of intelligence (Figure 5) are even higher than the agreement of video-based ratings of personality with self-reports and descriptions by acquaintances (Figures 1 and 2). This shows that inferences of intelligence from thin slices of behavior are accurate indeed. Moreover, the video-based ratings of intelligence are more strongly related to the LPS than to the APM, suggesting that verbal intelligence is particularly important for lay perceptions of intelligence. Because the tasks assigned to the participants included a card-sorting test, a memory test, and a nonverbal creativity test, this finding is unlikely to reflect a higher availability of cues for verbal than for nonverbal intelligence.

Shared meaning systems. According to WAM, same-act/cross-judge correlations reflect shared meaning systems and shared stereotypes, and we obtained mean same-act/cross-judge correlations of .28 at the level of single adjectives and of .37 at the level of four-item domain scores (Table 3, column 1). The difference between these two mean coefficients probably reflects differences in reliability that are lower for one-item than for four-item measures. Because WAM does not include a parameter for error of measurement, error is confounded with nonshared meaning systems. Thus, more reliable measures of personality impressions yield more appropriate estimates of the extent of shared meaning systems, and .37 is therefore a more appropriate estimate than .28.

A consensus estimate of .37 far exceeds estimates of about .15 that have been suggested by Kenny and his associates (Kenny, 1994; Kenny, Albright, Malloy, & Kashy, 1994). Moreover, the difference cannot be explained by different data-analytic approaches: If the video-based ratings of the present study are analyzed by two-way analysis of variance with judges and targets as factors, the average proportion of variance explained by target main effects is .31, whereas the average proportion of variance explained by main effects of judges is .10.

The main source of the different findings is probably that Kenny (1994) based his conclusions on round-robin analyses: Students first interacted in small groups and then mutually described all group members including themselves. Because the judges in those studies had to form impressions of several targets while actively participating in social interactions, their cognitive load was high and their judgment task therefore difficult. Consensus depends on the particulars of the judgment task, and we believe that an estimate of .15 is representative of a difficult task where the judges are not given the opportunity to concentrate on the behavior of one target at a time, whereas .31 or .37 are representative estimates of consensus if the judges are given that opportunity. Consensus for specific trait domains varies around this value, and it seems to be highest for Extraversion.

Stereotypes. According to WAM, consensus reflects not only shared meaning systems but also shared stereotypes. We therefore ran various analyses in which target gender and age were controlled. This, however, had only negligible effects, if any: Stereotypes related to gender and age were unimportant for consensus and self–other agreement, and their contribution to cross-situational consistency of personality impressions was minor. By contrast, the effects of stereotypes related to age and gender on the accuracy of ratings of intelligence were pronounced, reflecting that these stereotypes have a substantial kernel of truth that contributes to accuracy in ratings of intelligence.

Overlap. The present study testifies to the importance of behavioral overlap for the similarity of personality impressions by different judges: The agreement between judges who both observed the targets either outside or within the study was substantially higher than the agreement of self-reports and reports by acquaintances with experimenter reports, confederate reports, and impressions based on thin slices of behavior. Information overlap is the most reasonable explanation of the different levels of agreement between judges who relied on similar or dissimilar samples of behavior.

In particular, overlap was important for judgments of Conscientiousness. For this trait, agreement of the targets and their acquaintances with experimenters, confederates, and judges of thin slices was low. This is particularly surprising because other studies, where judges had less information on the targets' behavior, have shown higher levels of self–other agreement for Conscientiousness (Albright et al., 1988; Borkenau & Liebler, 1992b; Kenny et al., 1994; Norman & Goldberg, 1966). This suggests that Conscientiousness may be quite accurately inferred from physical appearance variables like refined appearance and formal dress (Borkenau & Liebler, 1995) but that the behavior observed in the present study conveyed misleading information on the everyday conscientiousness of the participants. If there are strong stereotypes and the behavior is not consistent, WAM predicts a dip in the curve that describes the effects of acquaintance on consensus

(Kenny, 1994). Our findings for Conscientiousness may well illustrate that phenomenon.

Cross-Situational Consistency

A strength of the present study is that it allows separation of cross-situational consistency from consensus in personality impressions, thus estimating the true consistency of personality apart from unreliability of measurement. Our estimates of cross-situational consistency varied from .35 for Openness to Experience to .61 for intelligence, with a mean of .43. This figure is much higher than the estimates of cross-situational consistency of .23 in Hartshorne and May's (1928) study and of .13 in Mischel and Peake's (1982) study, and it resembles the estimate of cross-situational consistency of .46 at the level of global factors in the Riverside Accuracy Project (Funder & Colvin, 1991). There are at least three reasonable explanations for why Hartshorne and May's, and particularly Mischel and Peake's, findings differ from Funder and Colvin's (1991) and ours: First, Mischel and Peake's extremely low consistency estimates of about .13 may reflect that compared with the other personality domains studied here, Conscientiousness is a personality domain that is less consensually inferred from behavior and less consistent across situations. Moreover, Hartshorne and May and Mischel and Peake focused on the frequencies that their research participants showed specific behaviors, whereas Funder and Colvin and the present study focused on how participants performed in assigned tasks. This suggests that behavioral style is more cross-situationally consistent than specific operant behaviors that may be influenced by many traits and situation-specific concerns (Ahadi & Diener, 1989; Allport, 1937). Third, the participants of the present study were considerably older than those in the Hartshorne and May and in the Mischel and Peake studies, and older persons may be more consistent than younger ones. In a recent meta-analysis, Roberts and DelVecchio (2000) concluded that the rank-order stability of personality increases from childhood until the 6th decade of life, and it may well be that a similar rule holds for the cross-situational consistency of behavior, although we are not aware of a study that has investigated this issue empirically.

Our estimates of cross-situational consistency do not differ between one-item adjective scales and four-item measures of personality domains; the mean is .43 in both analyses. By contrast, Funder and Colvin (1991) reported higher estimates at the level of common factors (.46) than at the level of single behavioral Q-sort items (.28). To some extent, this difference reflects that Funder and Colvin's factor scores are more reliable than their single Q-sort items and that their consistency estimates were not corrected for lack of consensus. The fact that Funder and Colvin obtained similar uncorrected estimates of cross-situational consistency at the level of common factors (.46) as we obtained for corrected estimates (.43) may reflect the fact that the participants of the present study were observed in a larger and more diverse sample of settings.

Mischel and Shoda (1995) claimed that the cross-situational consistency of individual differences in behavior is about .10, whereas the intraindividual stability of Situation × Behavior interactions is much higher and varies about $r = .40$. Therefore, they suggested studying the cognitive–affective units that underlie stable situation–behavior profiles that they refer to as a person's *personality signature.* Our findings do not imply that mediational processes should not be studied, but they show that personality consistencies may be found at two levels, at least: (a) at the level of stable Situation × Behavior interactions and (b) at the level of cross-situationally consistent impressions of personality.

Importance of Stereotypes

When we controlled for the effects of gender and age in the present study, we found that these variables are only weakly related to self-reports and ratings by others. Moreover, correlations with ratings by others are not systematically higher than correlations with self-reports. Particularly interesting are the findings for intelligence, because performance on intelligence tests is not biased by stereotypes. Thus, a comparison of the correlations of gender and age with (a) video-based ratings of intelligence and (b) performance on intelligence tests allows for a direct test of the kernel-of-truth hypothesis. Correlations of gender with ratings and with tests of intelligence are of the same size (about .20), whereas age is more strongly related to psychometric than to video-based ratings of intelligence. This finding indicates that at least for intelligence, video-based ratings do not overestimate the actual differences between groups. The same cannot be shown as clearly for the other traits, but it is reasonable to assume that effects of gender and age on trait impressions do not differ fundamentally between intelligence and the Big Five.

Moreover, we examined to what extent controlling for the effects of gender and age reduces the agreement between judges concerning the targets' personality traits. Note that these analyses underestimated the effects of similar inferences from behavior on consensus, because actual gender and age differences in personality and intelligence do exist. The correlation that is most strongly reduced by the control of gender and age is that between the two intelligence tests, suggesting that controlling for stereotypes by controlling for group membership may be pouring out the baby with the bathwater. For the agreement among judgments, this problem is less important, because controlling for gender and age hardly reduces estimates of agreement and cross-situational consistency.

Experimental research has shown that group membership is an important source of personality impressions if behavioral information is scarce but that its importance decreases rapidly as soon as additional behavioral information becomes available (Krueger & Rothbart, 1988). The present more naturalistic study supports this view as do findings by Borkenau and Liebler (1995), who found that the correlation between stranger ratings of physical attractiveness and of intelligence dropped from .63 to .28 as soon as observers viewed a sound film instead of a silent film of the strangers reading a standard text. Thus, the more cues for intelligence become available, the less intelligence is inferred from perceived attractiveness. In the context of attribution research, Heider (1958) observed that behavior engulfs the field (p. 54), and it seems that behavior also engulfs those stereotypes that are important for first impressions.

Obviously, there may have been other stereotypes—for example, stereotypes related to physical appearance—that contributed to consensus but were not controlled in the present study. Although this cannot be ruled out completely, there are at least two problems with explaining consensus by such uncontrolled stereotypes. First,

the physical appearance of the targets remains stable across tasks, whereas their behavior varies. Thus, if physical appearance stereotypes were important for consensus, high estimates of cross-situational consistency should be found. That, however, is not the case: The same-act/cross-judge correlation is—on average—more than twice as strong as the cross-act/cross-judge correlation, suggesting that the judges mostly rely on information that varies between tasks. Second, differences in appearance reflect differences in behavior to some extent, particularly if effects of gender and age are controlled. This was nicely illustrated in a study by Diener, Wolsic, and Fujita (1995). These authors found that the association between physical attractiveness and subjective well-being was stronger if ratings of attractiveness were based on pictures that showed the targets as they chose to appear for the study compared with pictures that showed the targets in their "natural beauty," that is, without jewelry and with their hair and clothes covered. To some extent, the attractiveness–happiness relation reflects that happy persons take more care of their appearance and are therefore perceived as more attractive. Beyond age, gender, and ethnicity, the empirical distinction between shared inferences from behavior and shared stereotypes is difficult because behavior influences appearance.

Diagnostic Tasks

The 15 tasks under study differ reliably in their diagnosticity for only two of six trait domains, and these domains are related to cognitive ability. Whereas interpersonal traits seem to be inferred equally well from various sorts of behavior, it seems that accurate inferences of Openness to Experience and of intelligence require observations of ability-demanding behavior or of behavioral residues like the books on one's bookshelf (Gosling et al., 2002) or the recordings that one buys (Rentfrow & Gosling, 2003). In the present study, Openness to Experience is most strongly related to ratings based on the pantomime task, where the participants have to invent multiple uses of a brick and demonstrate them by pantomime. Obviously, this task requires creativity both for inventing the multiple uses and then communicating them without use of words. By contrast, ratings of Openness that were based on the role-playing situations where the targets had to deal with interpersonal conflicts do not agree well with self-reports and acquaintance reports of Openness to Experience. The other correlations in Table 6 are consistent with the view that tasks requiring cognitive abilities (or revealing one's hobbies) are more diagnostic of one's Openness to Experience than tasks that require social skills.

Performance on intelligence tests is most strongly related to inferences from observing targets who read short phrases aloud, as in the headlines task. This suggests that the way people read standard text provides particularly diagnostic cues of their psychometric intelligence. This resembles findings by Borkenau and Liebler (1993), who found that inferences of intelligence based on observations of how strangers read a standard text allow for quite accurate judgments of their intelligence. This finding is interesting from a theoretical as well as from an applied perspective.

Concerning theory, the finding is consistent with the assumption that intelligent persons perform simple tasks more efficiently (Deary, 2000). According to this view, measured intelligence should not only be related to how persons solve complex problems but also to how efficiently they perform on everyday tasks like understanding and reading a short text. In this context, the strength of the correlation of judgments of intelligence from the headlines task with the performance measures of intelligence becomes important: Whereas the LPS is correlated at .53 with ratings from the headline task (.52 and .53 in the two subsamples), it is correlated at .65 with the APM. Moreover, if gender and age are controlled, the correlation of the LPS with the APM drops to .46, whereas the correlation with the headline task ($r = .51$) is hardly affected. The finding that the LPS is related to the headline task nearly as strongly as it is related to another test of intelligence shows that the headline task is a quite precise measure of intelligence.

Moreover, the strength of the correlation, in combination with the ease of administering this task, makes this an interesting finding from an applied perspective as well, especially when a short and cost-effective measure of intelligence is required or conventional tests of intelligence cannot be used. Certainly, this finding should be replicated before the replacement of intelligence tests with judgments of intelligence is considered.

Limitations

The main limitation of this study is the composition of the participant sample. First, they were twins, and twins constitute a somewhat special population. Second, they were mostly women, and a more balanced gender ratio would have been desirable. Third, the age range was quite large, but because most participants were in their 20s and 30s, the age distribution is not representative of the German population. We believe, however, that these are minor problems in the present context.

Another limitation is that the extent of information that was available to a judge was not systematically varied within judges. Rather, acquaintance was varied by aggregating different numbers of independent judgments. This allowed estimation of the cross-situational consistency of personality impressions, but it raises questions concerning the generalizability of our findings concerning the effects of acquaintance on agreement and accuracy in everyday acquaintance processes. Thus, the present study contributes new and interesting findings, but it also points to additional studies that have yet to be conducted.

References

Ahadi, S., & Diener, E. (1989). Multiple determinants and effect size. *Journal of Personality and Social Psychology, 56,* 398–406.

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 55,* 387–395.

Allport, G. W. (1937). *Personality: A psychological interpretation.* New York: Holt.

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology, 32,* 201–271.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111,* 256–274.

Angleitner, A., Ostendorf, F., & John, O. P. (1990). Towards a taxonomy of personality descriptors in German: A psycho-lexical study. *European Journal of Personality, 4,* 89–118.

Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology, 34,* 164–181.

Borkenau, P., & Liebler, A. (1992a). The cross-modal consistency of personality: Inferring strangers' traits from visual or acoustic information. *Journal of Research in Personality, 26,* 183–204.

Borkenau, P., & Liebler, A. (1992b). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology, 62,* 645–657.

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., & Liebler, A. (1995). Observable attributes as cues and manifestations of personality and intelligence. *Journal of Personality, 63,* 1–25.

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fuenf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae* [NEO Five-Factor Inventory by Costa and McCrae]. Goettingen, Germany: Hogrefe.

Borkenau, P., Riemann, R., Angleitner, A., & Spinath, F. M. (2001). Genetic and environmental influences on observed personality: Evidence from the German Observational Study of Adult Twins. *Journal of Personality and Social Psychology, 80,* 655–668.

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

Deary, I. J. (2000). Simple information processing and intelligence. In R. Sternberg (Ed.), *Handbook of intelligence* (pp. 267–284). New York: Cambridge University Press.

Diener, E., Wolsic, B., & Fujita, F. (1995). Physical attractiveness and subjective well-being. *Journal of Personality and Social Psychology, 69,* 120–129.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652–670.

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception.* San Diego, CA: 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., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology, 60,* 773–794.

Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology, 64,* 479–490.

Funder, D. C., & West, S. G. (1993). Consensus, self–other agreement, and accuracy in personality judgment: An introduction. *Journal of Personality, 61,* 457–476.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59,* 1216–1229.

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology, 82,* 379–398.

Harker, L. A., & Keltner, D. (2001). Expressions of positive emotion in women's college yearbook pictures and their relationship to personality and life outcomes across adulthood. *Journal of Personality and Social Psychology, 80,* 112–124.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character: I. Studies in deceit.* New York: Macmillan.

Heider, F. (1958). *The psychology of interpersonal relations.* New York: Wiley.

Horn, W. (1962). *Leistungspruefsystem* [Performance Test System]. Goettingen (Germany): Hogrefe.

Jones, E. E., Rock, L., Shaver, K. G., Goethals, G. R., & Ward, L. M. (1968). Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology, 10,* 317–341.

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review, 98,* 155–163.

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

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the Big Five. *Psychological Bulletin, 116,* 245–258.

Kenny, D. A., Horner, C., Kashy, D. A., & Chu, L. (1992). Consensus at zero acquaintance: Replication, behavioral cues, and stability. *Journal of Personality and Social Psychology, 62,* 88–97.

Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology, 55,* 187–195.

McCrae, R. R. (1982). Consensual validation of personality traits. Evidence from self-reports and ratings. *Journal of Personality and Social Psychology, 43,* 293–303.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52,* 81–90.

Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89,* 730–755.

Mischel, W., & Shoda, Y. (1995). A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102,* 246–268.

Murray, H. A. (1943). *Thematic Apperception Test manual.* Cambridge, MA: Harvard University Press.

Neubauer, A. C., Spinath, F. M., Riemann, R., Borkenau, P., & Angleitner, A. (2000). Genetic and environmental influences on two measures of speed of information processing and their relation to psychometric intelligence. *Intelligence, 28,* 267–289.

Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology, 4,* 681–691.

Paulhus, D. L., & Bruce, M. N. (1992). The effect of acquaintanceship on the validity of personality impressions: A longitudinal study. *Journal of Personality and Social Psychology, 63,* 816–824.

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.

Raven, J. C. (1958). *Advanced Progressive Matrices.* London: Lewis.

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure of personality correlates of music preferences. *Journal of Personality and Social Psychology, 84,* 1236–1256.

Riemann, R., Angleitner, A., & Strelau, J. (1997). Genetic and environmental influences on personality: A study of twins reared together using the self- and peer-report NEO-FFI scales. *Journal of Personality, 65,* 449–475.

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.

Spinath, F. M. (1999). *Validitaet von Fremdbeurteilungen: Einflussfaktoren auf die Konvergenz von Selbst- und Fremdbeuteilungen in Persoenlichkeitseinschätzungen* [Validity of stranger ratings: Influences on self–other agreement in descriptions of personality]. Lengerich, Germany: Pabst Science Publishers.

Spinath, F. M., Riemann, R., Hempel, S., Schlangen, B., Weiss, R., Borkenau, P., & Angleitner, A. (1999). A day in the life: Description of the German Observational Study of Adult Twins (GOSAT) assessing twin similarity in controlled laboratory settings. In I. Mervielde, I.

Deary, F. de Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 311–328). Tilburg, the Netherlands: Tilburg University Press.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change. *Journal of Personality and Social Psychology, 84,* 1041–1053.

Sturm, W., & Willmes, K. (1983). LPS-K—eine LPS-Kurzform für hirngeschaedigte Patienten; mit Anleitung zur psychometrischen Einzelfalldiagnostik [LPS-K—a short form of the LPS for patients suffering from brain damage with instructions for single-case diagnoses]. *Diagnostica, 29,* 346–358.

Received March 20, 2003
Revision received October 7, 2003
Accepted October 15, 2003 ■