Who should own the definition of personality?*

WILLEM K. B. HOFSTEE‡
University of Groningen, The Netherlands

Abstract

The averaged judgment of knowledgeable others provides the best available point of reference both for the definition of personality structure in general and for assessing someone's personality in particular. Self-judgments, as in personality questionnaires, are intrinsically deficient because judgment errors cannot be averaged out. The recommended procedure for assessing someone's personality is to give a personality questionnaire, phrased in the third person singular, to those who know the target best. This set may or may not include the target person as a judge.

INTRODUCTION

During the months preceding the elections for the Dutch parliament in 1989, we organized an election bet (Hofstee and Schiapman, 1990). Readers of a regional advertisement paper were invited to submit their predictions of the new composition of the parliament; a prize was put up for the closest forecasts. We announced to the readers that we would average their predictions, and that we expected this average to beat the opinion polls. Thus, a contest with the pollsters was superimposed upon the mutual contests among the respondents. As a criterion, we took the sum of the absolute discrepancies between the numbers of predicted and obtained seats, over the parties. According to this criterion, the averaged prediction of the participants was indeed superior to the pollsters' (for further details and considerations, see Hofstee and Schiapman, 1990). This finding was replicated for the 1994 elections.

The analogy between this demonstration experiment and the study of personality hinges upon the concept of self-presentation. The pollster asks the familiar question 'How would you vote if elections were held today?' The answer is likely to be determined by the impression you wish to make upon, or the message you wish to convey to, the pollsters—rather than a faithful introspection leading to a detached self-

*This paper contains the presidential address of the 7th Conference of the European Association for Personality Psychology (EAPP), Madrid, July 12–16, 1994.
‡All correspondence concerning this paper should be addressed to Willem K. B. Hofstee, University of Groningen, The Heymans Institute, Grote Kruisstraat 2-4, 9712 TS, Groningen, The Netherlands.

CCC 0890–2070/94/030149–14
© 1994 by John Wiley & Sons, Ltd.

Accepted 13 June 1994
prediction of your voting behaviour. With respect to personality questionnaires, Hogan (1994) argues that ‘Responding to an item is as much a social act as responding to a question from your mother, and its meaning should be understood in terms of the actor’s typical interpersonal goals rather than a hypothesized ... neuropsychic entity...’ (see also Johnson, 1994).

In contrast, the participants in our election bet tried to predict voting behaviour. No doubt one or two did not play along and gave in to wishful thinking or a counter-wishful strategy (trying to win the prize if their worst fears came true), but the majority had apparently tuned in on the predictive wavelength, which is precisely where we wanted them to be.

The thesis of this paper is that judgments of the average other person provide the best available point of reference for the definition of personality—both for the definition of personality in general and for the assessment of your or my personality: if you want to know what you really are like, ask a number of others.

The present approach should not be confused with social constructivism and related conceptions. I emphatically do not wish to argue that personality is in the eye of the beholder. As a student of personality, I subscribe to Funder’s (1991) dictum that traits are real. Society has commissioned to personality psychologists the task of finding out what people really are like. The study of social perceptions or narratives, in this context, is at best an auxiliary enterprise that makes us aware of potential distortions in the study of personality through judgments; at worst, these approaches deflect the study of personality from its mission.

Consequently, the title question of this paper, ‘Who should own the definition of personality?’, should be taken with a grain of salt. I concur with McCrae (1993) that ‘It is reality that “owns” the truth about personality, not any single operationalization.’ Being mortals, we do not have direct access to reality, so the best we can do is beat around the bush. In more official terminology, the study of personality should employ a construct-validation strategy (Funder, 1994), pitting methods against each other (self and peer reports, tests, behavioural observations, physiological measures), rather than taking one approach as its Archimedic point.

However, there are good reasons to focus upon a comparison of self and others’ reports. Self-reports in the form of personality questionnaires occupy the leading position in personality research and application. Physiological measurements—for which questionnaire scores are usually taken as criteria—and projective tests cannot serve as credible competitors. Behaviour observations (e.g. Funder and Sneed, 1993) may seem to constitute a serious alternative, but they presuppose an observer, so they are a special case of others’ reports. One might perform a comparative cost-benefit analysis of behaviour observation versus acquaintance ratings, which will probably come out in favour of the latter; but that comparison is outside the scope of this paper.

A central ingredient in the plea for others’ judgments as the best single measure of personality is the principle of aggregation. Over 80 years ago, Spearman (1910) and Brown (1910) independently published what has become known as the Spearman–Brown formula for reliability as a function of test length. A frequent application of the formula is to estimate the reliability of an averaged judgment as a function of the number of judges, thus taking judges as variables, instead of tests. An implication in terms of true-score theory is that a judgment is conceived as the sum of a true component (which is identical across judges for a particular target object
or person) and an error component, which is uncorrelated with the true component and with the error components of other judges.

Upon viewing human judgment from this psychometric perspective, the most striking overall finding is its unreliability: idiosyncrasy looms large. For example, in a study of peer judgments of grant proposals for scientific research (Hofstee, 1983) the average correlation between two independent judges appeared to be only 0.15; similar findings were presented by others (e.g. Scheerens and Beem, 1986). In this domain, therefore, the error component of individual judgments seems to cover some 85 per cent of the variance. In all fairness, it should be realized that there is much restriction of range in these selection situations, due to self-censoring, pre-screening, and the like. In my experience on such panels, clearly inferior grant applications were usually rejected unanimously, but any impression of large overall agreement among panel members is incorrect.

The example of peer review also serves to highlight the other side of the judgmental coin, which is validity. With respect to scientific plans and products, such as grant applications and journal submissions, few scientists would seriously propose replacing peer review with some objective criterion: even those psychometricians who have advocated doing away with human judgment in clinical diagnosis, personnel selection, or educational grading would have been unpleasantly surprised if their own manuscripts had been turned down on the basis of some algorithm. The obvious explanation of this self-excepting attitude is that scientists and practitioners alike, each in their own field, assume that human judgment represents the best criterion, fallible though those judgments may be.

This is not the place to go deeply into the question of whether the true quality of a manuscript or proposal, an applicant for a job, or an essay examination should be conceived as the average judgment of a population of competent judges. Rather than pursuing this general theme (see Hofstee, 1993), I shall focus on the question of whether it makes sense to define personality in terms of an average judgment. First, I review the many ways in which our conception of personality structure is based on averaged judgments. In the second and main part of this paper, I compare the relative merits of self and others' reports for assessing individuals' personalities.

THE DEFINITION OF PERSONALITY TRAITS

What do we mean by personality in general? Every teacher and textbook writer faces that question. Sooner or later, it translates into defining personality traits, for even if there is more to personality than just traits, no definition can do without an explicit or implicit reference to them. The definition of personality traits may be given in an abstract or 'intensional' manner. The alternative is an enumeration or 'extensional' definition. Both approaches have their strengths and weaknesses; probably the best strategy consists of an interplay between the two. However, the focus here will be on the enumerative or extensional definition.

The foremost example of enumeration in personality research ensues from the lexical approach. More than a century ago, Sir Francis Galton (1884) may have been first to scan a dictionary for trait terms. Just a little later, Rümelin (1890) nicely articulated the lexical motive: '... wir dürfen wohl mit Recht davon ausgehen, dass das die einfachsten, der Beobachtung zuerst und am häufigsten sich aufdrängen-
den Begriffen sind, welche die Sprache durch ein besonderes Wort auszuzeichnen
ein Bedürfnis empfindet’ (‘We may proceed from the assumption that language
experiences a need to express in a particular word those concepts that are the simplest,
that soonest and most often impose themselves upon the beholder’) (p. 397). The
agent in Rümelin’s version of the lexical hypothesis is ‘die Sprache’, or in more
contemporary terms, the language community. Thus, in the lexical tradition the
first demarcation of the concept of personality is provided by the sediment of a
judgmental process.

The selection of trait terms from the lexicon requires a second step. Even within
a word class such as adjectives, the selection is by no means automatic; the overlap
between two independent selections tends to be only some 50 per cent (Hofstee,
1990). Moreover, it is impossible to define objective criteria to decide, for example,
whether Conservative is merely an attitude or a trait, Merry merely a mood, Charming
merely a social effect, or Brilliant merely an evaluation. Again an average judgment
is needed, with a sufficient number of judges to secure a replicable result. One could
argue that the judges should be experts on personality, but that is a moot point;
empirically, lay and expert selections would probably be virtually indistinguishable.
To the extent that a particular expert’s selection—yours or mine—deviates systemati-
cally from the average lay person’s, the chances are that our idiosyncrasies are not
shared by fellow experts.

Averaged lay judgments subsequently play a dominant role in the structuring
of the trait domain. Inside and outside the lexical tradition, scientific decisions on
trait structure rely almost exclusively on judgmental data. Thus, decisions on trait
covariance are dictated by some average of people’s implicit personality theories,
as evidenced in such judgments. Individual investigators’ ideas on what goes together
with what would have a hard time competing with this average.

In summary, agreement can probably be reached on the statement that the definition
of personality structure is highly reflective of the average person’s implicit ideas.
What is scientific about it is procedural rather than substantive: the designs for
data gathering and processing are such that idiosyncrasies are filtered out. The sub-
stance of the definition, however, is provided by the average lay person. It may
be countered that my analysis relies heavily on the lexical approach to personality.
However, the Big Five paradigm, which epitomizes this approach, is broadly recog-
nized as an important point of reference for the definition of personality—whether
grudgingly (Pervin, 1994) or enthusiastically (e.g. Digman, 1991). Moreover, the
Big Five breakthrough came about precisely because the five-dimensional structure
in adjective ratings was recovered in personality questionnaires (see, e.g. Amelang
and Borkenau, 1982), which constitute the sediment of scientific conceptions of per-
sonality.

All this is not to say that we call upon the average person to provide theories.
On the contrary, we ask for a focus on concrete target persons. We do not try
to average abstract definitions of personality—if that would make sense at all. We
do not ask for direct or ‘internal’ (Wiggins, 1973, p. 336) judgments of trait covari-
tion, but infer trait structure from single ‘external’ ratings of actual persons. The
programme is to have judges focus on the reality out there, not on any stereotypes
inside their heads.

How should one evaluate this ‘bookkeeping’ approach to personality? Here, predi-
lections clearly diverge. There is the familiar argument that chemistry would not
have outgrown its infancy if chemists had clung to the natural language. On the
other hand, one may be suspicious of behavioural-scientific attempts to define realities
dissociated from real life. To me the issue seems to be a matter of strategy rather
than principle. I have outgrown the stage of generalized mistrust of common sense
and clinical judgment; I still think that individual judges are wrong most of the
time, but that is quite different from declaring the common component of their
judgments to be fundamentally mistaken.

The bookkeeping definition of personality should probably function as a stepping
stone. At the present stage, it would be clearly premature to replace the judgmental
definition of personality by some set of objective variables, for example, physiological
measures. In view of the low correspondence between the two sets, that would be
an unwise strategy. On the other hand, it is indeed unlikely in the long run that
the most powerful and efficient set of basic concepts of a branch of science can
be picked off the street. For example, it is hard to believe that varimax rotation
of principal components derived from modest sample sizes have given us the once-
and-for-all basic dimensions of personality, to be corroborated by future discoveries
of genetics. In summary, then, the consensual definition of personality should function
as something to be improved upon, not as something to be worshipped or
denounced.

THE PRIMACY OF OTHERS’ JUDGMENTS

Let us now consider people’s personalities: your personality and mine. How does
one assess personality? I shall argue that an average taken over the judgments of
a sufficient number of knowledgeable others will do the job. I should emphasize
in advance that I do not mean direct trait ratings, but ratings on a well constructed
questionnaire administered in the third person singular. The most obvious alternative
is a personality questionnaire administered in the first person singular. The argument
in favour of others’ judgments has many aspects, ranging from psychometric to
qualitative.

The psychometric argument

Eighty-four years after Spearman and Brown, the psychometric argument is unpro-
blematic. There is only one me, whereas there are many others who know enough
about me to provide a more reliable average judgment. Thus, other things being
equal, I am outnumbered and outperformed by the average other.

Personality questionnaires filled out by the target person are comparable to a
one-item test, or some minimal version of a matrix sampling design in which each
target person is rated by a different assessor. The implicit assumption seems to be
that the assessor is perfectly exchangeable and reliable. Let us examine that assump-
tion.

If a score on a personality questionnaire is considered as a piece of verbal behaviour,
its reliability may be gauged to be almost perfect. The appropriate thought experiment
is to have a person fill out the questionnaire, produce amnesia for the task by hypnosis,
and give a retest. Under these conditions, reliabilities in the upper nineties might
reasonably be expected.

However, we are not legitimately interested in verbal behaviour as such, but in
a person’s true score on a trait. Between the observed score and the trait score, a number of sources of error can be pinpointed. The first is idiosyncratic interpretations of the item. Johnson (1994) gives the example of the adjective Thoughtful, which may be conceived as Considerate by one respondent and Contemplative by another. Cattell (1946) spoke more bluntly about a ‘... lack of understanding, in the least intelligent quarter of the population, of what the questions and the words mean’ (pp. 342–343). Specification of items (as recommended, e.g. by Cattell, 1957; see also Goldberg and Kilkowski, 1985) diminishes the problem, but does not eliminate it. Also, the requirement of relevance imposes a limit upon the extent to which specification is feasible.

The second source of error is differential response bias: response distributions of judges differ in central tendency, dispersion, skewness, and so on. Judicious within-judge centring, standardization, or normalization of response distribution, meets the problem to some extent, but not completely, as the number of available judgments is finite. The third source is limited availability of cues due to limitations of the person’s perspective and memory. Last but not least, questionnaire responses suffer from self-presentational biases ranging from self-enhancement to self-diminishment (John and Robins, 1994). In summary, questionnaire responses are fraught with error. Replicating a self-rating would not help, for within a judge, the errors are systematic and thus will not cancel out.

Can we make a quantitative estimate of the true trait variance in questionnaire scores? As we cannot meaningfully carry out a reliability study, the estimate has to be based on their validity. The highest validities of questionnaire scores are with ratings of close acquaintances or spouses on a third-person version of the same instrument (see e.g. Borkenau and Liebler, 1993; Costa and McCrae, 1988; McCrae and Costa, 1987; Marsh and Byrne, 1993); the observed validities are of the order of 0.5–0.6. These values are inflated in at least three ways. First, the judgments are contaminated by communication prior to, if not upon, responding to the questionnaire. Second, specific (that is, unrelated to the trait) item variance contributes to the correlation as the instruments are the same. Third, both parties usually fill out the form at the same time, so temporal instability is not taken into consideration. Taking these sources of error into account, the proportion of trait variance that self and others’ ratings have in common should be about 0.4.

If self and other are exchangeable instruments for assessing personality, 0.4 is the best estimate of the proportion of trait variance in either of the two; in other words, the proportion of trait variance in personality questionnaires is of that order. A panel of six knowledgeable judges would be needed to attain a value of 0.8. As self is by definition single, there is no such hope for personality questionnaires in the first person singular.

Qualitative arguments in favour of self-reports

The first line of defence for self-reports consists of proclaiming their intrinsic superiority. For example, assuming that self-reports are twice as valid as others’ reports, the proportions of trait variance would be 0.56 and 0.28, respectively. Four others would be needed to achieve the same validity as the self-report, so cost-benefit considerations would enter into the discussion. However, let us examine such claims at some length.
To set the tone for the argument, imagine being asked whether you are a Modest person. From your own perspective, you simply cannot answer that question as your response would be paradoxical: if you would say ‘I am a modest person’, the fact that you say so entails an important immodesty; if you would say ‘No, I am not a modest person’, that would be disclaiming a desirable trait, which is what modesty is about (Hofstee, 1990). But the paradox disappears like snow before the sun if you bring yourself to taking the other person’s point of view. You can very well compare other people on Modesty, and if you succeed in taking a good hard look at yourself as an element of that set, there is no longer anything paradoxical about finding yourself relatively Modest or Inmodest. In fact, many people are even insensitive to the paradoxicality of answering such a question; that can only mean that they automatically adopt the other person’s position; so potent is the primacy of the external perspective on personality.

**Availability of cues**

Another argument is that the target person himself or herself has witnessed all the episodes that are relevant for judging his or her own personality, and no one else has. First, however, we know that the saliency and relevance of such episodes for judging personality tend to be played down in the actor’s perspective, as distinguished from the observer’s (Jones and Nisbett, 1971). Second, most of the richness of self-experience is wasted because of the limitations of the judge’s memory. Third, the superiority argument can only apply to one source of error, namely, cue sampling. With respect to two other sources, namely response bias and idiosyncratic interpretation of items, the self-rater is not in any special position. Self-presentational bias is specific to self-judgments.

**Privileged access**

Yet another argument in favour of self-report is privileged access: only the person himself or herself is in direct touch with his or her own feelings and emotions, motives and desires, attitudes and values, plans and projects, and so on. I do not intend to deny that people themselves know best about their secrets, their toothaches, or their food preferences. But secrets, pains, likes and dislikes are not central to the definition of personality. Nor are temporary states or other transient phenomena.

There is a more fundamental argument against a subjective definition of personality. The psychology of personality is a branch of science, and science aims at agreement. If we would emphasize subjective aspects, we would withdraw personality from scientific study. In a scientific context, personality is by definition a public phenomenon. The scientific study of pure subjectiveness is a contradiction in terms.

The spatial metaphor of privileged access pictures the person as an entity inside a black box constituted by an impenetrable skin; only I myself am capable of sinking some light into the glob of subjectiveness down there. Not only is this picture philosophically suspect, but also it is contradicted by many everyday instances in which the other person knows better what my mood is or what I am up to — by the criterion that I readily substitute his or her interpretation for mine.

To underline the privileged position of self-reports, McCrae (1993) gives an example in which ‘... we might be able to document that an individual had attended 2,456 parties in the last five years — yet he might rightfully claim on a self-report inventory that he was not extraverted. He would be correct, and an infinite number
of judges incorrect, if he did not enjoy going to parties’. However, there are two possibilities. The most likely one is that this person’s lack of enjoyment was visible. In that case, with an average of only four judges present per party, there have been 10 000 opportunities for observing his lack of extraversion, and the chances that on average he would be judged extraverted would be infinitesimal. In the other more interesting case, he would have systematically succeeded in convincing others that he was enjoying himself. Note that this is difficult to accomplish if you are not. He might still say, for some self-presentational reason, that he disliked parties, but by all reasonable standards the credibility of that self-report would be zero. He might be less extraverted in other situations, but that is not at issue here.

**Inner and outer traits**

In a stepped-down version of the privileged-access argument, a distinction is made between inner traits such as anxiety, and outer traits such as charm (Johnson, 1994). The superiority of self-report would be restricted to inner traits.

A series of studies by John and Robins (1993) leaves no doubt that self–other correlations are lower with less observable traits. But does that mean that self-report assesses trait variance unnoticed by the outside observer? A more parsimonious explanation is psychometric attenuation: both self-report and observation are less reliable with traits that are difficult to observe. John and Robins’ results strongly suggest that attenuation is operative: the pattern of self–peer and peer–peer correlations is very similar across traits. Interestingly, the trait *Emotional*, which would exemplify the privileged-access argument, has both the highest self–peer and the highest peer–peer correlation. John and Robins speculate that self and peer perception proceed through similar cognitive processes.

Another pattern of findings by John and Robins (1993) deals a heavy blow to the privileged-access tenet. First, agreement between judges is higher for neutral traits than for evaluative traits, that is, traits with high or low social desirability. Second, this effect is stronger for the agreement between self and peer than for peer–peer agreement. Third, self–peer agreement is systematically lower than peer–peer agreement, except with neutral traits. Together, these findings strongly suggest the following conclusion: under optimal circumstances—in the absence of self-presentational temptations—self-reports are at best exchangeable with single peer reports; however, with clearly desirable or undesirable items, self-report had better be covered with the cloak of charity. If you want to know what you really are like, you’d better ask a number of others.

**Catastrophic events**

Goldberg (1993), Pervin (1994), and Van den Hout (1993) have pointed to the insufficiency of the other person’s perspective in predicting dramatic acts of violence, such as suicide. This argument may be taken as a special case of the privileged-access argument, but it also introduces a separate topic: should personality psychology try to predict events?

I emphatically think not (see also Hofstee, 1993). Part of the mission of science is to rule out certain problems—like the quadrature of the circle, or the perpetuum mobile—rather than to keep trying to solve them. Concrete one-time events like the weather on my sixtieth birthday, the exchange rate of the dollar on the first of January in the year 2000, or the crumbling of the Berlin wall, are subject to
the butterfly effect of chaos theory. Some questions should be left to fortune-tellers, in our case, astrologists, cheirologists, and the like—or, more rationally, they should be left alone.

The distinction here is between dispositions and events. I fully acknowledge that persons may have a violent or a suicidal disposition, and that the study of such dispositions is scientifically and practically important. But with respect to dispositions, there is no merit in adopting the subjective point of view, as I have argued above.

**Flaws in others’ reports**

Another line of defence for self-reports consists of an attack on the competition. Agreement among others may be influenced by communication about the target person and by shared stereotypes; reporters may be biased about the person, especially if they are selected by that person; and others may have a limited perspective on the person due to their social relationships.

**Gossip**

First, if two or more others know the target person well enough to be eligible as judges, the chances are that they have been gossiping about that target’s personality. Therefore, part of their agreement may be spurious. However, the reverse may also be the case. Gossip may serve to articulate any dissensus about the person, spuriously lowering the agreement coefficient. Moreover, prior discussion need not make a judgment less valid. Finally, the argument is not specific: communication may also have taken place between the target person and others. Statistically speaking, self-other communication is even more likely, as there is no requirement that the others even know each other.

A way to study the issue is to register the amount of prior contact between judges of the same target person, including that person. That amount may range from none at all, as between acquaintances from different settings, to daily contact, as between spouses. Agreement between judges may be studied as a function of their mutual contact.

**Shared stereotypes**

Next, spurious agreement may arise through shared stereotypes concerning sex, age, race, physical characteristics, and the like. However, that possibility is largely academic if the set of judges is restricted to those who know the target person well. Stereotypes function as default options or prior probabilities that are largely overruled when we get to know a person better; denial of that mechanism would rest upon the assumption that traits do not exist, so that the perceiver would have no basis for a judgment other than stereotypes. Also, any risk of stereotyping is diminished if judges are asked about the target person in concrete terms (e.g. ‘usually turns up in time’, ‘talks a lot’, and the like).

Moreover, stereotypic variance is not necessarily invalid. One may invoke complicated social–psychological explanations for the fact that, on average, women are judged more *Emotional* than men, also by women. The explanation is that the difference merely reflects an (invalid) male stereotype, which has become internalized by women. We are not supposed to ask why men choose to find women "Emotional"
rather than, for example, Unemotional; nor should we inquire by what process women collectively arrive at internalizing false ideas about themselves. Mainly, we are not supposed to entertain the pedestrian view that, on average, women are somewhat more Emotional. Nonetheless, that is by far the most parsimonious explanation of the facts.

Pushing this a little further, are we really sure that associations between temperamental and physical characteristics in people are nothing but false stereotypes? If so, where did they come from? Can we rule out the possibility of any genetic linkage? I do not think that we have the least shred of evidence that we can. I do think that we have a problem of quite a different order: all individuals should be treated as equal. In my conviction, personality psychologists should even be among the first to subscribe to this ethical individualism. But we should not therefore deceive ourselves and our public by promoting explanations that may well prove to be false. Our assignment is to face the facts and put them at the disposal of the individual.

Social desirability bias

Restricting the set of judges to those who know the target person well enough may introduce social desirability bias: it would be difficult to find the kind of saint, who have tried to know and understand people they do not like. Moreover, in practice the judges are usually selected by the target person, so a derived self-presentation effect should be expected.

Note, first, that the bias is not as differential as it may be in self-report. John and Robins (1994) found evidence of substantial individual differences in self-enhancement versus self-diminishment, albeit in a laboratory setting. A complementary effect may be expected in judging others; in fact, the two tendencies were indistinguishable in the study of John and Robins. However, with several judges of a target these individual differences among judges should cancel out to some extent, which they cannot in self-report.

Second, the judges are not involved in self-presentation, even though they may have been assigned a role in that scenario. The difference is illustrated in the study of John and Robins (1993), in which agreement among peers was less sensitive to the social desirability of the item than self–peer agreement. The use of more factual and less inferential items may be expected to further reduce the problem of social desirability bias.

Partial perspectives

Finally, the basis for others' judgments of personality is often said to be limited, as they know the target only in certain roles and settings. Exceptional cases, exemplified by Woody Allen's chameleonic character Zelig, serve to illustrate a dramatic conceptions of personality, in which traits are purely interactive. But reality tends to be much more platitudinous. Our personality transpires through different settings and relations, as is illustrated by the agreement between judges. There is no doubt that different others know different facts about a person; but personality is about general trends, not about specific anecdotes.

A more productive way of phrasing the problem is to ask how much an additional judge adds to the trait variance. The general answer is contained in the signal-to-noise ratio, which is positive if the internal-consistency coefficient of ratings goes up upon adding that judge. We may try to predict judges' signal-to-noise ratios as a function
of class membership (self, spouse, family, friend, colleague, acquaintance, etc.), length and intensity of acquaintance, measures of their perceptiveness, and so on. In doing so, we should take into account the diminishing returns of added judges, in accordance with the Spearman–Brown function.

Identity or personality?

There remains a final point of discussion, which is much more philosophical in nature. To many people, a definition of personality in terms of judgments by others constitutes a denial of subjective autonomy. That reproach should not be shrugged off lightly. It puts into words why students and lay persons, who turn to psychology primarily to better understand themselves, may experience bitter disappointment. It also contains an explanation of the sustained revolt against the trait approach to personality. Traits are of things, in this case, persons taken as objects. Subjective experience is coloured by intentions, not traits (Pervin, 1994).

I plead in favour of distinguishing between the questions of 'Who we are' and 'What people are like', between identity and personality. Who we are, what our destination is, or our identity, are philosophical questions. I do not urge anyone to dismiss these questions. But, there should be a division of labour between the science of personality, on the one hand, and philosophy and theology, on the other.

CONCLUSION

My main conclusion is that there is no justification for the emphasis in trait psychology upon self-ratings. The only excuse is provided by the principle of the Drunkard's Search: long after midnight, intoxicated persons look for their lost doorway under the street lamp, not because they lost it there, but because of the good light. Personality psychologists ask for self-judgments because the respondent is at hand.

It does not follow that self-judgments are useless. Under certain conditions they may add to the trait variance component in an average judgment. One may explicitly instruct the subject to take an external position, rather than trying to invoke some subjective point of view. To drive home the point, we presently recommend and practice the writing of personality questionnaires in the third person singular (Hendriks, Hofstee and De Raad, 1993). A fringe benefit of this practice is that one form suffices for self and others' ratings. But my main recommendation is to turn to third persons of flesh and blood in assessing personality.

The scientific importance of taking a hard look at the quality of personality measurement cannot be stressed enough. We are not (and should not become) accustomed to the inferior quality of our instruments. A conspicuous example is behavioural-genetic research on personality. By convention, all unexplained variance is attributed to the within-families component of environmental influences. With heritability coefficients for personality traits of the order of 0.4, and a between-families environmental component of the order of zero, one would be led to the conclusion that within-family environmental influences explain the largest part of the variance. In view of the low proportion of trait variance in our measurements, that impression is unwarranted. In the words of Spearman (1910): '... an estimate of the correlation between two things is generally of little scientific value, if it does not depend unequivo-
cally on the nature of things, but just as much on the mere efficiency with which they happen to have been measured' (p. 271). By applying a correction for attenuation, one might come to the conclusion that heritability is close to perfect. But this is only a second-best solution, in view of the error in estimating the relevant statistics. The proper solution for a branch of science is to sharpen its tools (Bouchard, 1993). Personality measurements should be averaged over a sufficient number of items, points in time, and most importantly, judges.

My argument should not be read as a plea for direct trait ratings. I share Cattell's (1946) assessment that 'The human judge abstracts the essentials from masses of behavior observation, sometimes with a bias but in a way that no machine measurements or statistical device can yet accomplish' (p. 211), but I do not recommend using unnecessarily inferential and evaluative ratings of traits such as extraversion, agreeableness, conscientiousness, emotional stability, and (in the Cattellian tradition) culture. We should make it our business to know the items that are associated with each trait, and then put them in an inventory.

By this combination of a responsible instrument, multiple judges, and—hopefully—a more representative sampling along the time dimension, we may succeed in pulling the trait variance in personality measurement to over 50 per cent. This has hardly been achieved up to now. It is an essential condition for progress in the field, both in theoretical and applied contexts.

ACKNOWLEDGEMENTS

Many colleagues have reacted to an earlier version, part of which was presented at the First Conference of the German Society for Differential Psychology, Personality, and Psychodiagnostics at Bielefeld, 16–18 July, 1993. My thanks go to Alois Angleitner, Paul De Boeck, Peter Borkenau, Trudy C. G. Dehue, Hans J. Eysenck, David C. Funder, Lewis R. Goldberg, Hubert J. M. Hermans, Robert Hogan, Marcel Van den Hout, Oliver P. John, John A. Johnson, Ghedolph A. Kohnstamm, Paul A. M. Van Lange, Robert R. McCrae, Paul Meltzer, Ivan Mervielde, Theo (M.) Nijssse, Lars Nystedt, Lawrence A. Pervin, Rainer Riemann, Gérard Saucier, and André S. H. Tjoa, all of whom have contributed to the articulation of the arguments; and to John M. Digman for uncovering the references to the Spearman and Brown papers.

REFERENCES


