Journal of Personality and Social Psychology
1987, Vol. 52, No. 2, 409-418

0022-3514/87/$00.75

Differences Between Traits: Properties Associated With Interjudge Agreement

David C. Funder
University of Illinois at Urbana-Champaign

Kathryn M. Dobroth
Harvard University

The present study concerns the relation between properties of personality traits and the agreement with which they are applied to real individuals. Subjects rated the 100 personality items of the California Q-Set on nine subjective dimensions, six of which loaded highly on a first principal component. This factor was interpreted as reflecting each trait's "easy visibility" to an outside observer. Actual interjudge agreement in applying each trait to real individuals was assessed in two ways: Self–other agreement was assessed in two independent samples, and interpeer agreement was assessed in three samples. Impressive and stable agreement was found for most Q items. The traits that were applied to individuals with the greatest interjudge agreement were the same ones that seemed most easily visible and tended to be positively relevant to extraversion and negatively relevant to neuroticism (identified through a factor analysis by McCrae, Costa, & Busch, 1986). The results suggest (a) that traits defining extraversion are revealed relatively directly in social behavior and, therefore, are easy to judge, (b) that traits defining neuroticism are less visible and, so, are judged less accurately, and (c) that lay perceivers of personality are generally sensitive to this difference between traits.

An important kind of judgment in daily life concerns the personality traits of other people and ourselves. Evaluations of the people in our social environment are central to our decisions about who to befriend and avoid, trust and distrust, hire and fire, and so on. Moreover, descriptive judgments rendered by a subject's friends and acquaintances can provide a valuable tool for personality assessment and research (Funder, 1983; Funder & Harris, 1986; Moskowitz, 1986). A natural concern, therefore, is the degree to which such judgments might or might not be accurate.

On the one hand, modern research on social judgment has primarily focused on errors and shortcomings (e.g., Nisbett & Ross, 1980), which has produced widespread pessimism about people's ability to make accurate judgments of each other's personalities (cf. Christensen-Szalanski & Beach, 1984; Loftus & Beach, 1982). Although the relevance of research on error for accuracy issues can be questioned (Funder, 1987), informal observation is sufficient to establish that the judgments we make about each other in real life are frequently wrong. On the other hand, research has repeatedly demonstrated that peers' judgments of personality have an impressive ability to predict behavior even in the laboratory and sometimes across a span of a dozen years or more (Funder, 1983; Funder & Harris, 1986; Mischel, 1984; Moskowitz & Schwarz, 1982).

Rather than focus on *whether* personality judgments are valid, therefore, it might be more productive to investigate *when* they are valid. Although this was precisely the approach of early investigators such as Allport (1937), Guilford (1936), and Murray (1938; cf. Ozer, 1979), the latter question has lain relatively neglected in recent years. Modern research offers only a few hints about what factors might underlie differential accuracy.

There seem to be three principal candidates. The first is individual differences between judges. During the 1940s and 1950s, a great deal of effort was expended searching for the good judge of personality; unfortunately, most of this research contained serious methodological flaws (Cronbach, 1955; Hastorf & Bender, 1952), and the topic has all but disappeared in recent years (see Schneider, Hastorf, & Ellsworth, 1979, for a review). A recent exception is the study by Funder and Harris (1986), which demonstrates how social acuity is a complex combination of perceptual and social skills that is manifest in diverse ways.

A second candidate is individual differences among the persons judged. Bem and Allen (1974), for example, found that judgments by different raters agreed better about subjects who described themselves as consistent than those who described themselves as inconsistent on the trait in question. However, this finding is in dispute, with some investigators reporting at least partial replication (Campbell, 1985; Cheek, 1982; Kenrick & Stringfield, 1980) and others a failure to replicate (Chaplin & Goldberg, 1984).

Research on a third potential moderator of accuracy is almost nonexistent. That is the possibility that some traits can be judged more accurately than others. This possibility was examined in a preliminary way some years ago by Estes (1938). Subjects attempted to judge the personalities of stimulus persons viewed only on a brief movie film, and their accuracy was evaluated through comparison with judgments rendered of these persons by a panel of clinical judges. Estes found that, for example, inhibition–impulsion was judged more accurately than objectivity–projectivity. But no clear, overall pattern of results emerged, and it was also unclear whether the same traits that

This research was supported by National Institute of Mental Health Grants R01-MH40808 and R01-MH42427 to David C. Funder. We are grateful for the assistance of Paige Dennis and Elizabeth Gray.

Correspondence concerning this article should be addressed to David C. Funder, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, Illinois 61820.

were easiest to judge from the film would also be the easiest to judge in real life.

The topic has been almost completely neglected ever since. Yet a moment's reflection is sufficient to confirm that some traits do seem more easily visible and, therefore, more easy to judge than others. For example, some traits, such as *talkative,* refer to patterns of overt behavior that can be directly seen in a wide variety of situations. Others, such as *tends to fantasize,* refer to behaviors that cannot be seen directly, and arise in few interpersonal situations. Some traits, such as *ethically consistent,* must be evidenced by a large number of behaviors before they can be judged confidently; others, such as *verbally fluent,* require only a few behaviors to be confirmed.

The first purpose of the present study was to gather evidence about the characteristics of traits associated with the degree to which they seem easily visible. Several possibilities can be found in a recent, valuable study by Rothbart and Park (1986). These investigators had separate groups of subjects rate a large number of personality traits on eight dimensions relevant to their confirmability or disconfirmability. They included how easy it is to imagine specific, observable behaviors that would confirm or disconfirm the trait, how frequently occasions arise in daily life that allow confirming or disconfirming behaviors, how many confirming or disconfirming behaviors one would have to observe before considering the trait to be an accurate (or inaccurate) description of a person, and finally, each trait's prevalence in the population and its favorability.

Rothbart and Park found an interesting pattern of relations, mostly positive, between these dimensions across a set of 150 personality traits. However, their subjects did not rate directly how easy these traits would be to judge. Also missing from Rothbart and Park's data is any indication of what attributes of the content of traits might be related to these dimensions. Over the years, a large factor-analytic literature has determined that verbal descriptions of personality can be reduced to about five robust factors, usually named Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness (Goldberg, 1981; McCrae, Costa, & Busch, 1986; Norman, 1963). It seems unlikely that traits loading on these different factors would be equally easy for an observer to confirm or disconfirm. Neuroticism, for example, would seem relevant to traits that are less directly revealed in daily social interaction, and Extraversion would seem relevant to traits revealed by more visible, overt social behaviors. Data relevant to the relation between confirmability, subjective easiness, and the five dimensions of traits (as identified by McCrae et al., 1986) were gathered for the present study.

The second purpose of this study was to examine a more fundamental question: To what degree are these attributes of traits associated with the *actual* accuracy of personality judgment? Are the traits that seem the easiest to confirm or disconfirm actually judged more accurately? If so, this would be an encouraging finding, because it would suggest that people are sensitive to at least one of the factors that make their judgments more and less likely to be accurate. This would imply, in turn, that people might be most confident of their personality judgments when they are relatively likely to be correct.

The recent literature on decision making would seem to indicate that such sensitivity is unlikely. Several studies of human judgment in domains other than personality have concluded that people's confidence is, at best, uncorrelated with their accuracy (e.g., Einhorn & Hogarth, 1978; Fischoff, Slovic, & Lichtenstein, 1977; Goldberg, 1959; Oskamp, 1965). Kahneman and Tversky (1973) even claimed that "the factors which enhance confidence . . . are often *negatively* correlated with predictive accuracy" (p. 249, emphasis added).

Still, these findings must be balanced against those of a recent study conducted on judgments specifically relevant to personality. Epstein and Teraspulsky (1986) found that when estimating the degree of cross-situational consistency that would be found among different types of behaviors, individual subjects were most confident just when they were most likely to be accurate. Therefore, the issue merits further examination. The present study was designed to address directly the relation among a trait's content, subjective easiness to judge, other subjective properties, and the accuracy with which it is judged.

Accuracy is difficult to study because of the difficulty in finding a suitable criterion (cf. Funder, 1982a, 1987). Nonetheless, the issue is important and not completely intractable. The design of the present study was derived from the following prediction: If some traits can be judged more accurately than others, then agreement between judges applying these traits to real persons should be higher than for traits that are more difficult to judge accurately. This is not to say that interjudge agreement is sufficient to establish accuracy. Two judges can agree perfectly yet both be perfectly wrong. However, if two judges disagree then at least one of them must be wrong. For this reason, there are probabalistic grounds to expect judgments that agree to be more likely to be accurate than judgments that do not (cf. Funder & Van Ness, 1983). This expectation leads to the prediction that those traits judged most accurately will exhibit the best interjudge agreement.

An even more fundamental and pragmatic concern is that no superior criterion for accuracy is available, perhaps not even in principle. Although some psychologists might prefer objective behavioral criteria to the use of interjudge agreement, problems along that route were surveyed 50 years ago by Wolf and Murray in an analysis that remains apt today:

> Some psychologists . . . have attempted to devise tests . . . which will give objective measures of the traits. That few, if any of these tests have proved valuable is beside the point. The point is that it must be proved that the test tests the given trait. No one is seriously interested in a person's score on test *z* as such. It is only significant in so far as it measures the strength of variable *Z* as manifested in everyday life. To discover whether the latter is true one must rely upon the estimates of judges, which brings us back where we started. (Wolf & Murray, 1936, p. 364)

A considerable amount of data is needed to analyze traits according to the criterion of interjudge agreement. The central datum is a correlation coefficient (either Pearson or intraclass), and such coefficients are notoriously unstable when derived from the sample sizes (e.g., about 40) typical for psychological experiments. The present investigation, therefore, is based on five different samples in which subjects provided self-judgments of personality or were judged by two close acquaintances, with a combined total of 174 stimulus persons and 312 peer judges. This sample size was near the minimum necessary for the kinds of analyses to be reported in this article.

The purposes of this study were (a) to evaluate a set of 100 descriptive personality items (the California Q-Set; Block,

Table 1
Number of Subjects and Coefficient Alpha Values Prior to and Following Deletion of Unreliable Judges

Dimension	Original		Final	
	n	α	*n*	α
Ease of imagining confirming behaviors	10	.85	10	.85
Ease of imagining disconfirming behaviors	10	.67	9	.68
Occasions for confirming behaviors	9	.80	8	.83
Occasions for disconfirming behaviors	9	.68	7	.77
Number of instances to confirm	9	.80	9	.80
Number of instances to disconfirm	9	.63	9	.63
Frequency of trait in population	8	.73	8	.73
Favorability	9	.95	9	.95
Easy to judge	9	.87	9	.87
Mean α		.80		.81

1961/1978) along the eight dimensions introduced by Rothbart and Park plus subjective easiness to judge; (b) to assess the intercorrelations of these dimensions across this set of items and to determine the relation between these dimensions and the five factors of personality; (c) to assess the degree of self–other agreement and interpeer agreement by using the same personality items to describe several samples of real individuals; and (d) to examine the relation between the various dimensions of these items and their actual degree of interjudge agreement, both self–other and interpeer. The central hypothesis was that those traits that seem the most easily visible will manifest the best interjudge agreement.

Method

Overview

Nine independent groups of subjects rated the 100 items of the California Q-Set (Bem & Funder, 1978; Block, 1961/1978) on nine dimensions, eight of which were applied earlier to a different set of trait adjectives by Rothbart and Park (1986). The ninth dimension was subjective easiness to judge. Coefficient alpha was computed for each dimension to determine interrater reliability. The scale value for each trait on each dimension was obtained by summing across raters. The relevance of each Q-item to each of the five classic factors of personality was obtained from an analysis by McCrae et al. (1986).[1]

Indexes of interjudge agreement in applying each trait to real individuals were computed from five samples. Two separate groups of subjects provided self-descriptions with the use of the 100-item Q sort, and each subject was also judged by one or two peers, which allowed the calculation of two independent estimates of self–other agreement on each item. Similarly, Q-sort judgments of each subject were obtained from two close acquaintances in three separate samples, which allowed the calculation of three independent estimates of interpeer agreement on each item. The various estimates were combined into aggregate, summary indexes of the amount of self–other agreement, interpeer agreement, and total interjudge agreement.

Finally, these indexes of agreement were correlated with the ratings of the 100 Q items on each of the nine dimensions and the five sets of factor scores. The purpose was to detect properties of traits associated with interjudge agreement.

Subjects

Ratings of each of the 100 Q items on the nine dimensions were provided by 82 male and female Harvard University undergraduates, who were paid for their participation. Self-descriptions on the Q sort were obtained from 41 male and female Stanford University undergraduates and, later, from 64 Harvard undergraduates, all of whom were paid for their participation. In the first sample, two peers were recruited to judge 37 out of the 41 subjects; in the second sample, two peers evaluated 50 subjects out of 64; the remaining subjects were judged by one peer. Judgments were also obtained from two peers of each member of an additional sample of 69 Stanford undergraduates who did not provide self-descriptions. The combined subject sample consisted of 174 stimulus persons (105 of whom provided self-descriptions) and 312 peer informants.

Procedure

Trait ratings. The procedure for obtaining subjective judgments of nine trait properties closely followed that used by Rothbart and Park (1986), with two exceptions. One difference was that whereas the traits used by those investigators came from lists developed by Katz and Braly (1933), Anderson (1968), and Norman (1967), the traits in this study were the 100 items of the California Q-Set (Block, 1961/1978) as slightly modified by Bem and Funder (1978) for use with nonprofessionals. The Q set consists of 100 descriptive statements about personality, each typed on a separate card. For example, the first item reads "is critical, skeptical, not easily impressed," the second reads "is a genuinely dependable and responsible person," and so on. The set of items was developed over a period of years by a group of clinical and research psychologists, and it spans a wide spectrum of the personality domain. The Q set has been demonstrated to yield judgments useful for the prediction of behavior in a variety of contexts (see Bem & Funder, 1978; Funder, 1980b, 1982b, 1983).

The other difference between this study and the one by Rothbart and Park was that we asked a group of subjects to rate each item on the additional dimension of how easy to judge it seemed.

Subjects provided their ratings individually in a quiet room. The nine judgment tasks were distributed randomly across subjects—as each subject appeared, he or she was given the next questionnaire off the top of a randomly mixed stack. The instructions printed on each questionnaire were based as closely as possible on those used by Rothbart and Park (1986). Those pertaining to the one dimension not examined in the earlier study were as follows:

> We frequently use adjective traits to describe the characteristics of individuals or groups of individuals. For some traits that we use to describe people, judgments of whether someone has the trait or not are relatively easy. For other traits, their presence or absence is much more difficult to judge. For example, consider the trait "loud." It is ordinarily rather easy to determine whether someone has this trait or not. However, other traits, such as "has untapped potential" or "says only what he or she really believes" may be harder to judge for certain.
>
> For each of the . . . personality trait terms on the following pages, please rate how difficult or easy you think it would be to judge in another person.
>
> Please provide your judgment in the blank spot to the left of each trait, using an appropriate number from the following scale.

[1] We are grateful to Robert R. McCrae for providing us with these factor loadings.

Table 2
Intercorrelations Among Trait Properties

Trait property	α	1	2	3	4	5	6	7	8	9	10	11	12
1. Ease of imagining confirming behaviors	.85	—											
2. Ease of imagining disconfirming behaviors	.68	.59	—										
3. Occasions for confirming behaviors	.83	.65	.60	—									
4. Occasions for disconfirming behaviors	.77	.60	.64	.75	—								
5. Number of instances to confirm	.80	−.59	−.30	−.39	−.29	—							
6. Number of instances to disconfirm	.63	−.01	−.37	−.15	−.23	−.06	—						
7. Frequency of trait in population	.73	.20	.16	.31	.11	−.13	.11	—					
8. Favorability	.95	.10	.21	.52	.28	.19	−.14	.35	—				
9. Easy to judge	.87	.78	.62	.75	.67	−.57	−.11	.29	.28	—			
10. Self–other agreement	.46	.26	.34	.48	.37	−.08	−.11	.00	.43	.35	—		
11. Interpeer agreement	.54	.36	.20	.30	.21	−.28	−.08	−.10	.10	.31	.51	—	
12. Total agreement	.68	.36	.31	.45	.34	−.20	−.11	−.05	.31	.38	—	—	—

Note. For all analyses, $N = 100$, which was the number of Q items in the set used. For correlations of .20 or greater, $p < .05$, two-tailed. For correlations of .26 or greater, $p < .01$. For correlations of .34 or greater, $p < .001$, two-tailed.

Responses were given on a 9-point scale that ranged from *quite difficult* (1) to *quite easy* (9). Traits used as examples in the instructions did not appear in the actual list of traits to be rated. The questions asked for each of the nine scales were, following Rothbart and Park:

1. Imaginability of confirming behaviors: "How easy is it to imagine *specific, observable behaviors* that would provide confirmation of that trait?"

2. Imaginability of disconfirming behaviors: "How easy is it to imagine *specific, observable behaviors* that would *dis*confirm (provide evidence against) that trait?"

3. Frequency of occasions allowing confirming behaviors: "In the course of normal social interaction, *how frequently* do occasions arise that would allow for behaviors that *confirm* this trait?"

4. Frequency of occasions allowing disconfirming behaviors: "In the course of normal social interaction, *how frequently* do occasions arise that would allow for behaviors that *dis*confirm this trait?"

5. Number of behavioral instances required to confirm trait: "How many confirming behaviors would a person have to engage in before you would consider this trait to be an accurate description of that person?"

6. Number of behavioral instances required to disconfirm trait: "How many disconfirming behaviors would a person have to engage in before you would decide that the trait did not accurately describe that person?"

7. Prevalence of trait in the population: "How frequently would you expect to see this trait in the general population?"

8. Favorability: "How favorably or unfavorably would you regard a person who possessed this trait?"

9. Easiness: "How difficult or easy would it be to judge the degree to which another person had this trait?"

Each rating questionnaire began with 10 practice items, which were not scored, followed by the 100 items of the California Q-Set (Bem & Funder, 1978; Block, 1961/1978) in the order of their traditional numbering, which is random. The task typically required about 30 min.

Self-judgments and peers' judgments. In the current study as well as in previous studies conducted by Funder, the relevance of the items in the Q set to the personalities of real individuals was judged by the individuals themselves and by their close acquaintances. The procedure was similar in each case. A group of paid, volunteer subjects described their own personalities with the use of the Q sort (Bem & Funder, 1978; Block, 1961/1978).

The task of a judge who uses this instrument is to sort its 100 descriptive personality statements into a forced, approximately normal, 9-category distribution that ranges from *not at all characteristic* (1) to *highly characteristic* (9) of the person judged. The result is 100 scores that reflect the judged relevance of each item to the personality of the subject.

After completing the Q sort, each subject recruited two persons who knew him or her well to come to the laboratory and complete Q-sort descriptions of him or her. These "informants," who were also paid, typically were roommates or close friends. Care was taken to establish a comfortable rapport among subjects, informants, and experimenters and to provide assurances of the absolute confidentiality of all information provided. In particular, informants were told (truthfully) that their descriptions would not be made available to the acquaintances they described.

Funder (1980a) published an analysis of self–other agreement based on the first sample of self-ratings and peers' ratings, but did not examine interpeer agreement. The remaining data come from a study on another topic (Funder, 1982b) and from a study currently in progress and have not been previously published.

Results

Ratings of Trait Characteristics

Reliability. The first step in the data analysis was to construct reliable summary scores for each of the 100 Q items on the nine dimensions of interest. Within each of the nine dimensions, we calculated the correlation between each subject's ratings and the average ratings by all the other subjects. Following the practice of Rothbart and Park (1986), we eliminated any subject whose judgments correlated near zero (less than .10) with the average of the other judgments. This procedure led to the elimination of only 4 of 82 raters. The number of raters of

Table 3
Intercorrelations Among Trait Properties, With Favorability Partialed Out

	1	2	3	4	5	6	7	8	9	10	11
1. Ease of imagining confirming behaviors	—										
2. Ease of imagining disconfirming behaviors	.58	—									
3. Occasions for confirming behaviors	.70	.59	—								
4. Occasions for disconfirming behaviors	.60	.62	.74	—							
5. Number of instances to confirm	−.63	−.35	−.58	−.36	—						
6. Number of instances to disconfirm	.01	−.35	−.09	−.20	−.03	—					
7. Frequency of trait in population	.18	.10	.16	.01	−.21	.18	—				
8. Easy to judge	.79	.60	.74	.64	−.66	−.07	.21	—			
9. Self–other agreement	.24	.28	.33	.29	−.18	−.05	−.17	.27	—		
10. Interpeer agreement	.36	.19	.29	.19	−.30	−.06	−.15	.30	.52	—	
11. Total agreement	.34	.27	.35	.27	−.28	−.06	−.18	.32	—	—	—

Note. For correlations of .20 or greater, $p < .05$, two-tailed. For correlations of .26 or greater, $p < .01$, two-tailed. For correlations of .34 or greater, $p < .001$, two-tailed.

each dimension before and after elimination of unreliable raters and the alpha reliabilities of their average ratings before and after such elimination appear in Table 1. The alphas ranged from a high of .95 (favorability) to a low of .63 (number of instances needed to disconfirm), with a median of .80 (the median reliability obtained by Rothbart and Park was .81). The dimension easiness to judge, not included by Rothbart and Park, yielded an alpha reliability of .87. From Table 1 it can be seen that reliabilities were sufficient for present purposes, and retaining or eliminating the few unreliable raters made little difference.

Intercorrelations. The next step was to examine the intercorrelations of the various subjective ratings of the trait dimensions. Using the 78 reliable raters, we computed the average rating of each Q item on each of the nine dimensions. The intercorrelations between these dimensions as well as some other scores to be considered later appear in Table 2. Rothbart and Park (1986) found favorability to be highly correlated with the number of instances needed to confirm and disconfirm (in our study, these correlations were in the same direction, but much smaller). To correct for the possibility that some intercorrelations might be artifacts of social desirability, Rothbart and

Table 4
Correlations Between Five Factor Scores and Other Trait Characteristics

Variable	Frequency	Favorability	Self–other agreement	Interpeer agreement	Total agreement	Visibility
Factor scores						
Neuroticism	−.01	−.59	−.53	−.24	−.45	−.27
Extraversion	.23	.39	.29	.25	.31	.52
Openness	−.10	.14	.08	−.01	.05	−.11
Agreeableness	.27	.51	.07	.05	.07	.01
Conscientiousness	.03	.47	.21	−.05	.10	.01
Trait characteristics						
Frequency	—					
Favorability	.35	—				
Self–other agreement	.00	.43	—			
Interpeer agreement	−.10	.10	.51	—		
Total agreement	−.05	.31	—	—	—	
Visibility index	.25	.25	.39	.34	.42	—

Note. For all analyses, $N = 100$, which was the number of Q items in the set used. For correlations of .20 or greater, $p < .05$, two-tailed. For correlations of .26 or greater, $p < .01$. For correlations of .34 or greater, $p < .001$, two-tailed.

Table 5
Traits With Highest and Lowest Interjudge Agreement

No. Trait	Average total agreement	Self-other		Interpeer		
		Sample 1 (n = 41)	Sample 2 (n = 64)	Sample 1 (n = 37)	Sample 2 (n = 69)	Sample 3 (n = 50)
		Most agreement				
90. Is concerned with philosophical problems	.50***	.68***	.31*	.52***	.53***	.54***
84. Is cheerful	.43***	.42**	.36**	.35*	.45***	.58***
31. Regards self as physically attractive	.43***	.35*	.40**	.55***	.35**	.57***
28. Tends to arouse liking and acceptance	.41***	.50***	.41***	.45**	.31**	.40**
52. Behaves in an assertive fashion	.40***	.38*	.30*	.36*	.55***	.44**
4. Is a talkative individual	.40***	.40**	.33**	.38*	.37**	.59***
80. Interested in opposite sex	.40***	.39*	.21	.58***	.43***	.52***
62. Rebellious and nonconforming	.40***	.50***	.33**	.57***	.30*	.36**
33. Is calm, relaxed	.39***	.44*	.32**	.15	.50***	.48***
29. Turned to for advice and reassurance	.38***	.59***	.33**	.36*	.27*	.32*
91. Is power oriented	.37***	.40**	.22	.52***	.34**	.50***
81. Physically attractive	.36***	.36*	.28*	.51**	.39***	.36**
51. Values intellectual matters	.36***	.34*	.48***	.52***	.04	.42**
66. Enjoys aesthetic impressions	.36***	.46**	.20	.33*	.23	.64***
18. Initiates humor	.36***	.46**	.31*	.12	.40***	.41**
		Least agreement				
46. Engages in personal fantasy and daydreams	.05	.08	.00	.00	.16	.02
87. Interprets clear-cut situations in particularizing ways	.07	.18	.00	.00	.16	.02
89. Compares self to others	.07	.05	.00	.00	.25	.01
12. Tends to be self-defensive	.08	.02	.05	.19	.16	.02
34. Overreactive to minor frustrations	.08	.05	.00	.22	.15	.07
23. Extrapunitive; transfers and projects blame	.09	.10	.00	.22	.00	.25
10. Anxiety and tension produce bodily symptoms	.10	.01	.02	.28	.10	.19
69. Sensitive to demands	.10	−.22	.00	.15	.02	.43**
76. Projects own motives onto others	.10	−.10	.00	.36	.13	.15
13. Thin-skinned; sensitive to criticism	.11	.17	.00	.31	.09	.11
36. Subtly negativistic	.11	.16	.00	.15	.06	.30
77. Appears straightforward and candid	.11	−.05	.07	.19	.19	.16
50. Unpredictable and changeable	.11	−.06	.12	.22	.00	.31
30. Withdraws from adversity	.13*	.32*	.00	.26	.08	.10
70. Is ethically consistent	.13*	.10	.10	.14	.21	.11

* $p < .05$, two-tailed.
** $p < .01$, two-tailed.
*** $p < .001$, two-tailed.

Park calculated a second correlation matrix from which favorability was partialed. Following their practice, our intercorrelations with favorability partialed out appear in Table 3, but it can be seen that, as in the earlier study, such partialing made little difference to the general pattern.

In general, our findings replicated those of Rothbart and Park fairly well. Most of the dimensions were positively correlated with one another. The ease of imagining confirming and disconfirming behaviors and the frequency of occasions that allow confirming and disconfirming behaviors were intercorrelated particularly highly, with correlations even greater than those found by the earlier investigators (*Mdn* r = .60 in our study compared with r = .34 in the earlier study). The only notable failure to replicate Rothbart and Park concerns the dimensions of number of instances needed to confirm and disconfirm. Rothbart and Park found these two dimensions negatively correlated with each other (r = −.71) and correlated in generally opposite directions with the other dimensions. Although we also obtained a negative correlation between these two dimensions, the correlation was very small (r = −.06). Perhaps more importantly, we found both the number of behaviors needed to confirm and disconfirm to be negatively correlated with the other dimensions examined.

This complex pattern of intercorrelations can be simplified

Table 6
Traits Highest and Lowest in Judged "Easy Visibility" (Composite Index)

No. trait	Total visibility score	Actual agreement: Self–other agreement (df = 102)	Actual agreement: Interpeer agreement (df = 153)	Actual agreement: Total agreement (df = 254)
	Most visible			
43. Facially or gesturally expressive	36.93	.27**	.25**	.26***
81. Physically attractive	36.33	.31**	.41***	.36***
33. Is calm, relaxed	36.32	.37***	.42***	.39***
4. Is a talkative individual	35.50	.36***	.45***	.40***
56. Responds to humor	35.46	.37***	.16*	.27***
84. Is cheerful	35.25	.38***	.47***	.43***
92. Social poise and presence	34.73	.33***	.33***	.33***
52. Behaves in an assertive fashion	34.05	.33***	.47***	.40***
88. Personally charming	33.58	.26**	.25**	.26***
93. Sex typed	32.95	.34***	.26**	.30***
20. Has rapid personal tempo	32.90	.29**	.16*	.23**
80. Interested in opposite sex	32.05	.28**	.50***	.40***
77. Appears straightforward and candid	31.48	.04	.18*	.11
18. Initiates humor	31.27	.37***	.34***	.36***
28. Tends to arouse liking and acceptance	31.02	.45***	.37***	.41***
	Least visible			
22. Lack of personal meaning in life	9.79	.21*	.19*	.20**
46. Engages in personal fantasy and daydreams	10.66	.03	.08	.05
60. Has insights into own motives and behavior	10.95	.21*	.18*	.20**
44. Evaluates motives of others	11.99	.25*	.14	.19**
61. Creates and exploits dependency	12.59	.20*	.21**	.20**
16. Introspective	14.21	.18	.28***	.23**
47. Has readiness to feel guilt	14.74	.20*	.30***	.25***
36. Subtly negativistic	15.01	.06	.16*	.11
86. Repressive tendencies	15.10	.07	.25**	.16*
76. Projects own feelings onto others	15.31	.00	.19*	.10
9. Uncomfortable with uncertainties and complexities	15.66	.20*	.14	.17**
21. Arouses nurturant feelings	15.87	.08	.30***	.19**
75. Clear-cut consistent personality	16.05	.20*	.24**	.23**
72. Concerned with own adequacy	16.17	.12	.32***	.23**
39. Unconventional thought processes	16.41	.26**	.16	.21**

* $p < .05$, two-tailed.
** $p < .01$, two-tailed.
*** $p < .001$, two-tailed.

and illuminated by considering the dimension not examined by the earlier investigators: subjective easiness of judgment. It can be seen in Table 2 (and Table 3) that the subjective easiness with which a trait can be judged was highly correlated with how easy it is to imagine confirming and disconfirming behaviors, and how often occasions arise in daily life that allow confirming and disconfirming behaviors. The relevant intercorrelations ranged from .62 to .78. Moreover, subjective easiness was negatively correlated with the number of instances required to confirm and disconfirm the trait (rs = −.57 and −.11, respectively).

These findings reveal that seven out of the eight dimensions studied by Rothbart and Park are closely associated with the more simple dimension of how easy traits seem to judge. This association was particularly strong for five out of the six dimensions they presented as being theoretically important (i.e., excepting favorability and frequency). The median correlation between easiness and the first five dimensions in Table 2 was .67.

A principal-components analysis using a varimax rotation further clarified the relations among the judged properties of traits. As one would expect from their intercorrelations, the five properties just mentioned, plus subjective easiness, all loaded highly (between .65 and .90) on a first factor that accounted for almost half the total variance. All these properties were relevant to apparent observability and judgeability. Accordingly, the scores were combined (by using unit weights; "number of instances needed to confirm" loaded and was weighted negatively) into a composite index interpreted as revealing the apparent ease of observation or easy visibility of each Q item. The composite index has an alpha reliability of .90.

The next issue to be addressed was the relation between a trait's visibility and its relevance to the classic five factors of personality. The factor loadings calculated by McCrae et al. (1986) were correlated with the visibility index, and the results appear in the right-hand column of Table 4. Those traits load-

ing positively on Extraversion were particularly likely to seem easily visible (r = .52), whereas those traits defining Neuroticism were relatively unlikely to seem visible (r = −.27). The Q items that defined the Extraversion factor included "talkative," "arouses liking," "gregarious," and "socially poised." Neuroticism items included "thin-skinned," "basically anxious," and "irritable" (McCrae et al., 1986).

The findings reported thus far provide insight into what makes a trait *seem* visible and easy to judge. But an ultimately more important question still remains: Are the same traits that seem most easily visible *actually* easier to judge? This question can only be addressed by examining the accuracy with which traits are applied to real individuals.

Assessment of Interjudge Agreement

Within each of the two samples, the self-ratings on each of the 100 Q items were correlated with the mean ratings assigned by the two informants (or, in a few cases, one informant). The correlations were averaged, after being weighted by sample size and transformed via Fisher's z, into a composite index of self–other agreement. The two estimates of self–other agreement correlated with each other (r = .30), and their average yielded a set of scores that estimated self–other agreement with a reliability of .46.

Within each of the other three samples the agreement between the judgments on each Q item by the pairs of acquaintances was assessed via the intraclass correlation coefficient (Rosenthal & Rosnow, 1984). This is the coefficient that has to be calculated when the distinction between judges is arbitrary, and it can be considered the average of the correlations derived from all possible sets of X–Y pairings. It is interpreted in exactly the same way as a Pearson r, except that negative correlations are meaningless.

These correlations were averaged, again weighted by sample size and transformed via Fisher's z, to yield a composite index of interpeer agreement that had a reliability of .54. The index was averaged together with the index of self–other agreement, which was transformed by Fisher's z but not weighted by sample size (because each index was a best estimate of that kind of agreement and, thus, equally entitled to contribute to the composite of both kinds). This average, computed back into a correlation, was a general index of interjudge agreement, both self-peer and interpeer, that had a reliability of .68.

The 15 items with the highest general agreement and the 15 items with the lowest agreement are shown in Table 5 together with the agreement statistics derived from each of the five samples. An impressive amount of agreement was found, and those items with best agreement also manifested consistent agreement across independent samples. Overall, 87 out of the 100 Q items yielded significant interjudge agreement ($p < .05$, two-tailed).

Correlates of Interjudge Agreement

The final step in the data analysis was to correlate the three summary indexes of interjudge agreement with the various properties of the Q items. The complete results for each subjective dimension can be seen in the bottom three rows of Tables 2 and 3, and correlations involving the composite index of easy visibility appear in the bottom row of Table 4.

These intercorrelations revealed, first, that differences between traits in interjudge agreement were not very specific to whether the two judges being compared were the self and a peer, or two peers. The correlation between self–other and interpeer agreement was .51 ($p < .001$). Even more important, the traits that manifested the best agreement of both sorts tended to be the same ones that subjects regarded as most easily visible. This result is made concrete in Table 6, which displays the 15 traits with the highest and lowest composite visibility scores together with each trait's three composite agreement scores. The correlation between the composite-subjective-visibility score and the overall-composite-agreement score was .42 ($p < .001$).[2] When corrected for attenuation, this correlation increased to .54.

The relations between interjudge agreement and each trait's relevance to Neuroticism and Extraversion were as would be expected from the relation, shown earlier, between these factors and subjective visibility. Traits that loaded positively on Extraversion yielded relatively good interjudge agreement (r = .31, $p < .01$), and traits that loaded positively on Neuroticism yielded poorer interjudge agreement (r = −.45, $p < .001$). This result is congruent with the finding that subjects regard the traits that define Extraversion to be more visible than those that define Neuroticism.

Discussion

The findings of this study reveal that a trait seems most easily visible when (a) it is easy to imagine the behaviors that would confirm and disconfirm it, (b) there are many occasions that allow such behaviors, (c) only a few confirming behaviors are necessary to establish the trait, and (d) it seems subjectively easy to judge. There are also smaller tendencies for widely prevalent and more favorable traits to seem more visible.

Moreover, to a large extent subjective visibility is predictive of the actual accuracy with which a trait can be judged. This can be seen vividly by comparing the top half of Table 6, which displays the traits judged most easily visible, with the top half of Table 5, which displays the traits that are *actually* the most easily visible. Items appearing on both lists include "is calm, relaxed," "is cheerful," and "behaves in an assertive fashion." These traits seem relatively easy to observe and judge, and they are.[3] The other traits judged most visible also manifested, with one interesting exception, impressive interjudge agreement.

[2] Notice that favorability is also positively associated with most of these dimensions. However, that these correlations are not merely an artifact of favorability can be verified by a close examination of the partial correlations in Table 3. Partial correlations are not used in Table 4 because we believe that, in general, the favorability of an item is too fundamental an aspect of its meaning to be partialed out without distortion.

[3] These results are reminiscent of Kenrick and Stringfield's (1980) finding that the observability of a trait, assessed as the mean self-rating and others' rating of this property across individual subjects, was positively associated with interjudge agreement. The present results differ in that Kenrick and Stringfield focused on properties of *subjects* that made them harder and easier to judge; the present study is concerned with general properties of *traits* (also, see Cheek, 1982).

Subjects seemed to think that the degree to which someone appears straightforward and candid is easily visible; they were wrong.

The traits that seem the most difficult to observe similarly overlap with those that are, in fact, the most difficult; this includes *engages in personal fantasy and daydreams,* which obtained the lowest total agreement score and the second lowest subjective visibility score out of all 100 traits. Other traits that seem to be, and are, hard to judge include *is subtly negativistic: tends to undermine, obstruct, and sabotage* and *projects own motives and feelings onto others.*

At least two aspects of the content of traits appear relevant to both their subjective visibility and actual judgeability. Traits that load positively on Extraversion in the classic five-factor solution (McCrae et al., 1986), such as gregarious, generally refer to overt patterns of social behavior and so seem easily visible and are judged with greater accuracy. The traits defining the Neuroticism factor, such as *concerned with own adequacy as a person,* refer more often to intrapsychic states that must be inferred rather than directly observed. Such traits seem less easily visible and actually are judged with less accuracy by observers from daily life. The same traits that load positively on Extraversion tend to load negatively on Neuroticism and vice versa ($r = -.32$, $p < .001$, in McCrae et al.'s data), so the same principle may underlie both findings: Extraverted (and nonneurotic) personality characteristics are revealed by social behaviors that, by definition, are on public display; neurotic (and introverted) characteristics are relevant to private activities that are less observable.

Despite this fundamental difference between traits, lay judges agree well across a wide range of the personality domain. The very best items in Table 5 refer to general patterns of behavior, not single instances, and many require analysis of underlying motivations and even insight on the part of the judge. Some of the best agreement is found on traits such as *concerned with philosophical problems, power oriented,* and *genuinely values intellectual matters.*

Still, some psychologists would not be impressed by the interjudge agreement demonstrated in Table 5. The largest average correlation is .50, and as a group the best traits seem to yield agreement scores that range from about .40 to .30. However, two facts should be borne in mind. The first is that it is no longer tenable to believe that some traits just happen to yield higher agreement scores than do others. The stability of the higher correlations across different samples is impressively demonstrated by Table 5, and even more importantly, we have seen that independently specifiable properties of different traits go a long way toward predicting and explaining which traits will yield the best agreement. The traits at the top of Table 5 are the ones that are judged most accurately; they were not merely "lucky" in one particular sample.

The second fact is one that bears repeating in a variety of research contexts: Correlations in the range of .30 to .40 are larger than has traditionally been recognized. Although over the years it has become commonplace to square correlations such as these, to claim that they account for 9% or 16% of the variance, and to dismiss them as unimportant, a variety of empirical and statistical considerations are increasingly leading psychologists to acknowledge that such correlations reveal relations of important magnitude. Funder and Ozer (1983) have shown how several of the most important effects in the literature of experimental social psychology are of a size between .30 and .40. Rosenthal and Rubin (1982) demonstrated mathematically that a correlation of .40 will lead to a correct, dichotomous decision 70% of the time. In other words, a prediction based on a correlation of .40 will be correct more than twice as often as it is wrong. Finally, Ozer (1985) claimed that the widespread practice of squaring correlations is inappropriate in the first place. According to Ozer's analysis, when two judgments are both considered representatives of the same, unmeasured, latent variable, a correlation between them of .40 means that variable "accounts for" 40%, not 16%, of their total variance (cf. Tryon, 1929).

Against a background of contemporary research on social judgment that usually focuses on error and shortcomings, the conclusions of the present study seem remarkably optimistic. Different judges of the same personality, including the person in question, tend to agree with one another to an impressive degree on a wide variety of personality attributes. With this finding, the present study confirms and extends the results of studies such as Andersen (1984), Cheek (1982), Edwards and Klockars (1981), Funder (1980b), Goldberg, Norman, and Schwartz (1980), Hase and Goldberg (1967), McCrae (1982), Monson, Tanke, and Lund (1980), Paunonen and Jackson (1985), and Woodruffe (1985)

The other key finding of this study is more novel and, perhaps, more important. We are not accurate in our judgments of personality merely sometimes or when favored by chance. Rather, we are most likely to render accurate assessments of just those traits where the judgment seems easy.

References

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

Andersen, S. (1984). Self-knowledge and social inference: II. The diagnosticity of cognitive/affective and behavioral data. *Journal of Personality and Social Psychology, 46,* 294–307.

Anderson, N. H. (1968). Likeableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology, 9,* 272–279.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81,* 506–520.

Bem, D. J., & Funder, D. C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. *Psychological Review, 85,* 485–501.

Block, J. (1978). *The Q-sort method in personality assessment and psychiatric research.* Palo Alto, CA: Consulting Psychologists Press. (Original work published 1961)

Campbell, J. B. (1985, April). *Cross-situational consistency of personality attributes.* Paper presented at the meeting of the Eastern Psychological Association, Boston.

Chaplin, W. F., & Goldberg, L. R. (1984). A failure to replicate the Bem and Allen study of individual differences in cross-situational consistency. *Journal of Personality and Social Psychology, 47,* 1074–1090.

Cheek, J. M. (1982). Aggregation, moderator variables, and the validity of personality tests: A peer-rating study. *Journal of Personality and Social Psychology, 43,* 1254–1269.

Christensen-Szalanski, J. J. J., & Beach, L. R. (1984). The citation bias: Fad and fashion in the judgment and decision literature. *American Psychologist, 39,* 75–78.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of

others" and "assumed similarity." *Psychological Bulletin, 52,* 177–193.

Edwards, A. L., & Klockars, A. J. (1981). Significant others and self-evaluation: Relationships between perceived and actual evaluations. *Personality and Social Psychology Bulletin, 7,* 244–251.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85,* 395–416.

Epstein, S., & Teraspulsky, L. (1986). The perception of cross-situational consistency. *Journal of Personality and Social Psychology, 50.*

Estes, S. G. (1938). Judging personality from expressive behavior. *Journal of Abnormal and Social Psychology, 33,* 217–236.

Fischoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564.

Funder, D. C. (1980a). On seeing ourselves as others see us: Self–other agreement and discrepancy in personality ratings. *Journal of Personality, 48,* 473–493.

Funder, D. C. (1980b). The "trait" of ascribing traits: Individual differences in the tendency to trait ascription. *Journal of Research in Personality, 14,* 376–385.

Funder, D. C. (1982a). On the accuracy of dispositional versus situational attributions. *Social Cognition, 1,* 205–222.

Funder, D. C. (1982b). On assessing social psychological theories through the study of individual differences: Template matching and forced compliance. *Journal of Personality and Social Psychology, 43,* 100–110.

Funder, D. C. (1983). The "consistency" controversy and the accuracy of personality judgments. *Journal of Personality, 51,* 346–359.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75–90.

Funder, D. C., & Harris, M. J. (1986). On the several facets of personality assessment: The case of social acuity. *Journal of Personality, 54,* 528–550.

Funder, D. C., & Ozer, D. J. (1983). Behavior as a function of the situation. *Journal of Personality and Social Psychology, 44,* 107–112.

Funder, D. C., & Van Ness, M. J. (1983). On the nature and accuracy of attributions that change over time. *Journal of Personality, 51,* 17–33.

Goldberg, L. R. (1959). The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender–Gestalt Test. *Journal of Consulting Psychology, 23,* 25–33.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in the personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141–165). Beverly Hills, CA: Sage.

Goldberg, L. R., Norman, W. T., & Schwartz, E. (1980). The comparative validity of questionnaire data (16PF scales) and objective test data (O–A battery) in predicting five peer-rating criteria. *Applied Psychological Measurement, 4,* 183–194.

Guilford, J. P. (1936). *Psychometric methods.* New York: McGraw-Hill.

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin, 67,* 231–248.

Hastorf, A. H., & Bender, I. E. (1952). A caution respecting the measurement of empathic ability. *Journal of Abnormal and Social Psychology, 47,* 574–576.

Kahneman, A., & Tversky, D. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Katz, D., & Braly, K. W. (1933). Racial stereotypes of one hundred college students. *Journal of Abnormal and Social Psychology, 28,* 280–290.

Kenrick, D. T., & Stringfield, D. O. (1980). Personality traits and the eye of the beholder: Crossing some traditional philosophical boundaries in the search for consistency in all of the people. *Psychological Review, 87,* 88–104.

Loftus, E. F., & Beach, L. R. (1982). Human inference and judgment: Is the glass half empty or half full? *Stanford Law Review, 34,* 939–956.

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., Jr., & Busch, C. M. (1986). Evaluating comprehensiveness in personality systems: The California Q-Set and the five factor model. *Journal of Personality, 54,* 430–446.

Mischel, W. (1984). Convergences and challenges in the search for consistency. *American Psychologist, 39,* 351–364.

Monson, T. C., Tanke, E. D., & Lund, J. (1980). Determinants of social perception in a naturalistic setting. *Journal of Research in Personality, 14,* 104–120.

Moskowitz, D. S. (1986). Comparison of self-reports, reports by knowledgeable informants, and behavioral observation data. *Journal of Personality, 54,* 294–317.

Moskowitz, D. S., & Schwarz, J. C. (1982). Validity comparison of behavior counts and ratings by knowledgeable informants. *Journal of Personality and Social Psychology, 42,* 518–528.

Murray, H. A. (1938). *Explorations in personality.* New York: Oxford University Press.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* New York: Prentice-Hall.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66,* 574–583.

Norman, W. T. (1967). *2800 personality trait descriptors: Normative operating characteristics for a university population.* Unpublished manuscript, University of Michigan, Ann Arbor.

Oskamp, S. (1965). Overconfidence in cast-study judgments. *Journal of Consulting Psychology, 23,* 23–33.

Ozer, D. J. (1979, September). *Observer evaluations of personality: 1918–1937.* Paper presented at the meeting of the American Psychological Association, New York.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307–315.

Paunonen, S. V., & Jackson, D. N. (1985). Idiographic measurement strategies for personality and prediction: Some unredeemed promissory notes. *Psychological Review, 92,* 486–511.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis.* New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169.

Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology, 50,* 131–142.

Schneider, D. J., Hastorf, A. H., & Ellsworth, P. C. (1979). *Person perception* (2nd ed.). Reading, MA: Addison-Wesley.

Tryon, R. C. (1929). The interpretation of the correlation coefficient. *Psychological Review, 36,* 419–445.

Wolf, R., & Murray, H. A. (1936). An experiment in judging personalities. *Journal of Psychology, 3,* 345–365.

Woodruffe, C. (1985). Consensual validation of personality traits: Additional evidence and individual differences. *Journal of Personality and Social Psychology, 48,* 1240–1252.

Received December 27, 1985
Revision received May 9, 1986 ■