Journal of Personality and Social Psychology
1998, Vol. 74, No. 5, 1337–1349

0022-3514/98/$3.00

Do People Know How They Behave? Self-Reported Act Frequencies Compared With On-Line Codings by Observers

Samuel D. Gosling, Oliver P. John, and Kenneth H. Craik
University of California, Berkeley

Richard W. Robins
University of California, Davis

Behavioral acts constitute the building blocks of interpersonal perception and the basis for inferences about personality traits. How reliably can observers code the acts individuals perform in a specific situation? How valid are retrospective self-reports of these acts? Participants interacted in a group-discussion task and then reported their act frequencies, which were later coded by observers from videotapes. For each act, observer–observer agreement, self–observer agreement, and self-enhancement bias were examined. Findings show that (a) agreement varied greatly across acts; (b) much of this variation was predictable from properties of the acts (observability, base rate, desirability, Big Five domain); (c) on average, self-reports were positively distorted; and (d) this was particularly true for narcissistic individuals. Discussion focuses on implications for research on acts, traits, social perception, and the act frequency approach.

"You interrupted my mother at least three times this morning" exclaims Roger. "That's not true," responds Julia, "I only interrupted her *once*!" And so the discussion continues. Disagreements about who did and did not do what are commonplace in social interactions. When such disagreements arise, whom should we believe? Perhaps Julia was distorting the truth to paint a favorable picture of herself. Alternatively, Roger may remember that Julia interrupted his mother, when really the conversation was interrupted by a telephone call; or perhaps Julia was so caught up with what she was trying to say that she did not notice that Roger's mother had not finished speaking. When caught in such situations, many of us, convinced that we are right, wish that somehow past events had been recorded on videotape so that we could triumphantly rewind the tape and reveal the veracity of our own reports. Unfortunately, in everyday life, no such video is available.

Samuel D. Gosling, Oliver P. John, and Kenneth H. Craik, Department of Psychology and Institute of Personality and Social Research, University of California, Berkeley; Richard W. Robins, Department of Psychology, University of California, Davis.

Preparation of this article was supported, in part, by a University of California Graduate Fellowship and National Institute of Mental Health Grant MH 49255. We are indebted to the Institute of Personality and Social Research, where the Master's of Business Administration Assessment Program was conducted, and to Manfred Amelang, Peter Borkenau, Jonathan Cheek, David Funder, Robert Hogan, Robert McCrae, Debbie Moskowitz, and Delroy Paulhus for their helpful comments on an earlier version. We are also indebted to Melanie Ballatore, Jean Chang, Sharada Kohn, Susanne Koller, Ky-Van Lee, Wendy McGowan, Joanne Shinozaki, and Jill Warburton for their painstaking coding of the videotapes, and to the members of the Gordon Allport Society at the University of California, Berkeley, for providing the ratings of the act properties.

Correspondence concerning this article should be addressed to Samuel D. Gosling, Institute of Personality and Social Research #5050, University of California, Berkeley, California 94720-5050. Electronic mail may be sent to samiam@uclink.berkeley.edu.

In the present study, however, we compared individuals' reports of their behavior with observer codings of their behavior from videotapes. Specifically, participants interacted in a 40-min group discussion task and then reported how frequently they had performed a set of acts. Observers later coded (from videotapes) the frequency with which each participant had performed each act. Thus, this design allowed us to compare retrospective act frequency reports by the self with on-line act frequency codings by observers.[1] Specifically, we examined whether individuals can accurately report how they behaved in a specific situation, and when and why their reports are discrepant from observer codings of their behavior. Understanding the processes that lead to accurate judgments about act performances is fundamental to the study of social perception.

Research on Acts

Recent developments in personality psychology, such as the act frequency approach (AFA; Buss & Craik, 1983), have emphasized the connections between behavioral conduct and personality judgments. According to the AFA, personality dispositions are anchored in the everyday acts of persons. Dispositions or traits are conceptualized as cognitive categories of prototypical acts. Similarly, trait assessments constitute summary statements about the frequency of prototypical acts, that is, about act trends in the person's conduct over time.

The AFA treats acts as relatively objective, concrete individual events from which act-based trait assessments of individuals can be made (Buss & Craik, 1983; Craik, 1997; Davidson, 1980). Although it is implied that acts can be recorded fairly accurately by coders or observers, the AFA has left open exactly how act occurrences should be determined. AFA research dealing with trait assessment, as opposed to act-based conceptual analysis

[1] By *on-line codings*, we mean that observers coded and recorded acts as they occurred rather than relying on memory.

(Buss & Craik, 1983, 1987; Shopshire & Craik, 1996), has often used retrospective reports of act frequencies by self and others (Botwin & Buss, 1989, p. 989); this practice has been criticized by Block (1989) because it assumes the accuracy of these reports. In contrast, although the AFA aims to assess personality in terms of in vivo acts occurring in everyday situations, on-line act reports have been less frequently used (but see Borkenau & Ostendorf, 1987; Moskowitz, 1986, 1994).

Clearly, though, research psychologists do not have access to an omniscient "God's eye" view of every act or deed of an individual. Instead, we are left with sociohistorical evidence about act occurrences, ultimately based on fallible witnesses.

Thus, for example, Barker and Wright (1951) used behavior records of children's lived days to assess episodes relevant to dominance, aggression, and nurturance; however, no interobserver agreement analysis could be conducted because their method used a single observer (Barker & Wright, 1951). Using a beeper technique, Moskowitz (1994) obtained on-line self-reports of acts prototypic of interpersonal traits but parallel on-line observer reports were not available. Buss and Craik (1983) and Botwin and Buss (1989) have gathered retrospective act frequency reports from self and other but again without parallel on-line reports. Newcomb (1929), Borkenau and Ostendorf (1987), and Borkenau and Müller (1992) studied on-line and retrospective act frequency reports from observers but did not obtain reports from the self. In short, each of these methods served the purposes of each study, but collectively their use underscores the need for basic research that advances our understanding of the process of monitoring act frequencies.

One reason why most studies have relied on retrospective reports is the considerable conceptual and logistical difficulty posed by coding behavior in specific situations on-line. On-line coding of behavior is difficult and extremely time consuming compared to using human judges as intuitive data accumulators and integrators (Borkenau & Ostendorf, 1987). Yet, although collecting retrospective judgments requires far less resources, questions remain about how well individuals can code, observe, remember, and retrospectively report on their own and others' naturally occurring behavior.

The Present Research

The present research examined the following questions. First, to what extent do people agree about how often an act occurred? For example, do Julia's self-reports of her behavior agree with Roger's reports of her behavior, and will Roger agree with other observers about Julia's behavior? Second, what makes an act easy to judge? That is, are there some attributes or properties intrinsic to a given act that influence the degree to which both self and observer agree about its occurrence? Third, do people accurately report what they did in a particular situation? For example, did Julia really interrupt Roger's mother only once? Fourth, are self-reports of specific acts biased by a motive to self-enhance, and are some individuals more likely to self-enhance than others? For example, does Julia tend to exaggerate her desirable behaviors?

The present research builds on recent investigations of the determinants of agreement and accuracy in personality judgments. For example, John and Robins (1993, 1994) and Kenny (1994) found observer–observer agreement in trait judgments to be consistently higher than self–observer agreement. Furthermore, Funder and Colvin (1988) and John and Robins (1993) found trait properties, such as observability, social desirability, and location within the five-factor model (FFM) of personality structure (John, 1990), to be related to observer–observer and self–observer agreement in trait judgments. Finally, John and Robins (1994) found that self-judgments at the trait level are influenced by self-enhancement bias, which in turn is associated with individual variations in narcissism. Ozer and Buss (1991) have begun to address issues of this kind at the level of act frequency reports. They showed, for example, that agreement between retrospective observer and self act frequency reports is higher for acts associated with Extraversion but lower for acts associated with Agreeableness.

The present study extends this line of inquiry by examining determinants of agreement and accuracy using on-line act reports by observers and retrospective act reports by the self. On-line observer reports warrant study because in aggregated form they represent an important criterion for act occurrence. Retrospective self-reports warrant study because the self is an ever-present monitor of act occurrence and because the self enjoys a distinctive and, in certain respects, privileged vantage point for interpreting the nature of acts as they are performed. At the same time, however, self-reports are vulnerable to self-enhancement and other biases. Below we formulate hypotheses based on self-concept theory and previous research in the act and trait domains.

How Well Do People Agree About How Often an Act Occurred?

Two types of agreement can be distinguished: agreement between observers (observer–observer agreement) and agreement between observers and the targets' own self-reports of their behavior (self–observer agreement). Bem (1967, 1972) and other cognitive-informational self-theorists have argued that individuals perceive their own behavior in much the same way as external observers do; the way individuals perceive themselves should, therefore, correspond closely with the way they are perceived by others. This view suggests that self and observer reports of act frequencies should show substantial convergence, especially when the reports concern an interaction situation that is brief and clearly delimited.

In contrast, studies of global trait judgments (Funder & Colvin, 1997; John & Robins, 1993; Kenny, 1994) and evaluations of task performance (John & Robins, 1994) have shown that the self is a unique judge: Self-judgments tend to agree less with observer judgments than observers agree with each other. On the basis of this research, we predicted that self–observer agreement on act frequency reports would be lower than observer–observer agreement (Hypothesis 1).

Do Acts Differ in How Much Individuals Agree About Act Frequencies?

What makes an act easy to judge? To address this question, Ozer and Buss (1991) asked spouses to report how frequently they had performed a set of acts over the previous 3 months.

Agreement between spouses varied across acts and depended on a number of properties of the acts. For example, spouses showed relatively high levels of agreement about acts related to Extraversion (e.g., "I danced in front of a crowd") but relatively little agreement about acts related to Agreeableness (e.g., "I let someone cut into the parking space I was waiting for"). The Ozer and Buss study provides insights into act properties that might moderate interjudge agreement. Several studies have identified properties of traits that influence agreement, including the observability of trait-relevant behaviors, the social desirability of the trait, and the Big Five content domain of the trait judged. If acts are indeed the building blocks of personality, then the properties affecting agreement about traits may also affect agreement about acts, and findings for acts should therefore parallel those for traits.

Thus, drawing on trait research, we made the following predictions about acts. First, we predicted higher observer–observer and self–observer agreement for acts that are easily observed (Funder & Dobroth, 1987; John & Robins, 1993; Kenrick & Stringfield, 1980; Ozer & Buss, 1991) (Hypothesis 2a). Some acts refer to psychological events or processes within the mind of the actor that may not be directly observable (e.g., "I appeared cooperative in order to get my way"), whereas other acts are more easily observed from an external vantage point (e.g., "I sat at the head of the table"). Highly observable acts will be more salient to observers (who focus on visible behaviors) than to the self-perceiver, for whom internal experiences (e.g., intentions and motives) are also available (Funder, 1980). Whereas observable behavior is, in principle, available to both observer and self, less observable aspects of an act (such as intentions) are available primarily to the self and are potentially more salient than observable aspects of the act (Robins & John, 1997b; White & Younger, 1988). Thus, it seems unlikely that all acts can be coded with high reliability by even the most conscientious observers.

Second, we predicted higher agreement for acts that occur frequently (Funder & Colvin, 1991; Ozer & Buss, 1991) (Hypothesis 2b). If an act has a low base rate of occurrence, then observers are more likely to miss it over the course of an interaction. Moreover, on psychometric grounds, low base-rate acts will have less variance across targets, which will tend to reduce correlations between observers. Both observability and base rate involve informational factors that might limit agreement about act performances.

We also expected motivational factors to play a role. In particular, we predicted that agreement would be related to the social desirability of the act (Hypothesis 2c). However, trait research provides conflicting evidence about whether this relation will be linear or curvilinear. That is, Funder and Colvin (1988) and Hayes and Dunning (1997) found a positive linear relation, with higher agreement for more desirable traits. In contrast, the two studies reported by John and Robins (1993) showed a curvilinear relation, with higher agreement for evaluatively neutral traits and lower agreement for evaluatively extreme traits (either highly undesirable or highly desirable). The present study will examine the effects of desirability and evaluativeness on agreement in the act domain.

Fourth, extrapolating from earlier findings, we predicted higher agreement for acts related to Extraversion (Funder & Colvin, 1988; John & Robins, 1993; Kenny, 1994; Norman & Goldberg, 1966; Ozer & Buss, 1991) and lower agreement for acts related to Agreeableness (John & Robins, 1993) (Hypothesis 2d).

How Accurate Are Self-Reports of Act Frequency?

The accuracy of self-perception has been a long-standing concern in psychology (see Robins & John, 1997a, for a review). Many theorists are less than sanguine about the ability of people to perceive their behavior objectively. Hogan (1996), for example, spoke of the "inevitability of human self-deception" (p. 165), and Thorne (1989) observed that "due to self-deception, selective inattention, repression, or whatever one wishes to call lack of self-enlightenment, self-views may be less accurate than outsiders' views" (p. 157).

Assessing the accuracy of self-reports requires a criterion—a measure of "reality" against which self-perceptions can be compared. Given the absence of a single objective standard for evaluating the accuracy of global personality traits, the social consensus (i.e., aggregated trait ratings by others) has often been used as an accuracy criterion (e.g., Funder, 1995; Hofstee, 1994; Norman & Goldberg, 1966; Robins & John, 1997a). For example, much research on the accuracy of self-reports has compared self-ratings with judgments provided by peers (John & Robins, 1994; Kolar, Funder, & Colvin, 1996). However, some researchers have been skeptical of reports by such informants and have instead emphasized the need for direct behavioral observation (e.g., Kenny, 1994, p. 136). Hence, the present research focused on observer codings of act frequencies from videotapes in a specific interaction task. These codings provide a more objective measure of the behavioral reality in the task and can therefore serve as a criterion to evaluate accuracy and bias in self-reports of behavior in this task (Funder, 1995; Kenny, 1994; Robins & John, 1997b). We expected self-reported act frequencies to reflect, at least in part, the observed "reality" of participants' behavioral conduct. Thus, we predicted that the self-reports would show levels of accuracy similar to those found in trait research (Hypothesis 3). However, we did not expect the accuracy correlations to be uniformly high, so we also examined the properties of acts that might explain why accuracy is higher for some behaviors than for others.

Are Self-Reports of Act Frequency Biased?

Do individuals overreport socially desirable acts to enhance their self-views? Most self-concept theorists assume that people are motivated to maintain and enhance their feelings of self-worth (e.g., Allport, 1937; Greenwald, 1980; James, 1890; Rogers, 1959; Tesser, 1988). According to Taylor and Brown (1988, 1994) and others, most individuals have "positive illusions" about themselves, presumably stemming from the basic motive toward self-enhancement. Several studies have examined positive illusions by comparing self-reports to observer ratings of global personality traits, such as friendly and outgoing (Campbell & Fehr, 1990; Colvin, Block, & Funder, 1995; Lewinsohn, Mischel, Chaplin, & Barton, 1980). This research on trait ratings shows that, on average, individuals perceive themselves

somewhat more positively than they are perceived by others. If these positive illusions extend to perceptions of specific behaviors, then we would also expect individuals to show a self-enhancement bias in their act reports.

Current research on self-enhancement bias focuses on identifying boundary conditions (e.g., Sedikides & Strube, 1997). This research pursues two lines of inquiry. One line aims to identify properties of behaviors that influence the degree of self-enhancement (e.g., Dunning, Meyerowitz, & Holzberg, 1989). That is, are some acts overreported more than others? The other line aims to identify characteristics of persons associated with the tendency to self-enhance (e.g., John & Robins, 1994). That is, are some individuals particularly inclined to overreport their desirable behavior and underreport their undesirable behavior? Thus, the degree of bias may vary both across acts and across individuals. To understand this variability, we examined both properties of acts (e.g., how desirable the act is) and a characteristic of individuals (e.g., how narcissistic the individual is) that might predict bias.

Properties of acts may influence self-enhancement bias through two general processes. First, the social desirability of an act should activate motivational processes (e.g., self-esteem maintenance) that positively bias individuals' self-reports. Thus, we predicted that participants would paint a favorable portrait of themselves by overreporting their socially desirable acts relative to their undesirable acts (Hypothesis 4a).

Second, the observability of an act reflects how much information is available to self and other judges. In principle, the self has greater access to information about less observable aspects of acts (i.e., internal processes) than do observers, whereas both self and other judges have access to information about highly observable acts (i.e., acts involving overt behavior), and there may be some acts for which other judges have better access (cf. Funder, 1980). Thus, we predicted that individuals would overreport unobservable acts relative to the observer codings and underreport highly observable acts (Hypothesis 4b).[2]

Illusory self-enhancement is sometimes described as if it is present in all normal, psychologically healthy individuals: Taylor (1989) concluded that "normal human thought is marked not by accuracy but by positive self-enhancing illusions" (p. 7); Paulhus and Reid (1991) emphasized that "the healthy person is prone to self-deceptive positivity" (p. 307); and Greenwald and Pratkanis (1984) believed that self-enhancing biases pervade the "self-knowledge of the average normal adult of (at least) North American culture" (p. 139). However, John and Robins (1994) found self-enhancement bias in only 60% of their participants who evaluated their performance in a group discussion task more positively than did a group of independent observers. This finding raises the question of whether some individuals are particularly prone to positive illusions. As noted by John and Robins, the most theoretically relevant construct is narcissism. The *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*) criteria for the narcissistic personality include a grandiose sense of self-importance, a tendency to exaggerate accomplishments and talents, and an expectation to be recognized as "extraordinary" even without appropriate accomplishments (American Psychiatric Association, 1994). Research suggests that narcissistic individuals respond to threats to their self-worth by perceiving themselves more positively than is justified (Gabriel, Critelli, & Ee, 1994; John & Robins, 1994) and by denigrating others (Morf & Rhodewalt, 1993). Narcissists may be particularly prone to positively distorted self-evaluations because their inflated sense of self-importance is easily threatened. Thus, we predicted that narcissistic individuals will show more self-enhancement bias than nonnarcissistic individuals in their act frequency self-reports (Hypothesis 4c).

Method

Participants

Ninety Masters of Business Administration (MBA) students (41 women, 49 men) volunteered to participate in a personality and managerial assessment program. Because of technical problems, the videotapes of 2 participants were unusable; thus, the final N was 88. Their median age was 29 years, and on average they had more than 3 years of postcollege work experience. We collected data from two samples: 54 participants (26 women) in Sample 1 and 36 participants (15 women) in Sample 2.

Group Discussion Task

The group discussion task we used is a standardized exercise commonly used to assess managerial performance (e.g., Howard & Bray, 1988; Thornton & Byham, 1982). The task simulates a committee meeting in a large organization. Participants were randomly assigned to mixed-sex groups, with 6 members in each. Participants were told that the purpose of the meeting was to allocate a fixed amount of money to 6 candidates for a merit bonus. Each participant was assigned the role of supervisor of one candidate and was instructed to present a case for that candidate at the meeting; participants were seated at a round table and no leader was assigned. Participants received a realistic written summary of the employment backgrounds of all candidates, including salary, biographical information, and appraisals of prior job performance, and were given 10 min to review this information. They were instructed to start the meeting by each giving a 3- to 5-min presentation on the relative merits of their candidate. The groups had 40 min to reach consensus on how to allocate the merit bonuses. Instructions emphasized two goals: (a) obtain a large bonus for the candidate they represented and (b) help the group achieve a fair overall allocation of the bonus money. Thus, effective performance required behaviors that promoted the achievement of both goals. To permit subsequent coding of act frequencies, the task was videotaped with cameras mounted unobtrusively on the walls and focused on each participant's face and upper body.

Selection of Acts

We studied a total of 34 acts (20 acts in Sample 1 and 14 acts in Sample 2). We had two goals in selecting these acts: We wanted our findings to be relevant to previous AFA research, and we wanted the acts to be relevant to our group discussion task. Thus, in Sample 1, we selected 20 acts from a large set of acts generated by AFA procedures (Botwin & Buss, 1989) that seemed likely to occur in our task (e.g.,

[2] Note that this prediction does not imply that the self-reports are necessarily inaccurate, but only that they are biased relative to the observer criterion.

"Target issued orders that got the group organized").[3] In Sample 2, five psychologists familiar with the group discussion task generated a second set of 14 acts that refer to easily observable behaviors and occur often in this task (e.g., "Target outlined a set of criteria for determining how to allocate the money").

Self-Reports of Act Frequency

Immediately after completing the task, participants reported how frequently they had performed each act during the group discussion. The acts were worded in the first person (e.g., "I persuaded the others to accept my opinion on the issue"). Buss and Craik (1983; Buss, 1981) used a 4-step rating scale for their retrospective act frequency reports which extended over 3 months (0 = *act not performed,* 1 = *act performed rarely,* 2 = *act performed sometimes,* 3 = *act performed often*). To provide greater specificity in our 40-min task, we used a 4-point scale referring to the actual frequency of acts performed (0 = *not at all,* 1 = *once,* 2 = *two or three times,* 3 = *more than three times*).

Video-Based Observer Codings of Act Frequency

In Sample 1, four observers viewed the videotaped behavior of each participant and coded the frequencies of each of the 20 acts. In Sample 2, a second set of four observers coded the additional 14 acts for each participant. Both sets of observers were students at the same university but unacquainted with the videotaped participants. Acts were worded in the third person (e.g., "Target persuaded the others to accept his/her opinion on the issue"). Before viewing the videotapes, the observers watched four practice videotapes (which were not used in this research) to familiarize themselves with typical behavioral repertoires and the way the acts were manifested in the task.

Viewing order of the videotapes was randomized across observers. The observers coded each tape alone in a laboratory room. Coding the videotapes was a painstaking task. The observers went over 40 min of videotaped behavior, minute by minute, rewinding and reviewing the tapes whenever necessary. Each time one of the acts occurred, the observer recorded it on a scoring sheet. After scoring the whole tape, the observers tallied up the number of times they had recorded an act for the target participant, thus providing a measure of each participant's act frequency based on careful scrutiny of on-line behavior. The four observers coded participants' act frequencies with reasonable reliability; across the 34 acts, the average coefficient alpha reliability of the composited ratings was .69 (SD = .29).

Independent Variables: Properties of Acts

For each of the 34 acts, we measured four properties hypothesized to influence interjudge agreement and accuracy and bias in self-reported act frequencies.

Observability. Two facets of observability were rated by eight judges who were familiar with the group discussion task: *Noticeability* was defined by how well the act stands out from the stream of behavior (α = .89), and *high inferential content* was the degree of inference about internal thoughts and motivations required for an observer to be sure that the act has occurred (α = .96). Both judgments were made on 9-point rating scales. Across the 34 acts, mean ratings for noticeability and high inferential content were strongly negatively correlated (r = −.80). Therefore, we standardized both variables, reverse scored high inferential content, and combined the two ratings into one overall measure of observability. The most observable act was "Target reminded the group of their time limit"; the least observable act was "Target took the opposite point of view just to be contrary."

Social desirability. Using a 9-point scale (Hampson, Goldberg, & John, 1987), the judges also rated how socially desirable it was to perform each act in the group discussion. The mean ratings were used as an index of each act's desirability (α = .94). The most desirable act was "Target settled the dispute among other members of the group"; the least desirable act was "Target yelled at someone." Evaluativeness was measured by folding the 9-point scale such that 1 and 9 were recoded as 4, 2 and 8 were recoded as 3, and so on.

Base rate. The base rate of an act was the number of times the act was performed by any participant, on the basis of the observer codings. This index was computed separately for each observer and then composited; the mean alpha (averaged across the two sets of observers) was .83. The act with the highest base rate was "Target expressed her/his agreement with a point being made by another member of the group"; the act with the lowest base rate was "Target monopolized the conversation." Across the 34 acts, base rate correlated .10 (ns) with observability and .45 ($p < .05$) with social desirability (see Pratto & John, 1991), which, in turn, correlated .13 (ns) with each other.

Big Five personality domain. Acts in the group discussion task tend to be overt behaviors that are either interpersonal (e.g., negotiation and persuasion) or task-oriented (setting goals and organizing group activities; Bass, 1954). In terms of the Big Five personality domains, the interpersonal domains of Extraversion and Agreeableness and the task-focused domain of Conscientiousness were most relevant. In contrast, the other two Big Five domains (Neuroticism, Openness to Experience) refer primarily to an individual's covert experiences. Three expert judges rated the prototypicality of each act for each of the Big Five domains, with low ratings indicating the act was unrelated to that Big Five domain and high ratings indicating the act was highly related to either high or low pole. For example, the Extraversion rating for each act ranged from 0 (*act is unrelated to Extraversion or Introversion*) to 4 (*act is extremely prototypical of Extraversion or Introversion*). The alpha reliabilities of their composite judgments were high for Extraversion (.81), Agreeableness (.86), and Conscientiousness (.88) and somewhat lower for Neuroticism (.67) and Openness to Experience (.62). There were no prototypical examples of the Neuroticism and Openness to Experience domains. All acts had their highest mean prototypicality values on Extraversion, Agreeableness, or Conscientiousness, and therefore only these three Big Five domains will be examined in our analyses. "Target laughed out loud" was the most prototypical act for the Extraversion domain, "Target took the opposite point of view just to be contrary" for (low) Agreeableness, and "Target reminded the group of their time limit" for Conscientiousness. We used these continuous prototypicality ratings in our correlational analyses. For our analyses of variance (ANOVAs), the acts were classified independently by two of the judges into the Extraversion (α = .75), Agreeableness (α = .83), and Conscientiousness (α = .87) domains. The reliability of these classifications, computed across acts, suggests that the judges agreed about the Big Five content domain of each act.

Narcissism

We used the 33-item version of the Narcissistic Personality Inventory (NPI; α = 70; Raskin & Terry, 1988) to assess participants' level of narcissism. The NPI is the best validated self-report measure of overt narcissism for nonclinical populations (Raskin & Terry, 1988; see also Hendin & Cheek, 1997) and has been shown to predict psychologists' ratings of narcissism (e.g., John & Robins, 1994).

---

[3] Two of the 20 acts were edited slightly to make them more appropriate for our task. The Botwin and Buss (1989) act "I loudly corrected my friend's mistake" was rewritten to "I loudly corrected the mistake the previous speaker had made," and the act "I said 'OK' to every suggestion offered about my project" was changed to "I said 'OK' to every comment offered about my candidate." We are grateful to David M. Buss for providing us with his large set of acts.

Dependent Variables

Interjudge agreement: Observer–observer and self–observer agreement. To assess how much the observers agreed about the frequency of each act, we computed the correlation (across participants) between each pair of observers' video-based codings.[4] We then averaged the resulting six pairwise observer–observer correlations. This index reflects the average observer–observer agreement for each act.

To assess how much self and observer agreed about the frequency of each act, we computed the correlation (across participants) between the self-reports and video-based codings by each of the four observers. We then averaged the resulting four dyadic self–observer correlations. This index reflects the average agreement between self and a single observer and is therefore directly comparable to the dyadic observer–observer agreement index.

Accuracy and bias in self-reported acts. To assess accuracy and bias, we used the aggregated video-based observer codings as a behavior-based criterion measure of act frequency. Accuracy was defined by the correlation (computed across participants) between self-reports of act frequency and the observer criterion for act frequency. Bias was defined by the discrepancy between each participant's self-report and the observer criterion; positive values indicate that participants overreported how frequently they performed the act, and negative values indicate they underreported how frequently they performed the act. Bias can be computed both at the aggregate level (i.e., do individuals, on average, overreport or underreport some acts more than others?) and at the level of the individual person (i.e., do some persons overreport or underreport an act more than others?). Both accuracy and bias were computed separately for each act.

The dependent variables were computed separately for the acts in each sample. However, because the findings were similar in both samples, analyses across acts used the whole set of 34 acts.

Results and Discussion

Do Observers Agree More With Each Other Than They Do With the Self?

We first tested Hypothesis 1, which predicts that observer–observer agreement would be higher than self–observer agreement. Across the 34 acts, observer–observer agreement (M = .40, SD = .25) was significantly higher than self–observer agreement (M = .19, SD = .19), as shown by a t test for paired samples, $t(33) = 5.2$, $p < .001$, one-tailed. This effect held for 83% of the acts. In short, two observers generally agreed more about an act's frequency than did the self and an observer. One could argue that this difference is due to the fact that the observers had access to a videotaped record of the participant's behavior, whereas the participants' self-reports were made retrospectively. Would agreement be higher if participants had watched their act performances on videotape, just as the observers had? Using the same task as the present study, Robins and John (1997b) obtained self-ratings of performance in two conditions: retrospectively and after participants watched their own behavior on videotape. These video-based self-reports did not show greater agreement with observer judgments than did the retrospective self-reports.

We also found that acts eliciting high levels of observer–observer agreement also tended to elicit high self–observer agreement; the correlation between the two agreement indices across the 34 acts was .65, closely replicating the .64 value reported by John and Robins (1993) for trait ratings. In other words, when two observers agree about an act (or a trait), self and observer are also likely to agree.

Table 1
Correlations Between Act Properties and Interjudge Agreement on Act Frequency Reports (Computed Across the 34 Acts)

Act properties	Observer–observer agreement	Self–observer agreement
Observability	.38*	.34*
Base rate	.44*	.35*
Desirability	.52*	.46*
Evaluativeness	−.14	−.06
Prototypicality for Big Five domain		
Extraversion	.08	.32*
Agreeableness	−.27†	−.51*
Conscientiousness	.20	.38*
R (adjusted for shrinkage)	.67 (.55)	.77 (.69)

Note. Numbers in this table are correlations computed across the 34 acts. For example, the correlation of .38 between observability and observer–observer agreement indicates that more observable acts tended to elicit higher levels of agreement than less observable acts. Similarly, the correlation of −.51 between Agreeableness and self–observer agreement indicates that acts from the Agreeableness domain (i.e., prototypical examples of either Agreeableness or Disagreeableness) tended to elicit lower levels of self–observer agreement than acts unrelated to Agreeableness.
† $p < .10$ (marginally significant). * $p < .05$.

What Makes an Act Easy to Judge? Effects of Observability, Social Desirability, Base Rate, and Big Five Domain

The level of agreement varied substantially across acts, ranging from −.08 to .88 for observer–observer agreement, and from −.12 to .62 for self–observer agreement. Why are some acts judged more consensually than others? To address this question, we correlated the act properties with observer–observer and self–observer agreement across the 34 acts. These across-act correlation coefficients are given in Table 1. As predicted by Hypotheses 2a, 2b, and 2c, observability, social desirability, and base rate of the acts were all positively and substantially correlated with both observer–observer and self–observer agreement. The observability effect is consistent with Ozer and Buss's (1991) research on acts, as well as with Funder and Dobroth's (1987) and John and Robins's (1993) research on traits. The positive linear relation between social desirability and agreement is consistent with Funder and Dobroth (1987) and Hayes and Dunning (1997). However, we did not find the

[4] We computed agreement using both the Pearson product–moment correlation and the intraclass correlation (Shrout & Fleiss, 1979). The two types of correlation were almost perfectly correlated (r = .98), suggesting that our findings would not be affected by the method of computing the correlation. We therefore retained the Pearson correlation so as to make our results comparable to previous research on interjudge agreement. All computations involving correlation coefficients were done using Fisher's r-to-Z transformation.

evaluativeness effect reported by John and Robins, who found that both extremely negative and extremely positive traits elicit lower levels of agreement.[5] In summary, acts that were observable, desirable, and occurred relatively frequently were judged with relatively more agreement than acts that were difficult to observe, undesirable, and relatively infrequent.[6]

Table 1 also shows the correlation between Big Five content domain and interjudge agreement. These correlations are generally consistent with Hypothesis 2d. Self–observer agreement correlated positively with act prototypicality for both Extraversion and Conscientiousness and negatively with prototypicality for Agreeableness, indicating that self–observer agreement was higher for acts from the Extraversion and Conscientiousness domains and lower for acts from the Agreeableness domain. The same pattern was found for observer–observer agreement, but the correlations did not reach conventional levels of significance. These findings are generally consistent with previous research in both the act and trait domains. However, there were two differences. First, we did not find the Extraversion effect for observer–observer agreement found in several previous studies (e.g., John & Robins, 1993; Kenny, 1994). Second, we found a Conscientiousness effect for self–observer agreement that has not been found in previous research.

To examine how well the act properties jointly predicted observer–observer and self–observer agreement, we conducted a multiple regression analysis.[7] After adjusting for the effects of shrinkage, we were able to predict inter-act differences in observer–observer agreement with an R of .55 and differences in self–observer agreement with an R of .69. Thus, observability, base rate, social desirability, evaluativeness, and Big Five domain predicted a substantial proportion of the total variation in agreement across acts. These Rs are similar in magnitude to those reported by John and Robins (1993). In summary, there are a number of properties that together help us understand how well individuals agree, whether one is examining how often individuals perform an act or how they rate on a personality trait.

How Accurate Are Self-Reports of Act Frequency?

To examine accuracy, we correlated the self-reported act frequencies with the aggregated observer codings. Across all 34 acts, the mean correlation was .24 (SD = .26). However, this value underestimates the accuracy of the self-reports because for some acts the video-based observer codings were not highly reliable. Thus, as a fairer test, we considered only those 12 acts that observers coded with high reliability (i.e., those with an alpha above .80).[8] Consistent with Hypothesis 3, the self-reports showed a significant level of accuracy, with a mean correlation of .40 (SD = .26; see Table 2).

However, the accuracy correlations varied considerably even within this subset of highly reliable acts, ranging from a high of .72 to a low of .03. Table 2 presents the Big Five classifications of these 12 acts. The 4 acts with the highest accuracy correlations (mean r = .61) were all from the Extraversion domain, whereas the 5 acts with the lowest accuracy correlations (mean r = .16) were all from the Agreeableness domain; the 3 Conscientiousness acts fell in between, with a mean r of .44. To test whether this pattern of findings could be attributed to chance, we conducted a one-way ANOVA on the accuracy correlations; Big Five domain (Extraversion, Conscientiousness, Agreeableness) was the independent variable and the acts served as the unit of analysis (i.e., n = 12 observations). The effect of Big Five domain was significant, $F(2, 9) = 19.9$, $p < .001$. Thus, even for acts coded with high reliability, the accuracy of self-reports varied widely, and this variation was related to predicted differences among Big Five domains. Apparently, then, the use of retrospective self-reported act frequencies is a promising methodology for acts from the Extraversion domain but less promising when the acts come from the Agreeableness domain. Future research should examine the usefulness of retrospective self-reported act frequencies for the two Big Five domains not included in the present research, Neuroticism and Openness.

Are Self-Reports of Act Frequency Biased?

The accuracy findings suggest that individuals' reports of their behavior are at least partially based on "reality." However, the congruence between self-reports and codings of behavior is far from perfect, suggesting that individuals are doing more than merely reporting what they did. Thus, there may be systematic biases in act reports. One possibility is that this bias reflects self-enhancement motivation: Individuals can enhance their self-worth by overreporting desirable acts and underreporting undesirable acts. Thus, Hypothesis 4a predicts that the desirability of an act will be associated with overreporting of that act. A second possibility is that the bias reflects the salience of information to the self: Some acts are more easily observed from an external perspective than from the self perspective (e.g., we may not be aware of nonverbal cues that are apparent to others). Thus, Hypothesis 4b predicts that highly observable acts will be underreported by the self, and that less observable acts will be overreported (relative to the observer codings).

---

[5] As Hayes and Dunning (1997) argued, the curvilinear evaluativeness effect may depend on the particular sampling of acts or traits. That is, the relation between desirability and agreement may be positive and linear unless the study includes both clearly neutral and extremely positive stimuli. The present set of acts included few neutral acts and only one extremely positive act, and thus may not provide an appropriate test of the evaluativeness hypothesis.

[6] Note that these effects held for both observer–observer and self–observer agreement. However, given the substantial correlation between these two types of agreement, the effects are not independent, as shown by partial-correlation analyses.

[7] To test whether these act properties had independent effects on agreement, we conducted partial-correlation analyses. We partialed observability in the first analysis, base rate in the second, and social desirability in the third. Together these analyses showed that the observability effect and the social desirability effect remained significant even when the effects of the other two variables were partialed out. The base-rate effect was also independent of observability but dropped below conventional levels of significance when social desirability was partialed. In summary, observability and social desirability had significant and independent effects on agreement, whereas the effects of base rate depended, at least in part, on the desirability of the acts. Partial correlation analyses further showed that the Big Five effects were largely independent of the effects of observability, base rate, and social desirability.

[8] We repeated these analyses using less stringent alpha cutoffs of .60 and .70, and the findings did not change significantly.

Table 2
The 12 Most Reliably Coded Acts ($\alpha > .80$) Ranked by Their Self–Observer Validity

Act	Big Five domain	Self–observer validity
Told joke to lighten tense moment	E	.72
Made humorous remark	E	.60
Took charge of things at the meeting	E	.57
Laughed out loud	E	.52
Outlined set of steps thought group should follow	C	.45
Pointed out the distinction between a merit bonus and salary increase	C	.45
Reminded group of time limit	C	.41
Said was willing to lower the money recommending for own candidate	A	.32
Expressed agreement with another group member	A	.31
Pointed out possible effects on employee morale	A	.08
Interrupted someone else	A	.07
Suggested they give some money to every candidate	A	.03
M		.40
SD		.26

Note. The act descriptions have been slightly abbreviated. All acts are desirable (i.e., rated above 6 on the 9-point social desirability scale) except "Interrupted someone else," which was undesirable (mean desirability = 2.8), and "Laughed out loud," which was relatively neutral (mean desirability = 5.6). E = Extraversion; A = Agreeableness; C = Conscientiousness.

To test these hypotheses, we first computed an index of bias (i.e., under- vs. overreporting) in the self-reports. Specifically, we conducted a multiple regression across the 34 acts predicting self-reported act frequency from the mean observer codings.[9] The mean observer codings represent the observable "reality" component in self-perception and accounted for almost 30% of the variance in mean self-reported act frequencies ($\beta = .54$, $p < .05$). The residual variance represents bias in the self-reports. The standardized regression residuals were retained to index bias; positive residuals reflect overreporting by the self and negative residuals represent underreporting by the self relative to the video-based observer codings. Next, to predict bias in self-reports, we conducted a second multiple regression in which we entered the social desirability and the observability of the acts. Consistent with Hypotheses 4a and 4b, both variables had significant effects (βs = .44 and −.39, respectively, both ps < .05). Overall, the multiple correlation was .80, with a shrinkage-adjusted R^2 of .60. Thus, we were able to predict what participants said they did from what coders said they did and from what kind of acts they were reporting (see Figure 1). More specifically, although participants' reports of what they did were strongly linked to their observable behavior (as captured on videotape), their reports also depended on the desirability and observability of the behavior. The desirability and observability effects suggest that both motivational and informational factors bias what individuals report they did in a situation.

Individual Differences in Self-Enhancement Bias

So far we have shown that certain kinds of acts, namely desirable acts and less observable acts, elicit overreporting relative to the video codings. Now we turn to the question of whether certain kinds of individuals give biased reports of their behavior. To establish the existence of such individual differences, we examined for each desirable act the percentage of individuals whose self-reported act frequencies were greater than, less than, and the same as the observer codings. Averaging the percentages across the 15 desirable acts, 57% of the participants overreported (i.e., showed self-enhancement bias), 24% underreported (i.e., self-diminishment bias), and 19% were exactly accurate. That is, 43% of the participants failed to show the general self-enhancement effect (Taylor & Brown, 1988). Clearly, then, individuals show substantial differences in self-perception bias, suggesting that the self-enhancement tendency should not be treated as a general law of social behavior (John & Robins, 1994).

To test the prediction that narcissism will predict these individual differences in self-enhancement, we computed a self-enhancement index based on the degree to which participants overreported their desirable acts plus the degree to which they underreported their undesirable acts. Consistent with Hypothesis 4c, the NPI correlated .27 ($p < .05$) with this self-enhancement index.[10] Analyses of individual acts revealed that narcissists were particularly inclined to overreport desirable acts such as "I took charge of things at the meeting" and "I made an argument that changed another person's mind." The tendency for narcissistic individuals to exaggerate the frequency with which they performed desirable acts provides further support for the link be-

[9] Because the acts differ in how reliably they were coded by the observers, we first removed these reliability differences using regression and retaining the residuals. Thus, our findings cannot be due to reliability differences.

[10] We also computed the correlation between narcissism and bias separately in the two samples and then averaged these correlations (weighted by sample size). This averaged correlation was .28, almost exactly the same as when the self-enhancement index was standard scored and combined across the two samples.

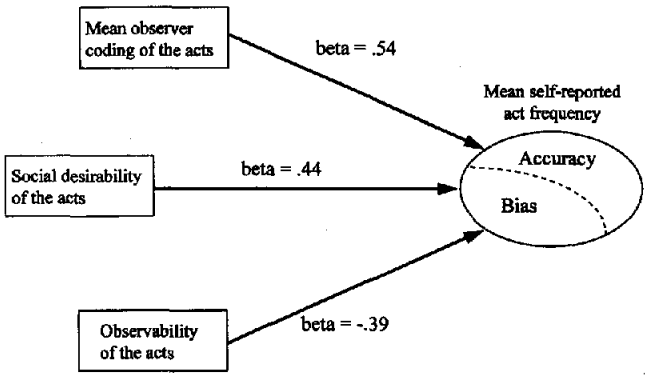

Figure 1. Which acts are people most likely to overreport? Determinants of mean self-reported act frequencies.

tween narcissism and positive illusions about the self (John & Robins, 1994).

General Discussion

This research addressed a fundamental question about self-perception: Do people know how they acted in a particular situation? We compared individuals' reports of how frequently they performed a set of acts with observer codings of their behavior from videotapes. We found that for some acts there is a clear consensus about how often the act occurred whereas for other acts individuals simply do not agree. We explored several factors that might account for these differences and found that individuals tend to agree about acts that are observable, desirable, frequently occurring, and are from the Extraversion and Conscientiousness (rather than the Agreeableness) domains. We also examined how accurately people report on their behavior and whether their reports are positively biased. We found that individuals' recollections of their behavior showed some correspondence with codings of their behavior, but the degree of correspondence varied systematically across acts. Finally, we found a general tendency toward self-enhancement bias in the act self-reports, but the degree of bias depended on both the individual act and the individual person. Specifically, self-enhancement was greatest for acts that were highly desirable and difficult to observe and for persons who were particularly narcissistic.

What can these findings tell us about the disagreement between Julia and Roger regarding how many times she had interrupted his mother that morning? First, we can expect less agreement between self and other, Julia and Roger, than between Roger and another observer. Second, however, for both Julia and Roger, the amount of agreement will depend on the specific act being monitored; we would expect relatively low agreement because "interrupting another person" is an undesirable and disagreeable act. Third, given that act self-reports are susceptible to self-enhancement bias, we would expect Julia to underestimate how often she had in fact interrupted Roger's mother, especially if she has narcissistic tendencies. In short, our analysis suggests that their disagreement may resist easy resolution.

We now move beyond the rather specific context of Julia and Roger's disagreement and turn to the wider implications of the findings. Specifically, we discuss how the findings compare with previous research in the act and trait domains, and what they imply for act-based personality assessment as advocated by the AFA.

Comparison of Act and Trait Research on Agreement and Accuracy

Research on acts. Only a few studies have examined observer–observer and self–observer agreement in the act domain. Table 3 summarizes the results of six such studies, including the present one. In each study, judges provided reports of specific behaviors that occurred over a specified period of time. However, the studies differed in several important respects. First, different types of judges were used. As shown in Table 3, only half of the studies obtained self-reports. Whereas all the other studies used observers who were previously unacquainted with the targets, Ozer and Buss (1991) relied on spouses as observers. Spouses know the target well and thus might base their reports on a much broader knowledge base; however, they are also more likely to be emotionally invested in their ratings and thus more likely to be biased. Second, most of the studies used retrospective observer reports whereas we used on-line codings of behavior from videotapes. Third, the level of analysis varied from extremely molecular acts such as "I sang a song in front of the group" to more molar behaviors, such as "acted aggressively." Finally, the studies differed markedly in the duration of the observational period. The 3-month period studied by Ozer and Buss was much longer and less clearly circumscribed than the interaction tasks used in Borkenau and Ostendorf (1987), Funder and Colvin (1991), and the present study.

To integrate the findings from this diverse set of studies, we combined the findings across the studies.[11] This meta-analysis shows an impressive degree of convergence across studies. Across the four studies that reported observer–observer agreement, the average (weighted by number of targets rated) was .64 (SD = .07) for alpha reliability, and .28 (SD = .09) for pairwise (unaggregated) observer–observer agreement. Similarly, self–observer agreement averaged .21 (SD = .08) across the three relevant studies; this lower value further supports the generality of John and Robins's (1993) hypothesis that self–observer agreement is lower than observer–observer agreement. These estimates of agreement may appear small, but it is important to consider that they involve single behaviors reported by a single judge (self or observer). Multiple-act composites (or act trends; cf. Buss & Craik, 1983) take advantage of the appreciable effects of aggregation across acts and yield considerably higher levels of agreement. Note that variability in the findings—expressed by the standard deviations across studies—was small, suggesting that the overall means in Table 3 can be assumed to represent the typical findings in this literature reasonably well. Overall, the findings from our own study seem to fit well with the conclusions from the meta-analysis: We

[11] Note that the alpha reliabilities reported in Table 3 are derived from varying numbers of judges (4–6). Therefore, to make the agreement findings comparable across studies, Table 3 also presents the unaggregated mean pairwise agreement correlations, which are, of course, considerably lower.

Table 3
Mean Interobserver and Self–Observer Agreement Correlations in Studies of Ratings of Single Acts and Traits

Study	Interobserver α	Interobserver r	Self–observer r	No. targets	No. judges	Observational base	Example of behavior description used
			Behavior				
Present study	.69	.40	.19	90	4	40-min on-line	Reminded group of time limit
Borkenau & Ostendorf (1987)	.65	.27[a]		48	5	50-min on-line/ retrospective	Changed the subject
Borkenau & Müller (1992)	.53	.18[a]		24	5	50-min retrospective	Acted "aggressively"
Funder & Colvin (1991)	.64[b]	.23[a]		140	6	5-min retrospective	Offered advice to partner
Kolar et al. (1996)			.12[c]	140	6	5-min retrospective	Tends to proffer advice (self) Offered advice to partner (observer)
Ozer & Buss (1991)			.28	186	1	3-month retrospective	Interrupted a conversation
Across studies							
M	.64	.28	.21				
SD	.07	.09	.08				
			Traits				
John & Robins (1993)							
Study 1	.57[a]	.25	.19	50	4	Peers	Critical
Study 2	.53[a]	.22	.20	218	2–4	Peers	Talkative–quiet

[a] Estimated from alpha coefficients according to Spearman–Brown prophecy formula. [b] Median alpha across 62 acts. [c] Self-ratings on California Adult Q-Sort items correlated with behavioral Q-sort ratings by six observers.

found slightly higher-than-average observer–observer agreement but average levels of self–observer agreement.

Research on traits. Our findings for acts show striking parallels with previous research on traits. For example, John and Robins (1993) reported two studies of peer–peer and self–peer agreement on a wide range of personality traits (e.g., talkative, critical). Mean peer–peer agreement was .23, indicating slightly lower levels of observer agreement than in act research (r = .28). The mean self–peer correlation was .20, similar to the self–observer agreement correlation of .21 found in the three act studies reviewed here. Thus, both act and trait studies show that self-ratings generally agree less with ratings by observers than observer ratings agree with each other (John & Robins, 1993, 1994; Kenny, 1994). Finally, both act and trait research suggests that self-perceptions, whether on specific acts or global traits, show some self-enhancement bias when compared with observer judgments.

However, the resemblance between the act and trait domains is not perfect. For example, in the present study, we found evidence that acts from the Conscientiousness domain elicit higher levels of self–observer agreement, whereas this effect has not been found for traits. In contrast to the trait findings (Funder & Colvin, 1988; John & Robins, 1993; Kenny, 1994), the present act research did not find strong support that Extraversion acts show higher observer–observer agreement. The positive linear relation between act desirability and both observer–observer and self–observer agreement is consistent with trait findings from Funder and Colvin and from Hayes and Dunning (1997) but not with John and Robins (1993). Finally, we found that the base rate of an act was an important predictor of both types of agreement, an effect not yet tested in trait research where the concept of base rate is less directly applicable.

Overall, then, there appear to be both similarities and differences between agreement on acts and agreement on traits. Clearly, an important avenue for future research concerns the psychological roots of these similarities and differences. Such research will need to take into account differences in the way act and trait judgments are made. One might expect judges to agree more about acts than about general personality traits because many acts are directly observable (Buss & Craik, 1980, 1983; Kenny, 1994), whereas traits represent summary impressions of multiple-act occurrences. Thus, trait inferences require first perceiving specific behaviors and then abstracting them into trait ascriptions. On the other hand, agreement may be higher for traits because trait inferences are typically based on a diverse set of relevant behavioral episodes. The broader observational base of traits means that observers are less likely to miss all of the many trait-relevant behaviors than they are to miss a specific performance of a single act. For example, it would be perfectly plausible for some judges to miss an instance of the specific act of "interrupting someone." It is less plausible that a judge will miss all disagreeable behaviors in the situation (including, among others, "interrupting someone," "loudly correcting someone's mistake," and "insisting on having the last word"). The present findings indicate higher observer–observer agreement for acts than for trait ratings by peers, thus suggesting that the greater observability of acts may outweigh the greater breadth of traits in determining agreement among observers.

In addition, act and trait reports may differ because they derive from two different forms of memory. Specific behaviors are encoded in episodic memory whereas representations of traits are encoded in semantic memory (Klein & Loftus, 1993). Consequently, judgments about acts require recall of specific behavioral instances (i.e., episodic memory) and are likely to proceed through a different cognitive process than judgments about traits, which require retrieval of abstract, generalized information about a person (i.e., semantic memory). One implication of this distinction is that self-perception bias may occur either at the initial stage of encoding behavior into episodic memory or at the stage when memories of specific acts are generalized into semantic knowledge as trait representations (e.g., by selectively attending to desirable episodic memories). The present findings imply the former—that act perceptions themselves are biased. Thus, self-judgments about traits may be biased just because self-judgments about acts are biased. Clearly, however, our findings do not exclude the possibility that bias also exists when semantic knowledge about the self is formed. In summary, understanding the processes by which perceptions of act occurrences are translated into trait judgments will help elucidate the factors that cause accuracy and bias in personality impressions.

A research program on the process of act monitoring should be linked to models of interpersonal perception. For example, Kenny's (1994) model addresses why individuals agree (or disagree) with one another in their perceptions of others. Kenny proposed nine parameters that contribute to agreement (or consensus) between judges, including acquaintance, information overlap, and similar meaning systems. Thus, we have two accounts of why judges agree with one another: Kenny's model, which points to aspects of the perceiver and the context of perception, and the present findings, which point to properties of the acts under scrutiny. To what extent do these two accounts overlap?

We suggest that the act property of observability is related to Kenny's (1994) "similar meaning system" (the extent to which an act is given the same meaning by two perceivers). Highly observable acts tend to require less inference to judge their occurrence than do less observable acts. Thus, the more observable an act is, the more likely will judges attach the same meaning to it. Conversely, judges will be more likely to disagree about the meaning of an act requiring a great deal of inference about the target's internal thoughts and feelings. Thus, according to Kenny's model, observability should indeed be related to interjudge agreement, and the present study is consistent with this prediction. This example illustrates how Kenny's model can be applied to retrospective reports and on-line codings of acts, and an attempt to integrate Kenny's model with research on the properties of acts (and traits) should prove fruitful (see Robins & John, 1996).

Implications for Act-Based Trait Assessment

The present study has some implications for the feasibility and practice of act-based personality assessment using both on-line and retrospective act frequency reports. Our findings for on-line act reports showed levels of interobserver agreement that were reasonably high for the majority of acts, and, indeed, slightly higher than that obtained for trait ratings. These results support the feasibility of this fundamental mode of act-based trait assessment. Furthermore, Borkenau and Ostendorf (1987) studied a situation similar to that used in this research and found substantial accuracy for retrospective observer act reports. Finally, it is important to keep in mind that our findings focus on reports of single acts and do not benefit from aggregation across acts. Thus, reliability and validity of both on-line observer and retrospective self-reports would be substantially higher for the multiple-act indices advocated by the AFA (Cheek, 1982).

However, the present findings suggest some limitations of retrospective self-reports as surrogates for on-line codings of act frequency. Although we found some degree of correspondence between self-reports and aggregated on-line act reports by observers, the more salient finding was the great variability in self–observer agreement across acts. For some acts, self-reports appear to correspond with the on-line observer codings (i.e., Extraversion) but for other acts self-reports do not (i.e., Agreeableness). Furthermore, our results indicate that the operation of self-enhancement bias, previously found for trait ratings, cannot be avoided at the act report level. Finally, unlike observer reports, self-reports of acts have the intrinsic limitation that aggregation across "multiple selves" is not possible (Hofstee, 1994).

One interesting question arising from our study is whether act frequency self-reports would be more accurate if they were made on-line rather than retrospectively, allowing individuals to rewind and review a videotape of their behavior (Robins & John, 1997b). Unfortunately, such a procedure would be problematic if these on-line self-reports were found to be influenced by retrospective judgments made immediately after the task. Another promising approach would be to supplement retrospective act reports with more or less on-line act reports using beeper technology (Moskowitz, 1994).

The present findings suggest that some practical challenges remain to be overcome to fully implement the AFA and realize its envisioned theoretical potential. For example, valid retrospective self-reports are difficult to obtain for acts related to Agreeableness, results consistent with those reported by Ozer and Buss (1991). These findings for acts parallel those for trait ratings, thus indicating that the problem may reside not with act monitoring per se but rather with the distinctiveness of self–other perspectives in this behavioral domain. Thus, the implications of these findings pertain not just to AFA assessment methods but more generally to method effects in construct validation (e.g., Ozer, 1989). In particular, researchers should specify what kinds of method effects should be expected given the conceptual definition of the particular trait construct in question.

Many act frequency studies have used "two separate data sources to assess act performance, thus circumventing the limitations of self-report noted by Block" (Botwin & Buss, 1989, p. 989). Of course, any single method is limited; AFA research should continue to use multiple methods to gather act data. Most notably, the range of AFA methods can be expanded by a return to the basic formulation of the AFA, which highlights the formidable task of observing and coding situated acts as they occur in specific everyday contexts (Craik, 1993, in press). The present findings are encouraging: Even for single acts, the on-line ob-

server reports aggregated across four observers had a mean reliability of .69. Act-based dispositional analysis entails an additional aggregation of trends across trait-prototypic acts, which can be expected to improve further the reliability of this form of act-based trait assessment.

In conclusion, a greater understanding of when and why individuals can accurately report what they and others did in a situation should be the goal of further psychological research. Not only can such research inform studies that use observer and self-report methods, but it can also illuminate the processes that underlie disagreements in such domains as romantic relationships, conflict resolution, and negotiation.

References

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

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barker, R. G., & Wright, H. F. (1951). *One boy's day: A specimen record of behavior.* New York: Harper.

Bass, B. (1954). The leaderless group discussion. *Psychological Bulletin, 51,* 465–492.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74,* 183–200.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). New York: Academic Press.

Block, J. (1989). Critique of the act frequency approach to personality. *Journal of Personality and Social Psychology, 56,* 234–245.

Borkenau, P., & Müller, B. (1992). Inferring act frequencies and traits from behavioral observations. *Journal of Personality, 60,* 553–573.

Borkenau, P., & Ostendorf, F. (1987). Retrospective estimates of act frequencies: How accurately do they reflect reality? *Journal of Personality and Social Psychology, 52,* 626–638.

Botwin, M. D., & Buss, D. M. (1989). Structure of act-report data: Is the five-factor model of personality recaptured? *Journal of Personality and Social Psychology, 56,* 988–1001.

Buss, D. M. (1981). *The act frequency analysis of interpersonal dispositions.* Unpublished doctoral dissertation, University of California, Berkeley.

Buss, D. M., & Craik, K. H. (1980). The frequency concept of disposition: Dominance and prototypically dominant acts. *Journal of Personality, 48,* 379–392.

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review, 90,* 105–126.

Buss, D. M., & Craik, K. H. (1987). Acts, dispositions, and clinical assessment: The psychopathology of everyday conduct. *Clinical Psychology Review, 6,* 141–156.

Campbell, J. D., & Fehr, B. (1990). Self-esteem and perceptions of conveyed impressions: Is negative affectivity associated with greater realism? *Journal of Personality and Social Psychology, 58,* 122–133.

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.

Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology, 68,* 1152–1162.

Craik, K. H. (1993). Accentuated, revealed, and quotidian personalities. *Psychological Inquiry, 4,* 278–289.

Craik, K. H. (1997). Circumnavigating the personality as a whole: The challenge of integrative methodological pluralism. *Journal of Personality, 65,* 1087–1111.

Craik, K. H. (in press). The lived day of an individual: A person-environment perspective. In W. B. Walsh, K. H. Craik, & R. Price (Eds.) *New directions in person-environment psychology* (2nd ed.). Hillsdale, NJ: Erlbaum.

Davidson, D. (1980). *Essays on actions and events.* Oxford, England Clarendon Press.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self serving assessments of ability. *Journal of Personality and Social Psychology, 57,* 1082–1090.

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

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

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

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology, 60,* 773–794.

Funder, D. C., & Colvin, C. R. (1997). Congruence of self and others judgments of personality. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 617–647). New York: Academic Press.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52,* 409–418.

Gabriel, M. T., Critelli, J. W., & Ee, J. S. (1994). Narcissistic illusion in self-evaluations of intelligence and attractiveness. *Journal of Personality, 62,* 143–155.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist, 35,* 603–618.

Greenwald, A. G., & Pratkanis, A. R. (1984). The self. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 3, pp. 129–178). Hillsdale, NJ: Erlbaum.

Hampson, S. E., Goldberg, L. R., & John, O. P. (1987). Category breadth and social-desirability values for 573 personality terms. *European Journal of Personality, 1,* 241–258.

Hayes, A. F., & Dunning, D. (1997). Construal processes and trait ambiguity: Implications for self–peer agreement in personality judgments. *Journal of Personality and Social Psychology, 72,* 664–677.

Hendin, H. M., & Cheek, J. M. (1997). Assessing hypersensitive narcissism: A reexamination of Murray's Narcicism scale. *Journal of Research in Personality, 31,* 588–599.

Hofstee, W. K. B. (1994). Who should own the definition of personality *European Journal of Personality, 8,* 149–162.

Hogan, R. (1996). A socioanalytic perspective on the five-factor model In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 163–179). New York: Guilford Press.

Howard, A., & Bray, D. W. (1988). *Managerial lives in transition: Advancing age and changing times.* New York: Guilford Press.

James, W. (1890). *The principles of psychology.* Cambridge, MA: Harvard University.

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). New York: Guilford Press.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61,* 521–551.

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-percep

tion: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology, 66,* 206–219.

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

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.

Klein, S. B., & Loftus, J. (1993). The mental representation of trait and autobiographical knowledge about the self. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *Advances in social cognition* (Vol. 5, pp. 1–49). Hillsdale, NJ: Erlbaum.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64,* 311–337.

Lewinsohn, P. M., Mischel, W., Chaplin, W., & Barton, R. (1980). Social competence and depression: The role of illusory self-perceptions. *Journal of Abnormal Psychology, 89,* 203–212.

Morf, C. C., & Rhodewalt, F. (1993). Narcissism and self-evaluation maintenance: Explorations in object relations. *Personality and Social Psychology Bulletin, 19,* 668–676.

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. (1994). Cross-situational generality and the interpersonal circumplex. *Journal of Personality and Social Psychology, 66,* 921–933.

Newcomb, T. M. (1929). *The consistency of certain extrovert–introvert behavior patterns in 51 problem boys.* New York: Columbia University.

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

Ozer, D. J. (1989). Construct validity in personality assessment. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 224–234). New York: Springer-Verlag.

Ozer, D. J., & Buss, D. M. (1991). Two views of behavior: Agreement and disagreement among marital partners. In D. J. Ozer, J. M. Healy, Jr., & A. J. Stewart (Eds.), *Perspectives in personality* (Vol. 3, pp. 91–106). London: Jessica Kingsley.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology, 60,* 307–317.

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61,* 380–391.

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and some further evidence of its construct validity. *Journal of Personality and Social Psychology, 54,* 890–902.

Robins, R. W., & John, O. P. (1996). Toward a broader agenda for research on self and other perception. *Psychological Inquiry, 7,* 279–287.

Robins, R. W., & John, O. P. (1997a). The quest for self-insight: Theory and research on accuracy and bias in self-perception. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 649–679). New York: Academic Press.

Robins, R. W., & John, O. P. (1997b). Self-perception, visual perspective, and narcissism: Is seeing believing? *Psychological Science, 8,* 37–42.

Rogers, C. R. (1959). A theory of therapy, personality, and interpersonal relations, developed in the client-centered framework. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3, pp. 185–256). New York: McGraw-Hill.

Sedikides, C., & Strube, M. J. (1997). Self-evaluations: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. *Advances in experimental social psychology* (Vol. 29, pp. 209–269). New York: Academic Press.

Shopshire, M. S., & Craik, K. H. (1996). An act-based conceptual analysis of obsessive-compulsive, paranoid, and histrionic personality disorders. *Journal of Personality Disorders, 10,* 203–218.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Taylor, S. E. (1989). *Positive illusions: Creative self-deception and the healthy mind.* New York: Basic Books.

Taylor, S. E., & Brown, J. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193–210.

Taylor, S. E., & Brown, J. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin, 116,* 21–27.

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 181–227). New York: Academic Press.

Thorne, A. (1989). Conditional patterns, transference, and the coherence of personality across time. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 149–159). New York: Springer-Verlag.

Thornton, G. C., & Byham, W. C. (1982). *Assessment centers and managerial performance.* San Diego, CA: Academic Press.

White, P. A., & Younger, D. (1988). Differences in the ascription of transient internal states to self and other. *Journal of Experimental Social Psychology, 24,* 292–309.

Received December 20, 1996
Revision received June 16, 1997
Accepted June 16, 1997 ■