Journal of Personality and Social Psychology
1987, Vol. 52, No. 3, 626-638

Retrospective Estimates of Act Frequencies: How Accurately Do They Reflect Reality?

Peter Borkenau and Fritz Ostendorf
University of Bielefeld, Federal Republic of Germany

Several authors have argued that memory-based reports about dispositional characteristics of people exhibit a correlational structure that is unrelated to the co-occurrences of pertinent on-line recorded behaviors. We hypothesize, however, that the empirical evidence, deemed to be in favor of this challenging hypothesis, is due to the neglect of act overlap among the behavior classes under investigation. The present study was conducted in order to examine this hypothesis. Eight groups, each composed of 6 male actors, were videotaped while discussing controversial topics. The tapes were later shown to five judges who retrospectively estimated, with respect to 16 behavior categories, the individual act frequencies after viewing each discussion in its entirety. Four other judges classified each activity, immediately after observing it, by using one of two coding schemes. Two judges made forced choices; that is, they assigned a given behavior to exactly one category. Two other judges rated the prototypicalities of each activity with respect to each of the 16 behavior categories. The latter coding scheme led to substantially higher correspondences between the correlational structures of the retrospectively estimated and the on-line-recorded act frequencies. The results support a systematic overlap hypothesis rather than a systematic distortion hypothesis.

The accuracy of personality ratings is one of the most controversial issues in psychology. Personologists emphasize the correctness of their subjects' reports about their own or their significant others' trait positions, typical behaviors, feelings, wishes, or attitudes. They usually interpret the major factors that emerge from factor analyses of trait ratings and questionnaires as factors of personality (Cattell, 1946; Eysenck & Eysenck, 1969; Guilford, 1975; McCrae & Costa, 1987; Norman, 1963). Thus, subjects' self-reports or reports by knowledgeable informants are taken as an accurate reflection of reality.

It is not difficult, however, to demonstrate inaccuracies in retrospective reports of person variables. Humans obviously differ in their perception of events, and human memory is far from perfect. Accordingly, reports about behavior do not completely agree with one another. Moreover, they cannot be expected to coincide with reality. Studies about the heuristics and biases of human judgment have yielded systematic differences between scientific rules of inference and the implicit rules used by laymen (Nisbett & Ross, 1980; Ross, 1977; Tversky & Kahneman, 1974). Accordingly, discrepancies between the judgments about persons and the "real state of affairs" seem highly probable. Thus, personologists are confronted with the question of how accurately trait ratings and questionnaire scores reflect reality. However, when discussing the accuracy of personality ratings in more detail, it is crucial to distinguish between the problem of validity and that of structural fidelity (Loevinger, 1957).

Validity of Trait Ratings

Validity, as used here, refers to the relation, across subjects, between a measure of a latent variable and the latent variable itself. Thus, we use the term in the sense of "construct validity" (Loevinger, 1957). However, a problem appears immediately: Which measure should one take as an accurate index of the latent variable? In this respect, it is usually argued that dispositional constructs should predict multiple-act criteria more accurately than single-act criteria, the former being the more appropriate referents for dispositional constructs (Alston, 1975; Buss & Craik, 1983; Herrmann, 1973). This relation was indeed demonstrated in a number of studies with respect to both self-reported (Buss & Craik, 1980, 1981; Fishbein & Ajzen, 1974; Jaccard, 1974) and more objectively recorded behaviors (Aries, Gold, & Weigel, 1983; McGowan & Gormly, 1976; Weigel & Newman, 1976). Buss and Craik (1983, 1985) further elaborated the relations between dispositional constructs and observable behavior. According to their act-frequency approach, traits denote individuals' dispositions to show a high frequency of those acts that are prototypical instances of the respective trait. Their approach encompasses three issues: (a) valid assessment of act frequencies, (b) estimation of the prototypicalities of acts for the trait under study, and (c) formation of multiple-act criteria. We believe that the first of these problems deserves more attention than it has received in the past. Accordingly, a necessary step in the process of validating dispositional constructs is the study of the relation between retrospective frequency ratings and behavior counts. This is the first aim of the present study.

This research was supported by Grant Az. Bo 774/1-1 from Deutsche Forschungsgemeinschaft to Peter Borkenau.

We gratefully acknowledge the contributions of Ulrich Baumgartner, Kristina Herzig, Corinna Rüting, and Angelus Stulik, who painstakingly performed the laborious on-line judgment tasks. We thank Daniele Kammer, Rainer Riemann, Robert A. Wicklund, and two anonymous reviewers for their helpful comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Peter Borkenau, Abteilung Psychologie, Universität Bielefeld, Postfach 8640, D-4800 Bielefeld, Federal Republic of Germany.

Structural Fidelity of Personality Ratings

The *halo effect* was one of the early findings of scientific psychology. Thorndike (1920) encountered suspiciously high correlations between ratings that comprised evaluative connotations. For example, several ability ratings were significantly more highly correlated than comparable objective test scores. Later studies indicated that there are "logical presuppositions in the minds of the raters" (Newcomb, 1931), which are of a more elaborate nature than a purely evaluative halo (Asch, 1946; Newcomb, 1931). In more recent research, at least two factors have been demonstrated to foster halo effects: semantic similarity (Berman & Kenny, 1976; Borkenau, 1986a; Chapman & Chapman, 1967, 1969) and attribute distinctiveness (Hamilton, Dugan, & Trolier, 1985).

However, it is one thing to demonstrate that halo effects *may* occur; it is another matter to claim that halo effects are so pervasive in personality ratings that no individual difference theory of personality is necessary to account for the findings of factor analytic personality research. The latter view was taken by Shweder (1975). Shweder and D'Andrade substantiated this challenging hypothesis by reanalyzing several studies. In all these studies (D'Andrade, 1974; Shweder, 1975; Shweder & D'Andrade, 1980), behaviors had been recorded on-line by using several categories. The numbers of entries for all categories were then correlated across subjects. Thus, intercorrelations were established between on-line-recorded behavior frequencies. Shweder and D'Andrade dealt with these coefficients as if they mirrored the true behavior relations, a standard against which the structural fidelity of memory-based behavior ratings could easily be tested. They did this by comparing the size and the rank order of the correlations among the on-line behavior counts, on the one hand, and by comparing the size and the rank order of the correlations among the retrospective frequency judgments, on the other hand. They found that the intercorrelations of the on-line behavior counts were generally lower and that the structural correspondences between the on-line-recorded and the memory-based behavior frequencies were weak. Moreover, when the semantic similarity relations among the terms descriptive of behavior were also taken into account, stronger relations were found between rating covariations and semantic similarities than between the structures of rated and on-line-recorded behavior frequencies. From this pattern of results, Shweder and D'Andrade derived a systematic distortion hypothesis that began from the premise that, when estimating retrospectively, judges are not able to remember the behavior frequencies with precision. Furthermore, and more important, the errors that the judges make are not random, but point systematically in the direction of the semantic similarity relations among the category descriptors (Shweder, 1982).

There are several implications of this hypothesis. One implication is that there should be a reciprocal relation between the accuracy of retrospective judgments and the amount of systematic distortion, that is, a *reverse error of attenuation* (Thorndike, 1920). This prediction, however, was not confirmed in several studies wherein subjects had to learn and, later on, to remember artificial characters (Berman & Kenny, 1976; Borkenau, 1986a; Cantor & Mischel, 1979). In these studies, it turned out that the correlational error was quite constant and independent of the time elapsed between encoding and recognition. Thus, contrary to D'Andrade's (1974) claim, the duration of the retention interval does not seem to be the crucial variable. Moreover, in the three aforementioned studies, the accuracy of the character descriptions declined with the passage of time, but the correlational error did not increase. Thus, the prediction of a reverse error of attenuation was not confirmed.

But how, then, can the discrepancies in structure between the on-line-recorded behavior frequencies and the memory-based ones be explained? One explanation may be that the on-line-recorded behavior frequencies were less reliable, a point emphasized by Block, Weiss, & Thorne (1979). However, we want to pursue here the *systematic overlap hypothesis*, as suggested by Borkenau (1986b). This hypothesis claims that the intercorrelations among retrospective act-frequency ratings are, in large part, due to meaning overlap among the dispositional categories; the more semantically similar two terms descriptive of behavior are, the more they pertain to overlapping features of personality and, thus, to overlapping act universes. Because D'Andrade (1974) and Shweder (1975) reanalyzed studies of the type wherein each on-line-recorded behavior was mapped to exactly one category immediately after observation, the structural relations attributable to meaning overlap among the behavior categories may have been eliminated. The corresponding correlations may, therefore, have been attenuated (Romer & Revelle, 1984). In the studies by Borkenau (1986b), which investigated classifications of verbally described acts to five traits, the systematic overlap hypothesis was strongly supported. It turned out that the semantic similarity of two trait-descriptive terms was higher the more that the same acts were regarded as instances of both. However, in order to demonstrate that such puzzling results, as reported by Shweder and D'Andrade, emerge from a neglect of act overlap, it is necessary to resort to *on-line* behavior counts in which each act has to be classified with respect to several of the categories under study.

Shweder and D'Andrade's hypothesis is problematic in one additional respect: Whereas the systematic distortion hypothesis is directed at the covariations among *trait* ratings, the studies referred to investigated correlations among *behavior* frequencies. Comparable relations at the trait and behavior level are thereby assumed. This assumption has seriously been challenged by Semin and Greenslade (1985). The present study is directed, for two reasons, at the level of behavior. First, one of its purposes is to show that results such as those reported by Shweder and D'Andrade are encountered only if act overlap among categories is neglected. Therefore, we chose categories at a level of inclusiveness similar to theirs. Second, the relations between the semantic similarities of traits and their degree of act overlap, as predicted by the systematic overlap hypothesis, have already been confirmed in an earlier study (Borkenau, 1986b). Therefore, we now intend to verify these relations for categories that are descriptive of behavior. This is the second purpose of the present study.

Overview of the Present Study

Eight groups, each composed of 6 male students, were videotaped as they discussed controversial topics. The videotapes were later analyzed, with regard to 16 behavior categories, for the act frequencies of the 48 single participants.[1] Retrospective frequency judgments as well as on-line behavior counts were obtained. The *retrospective ratings* were done after each discussion had been shown, in its entirety, to the judges. The judges were then requested (a) to rank order the 6 discussants with respect to how frequently they had exhibited a behavior that was an example of each of the 16 behavior categories and (b) to estimate the absolute frequency of pertinent behaviors for each discussant and each of the 16 kinds of behavior.

On-line behavior counts were obtained by first subdividing the eight discussions into 15-s units of observation. For each of the 15-s units, it was then decided who of the 6 discussants had spoken during the respective time interval. Using such a combined time and event sampling, we identified 3,696 activities of single actors, that is, an average of 77 activities per discussant. These activities were subsequently assigned to the 16 classes of behavior with the use of two coding schemes.

A simple classification coding scheme was applied in order to demonstrate the replicability of the results reported by Shweder and D'Andrade and to establish a standard against which the second on-line coding scheme could be compared. Two independent judges had to classify each of the 3,696 activities into one (and only one) of the 16 classes of behavior under study. Thus, for example, the judges were not allowed to assign a behavioral sequence into the category *criticizes* and to the category *contradicts* or *disapproves*. Accordingly, no act overlap was possible, even for very similar categories.

Two other independent judges rated the 3,696 activities by using a *prototypicality coding scheme* that allowed for meaning overlap among the behavior categories. They were presented with the eight discussions 16 times each (i.e., once for each category). For each presentation, they were instructed to judge the prototypicality of the 3,696 activities with respect to the one category at issue. It was expected that, with respect to the 16 categories, the prototypicality ratings would be positively intercorrelated across activities for semantically similar categories (e.g., criticizes vs. contradicts) and negatively intercorrelated for dissimilar categories (e.g., criticizes vs. agrees). Consequently, we expected that the structural correspondences between the on-line-counted act frequencies and the retrospectively estimated ones would be more pronounced when the prototypicality coding scheme (compared with the simple classification coding scheme) was applied in the scoring of the single discussants' activities.

Method

Semantic-Similarity Judgments

Twenty students (10 female, 10 male) judged the semantic similarities among 16 behavior categories with the use of a 7-point rating scale (−3 = *antonyms* and 3 = *synonyms or near synonyms*). The 120 trait pairs were randomly presented on a video screen to each judge.

Behavior Setting

Male actors who agreed to be videotaped were recruited by a leaflet distributed to all students at the University of Bielefeld, Federal Republic of Germany. Students interested in participating in the study were asked to choose one topic from among six, all of which were controversial in the general public at the time the study was conducted (e.g., speed limits on German highways), and to indicate their attitude with respect to this topic. On the basis of this information, the experimenter, in order to stimulate lively debates, formed groups of 6 participants who advocated different opinions. The payment of the actors was made dependent on the quality of their discussion. All participants received $4 for their participation; additionally, they could earn $12 if their group provided the best debate among the eight groups under study. The actors were instructed to discuss their respective viewpoints during the initial 25 min and to use the following 25 min for jointly writing a memorandum of agreement. This memorandum was to summarize the aspects of the problem about which they agreed. Each of the 50-min sessions was tape-recorded in its entirety and later analyzed.

The discussants were seated on two sides of a square table so the faces of all actors could be videotaped during the whole session. A name card with a pseudonym was placed in front of each actor in order to provide observers with an identification of each discussant. The same six pseudonyms were used for each of the eight groups. After about 50 min, the discussion was interrupted by the experimenter.

Coding of the Behavior Sequences

Retrospective ratings. Five student observers (2 female, 3 male), unacquainted with the actors, viewed each of the eight discussions in its entirety in a different random order. Before viewing the first tape, they were informed about the details of their rating task. Then, after presentation of each 50-min discussion, they were instructed to rank order the 6 actors with respect to the frequencies with which they had exhibited each of the 16 classes of behavior. To this end, they had to write the six pseudonyms in the appropriate order. The instructions were as follows:

> Now, please judge the behavior of the single discussants. To this end, you are provided with a list of terms useful in describing people's behavior in a discussion. An example could be "interrupts the speech of another participant." Your task is to remember how frequently each of the participants has acted in a way that may be appropriately described by this term. Afterward, please rank order the discussants, indicating which of the 6 participants has shown corresponding behavior most frequently, who is in the second place, etc.

To encourage the independence of individual judgments, each of the 16 behavior categories was then presented on a different page of a booklet. At the top of each page, the type of behavior currently at issue was written, followed by the phrase "This behavior was shown by" and six empty lines, one for each of the pseudonyms of the 6 discussants. These six lines were designated as "most frequently," "second place," and so on. The order of the behavior categories was randomized and, therefore, differed for each judge. When the raters had completed this booklet, it was collected, and another one was distributed. The second booklet asked the judges "to indicate how frequently the single participants have acted in the respective way." A different behavior category was written at the top of each page, followed by the question "How frequently has

[1] The English translations of the rating categories are reported in several tables. The categories were selected according to the following criteria. First, near synonyms and antonyms had to be incorporated in order to demonstrate meaning overlap among categories. Second, in order to minimize assignments to a residual category, each reasonable activity that could be expected to occur during a discussion should be classifiable to at least one category. It is reasonable to suppose that Bale's coding scheme would have satisfied the first criterion more than the second.

The German terms used were unterstützt, greift Beiträge anderer auf, scherzt, vermittelt, sucht Ausgleich, stimmt zu, schlägt vor, leitet die Diskussion, kritisiert, informiert, erklärt, schweift vom Thema ab, fragt nach Meinungen, widerspricht, lehnt ab, and macht lächerlich.

each of the participants shown the corresponding behavior during the discussion?" The pseudonyms of the 6 discussants were listed below this. The judges indicated their estimate by writing a number following each name. Thus, they literally guessed the behavior frequencies. Once again, the 16 categories were presented on separate pages, the order of which was randomized for each judge.

Simple classifications. Two student judges (1 female, 1 male) viewed the eight discussions in a different random order. They were presented with tapes on which the discussions had been subdivided into 15-s units of observation. Each 15-s sequence was followed by a 10-s still. The judges were instructed to stop the tape when a still appeared on the screen and to judge the behavior of the discussants who had been verbally active during the preceding 15 s. Afterward, they were to restart the video recorder and view the next sequence until the next still appeared and so on.

The judgments were made in booklets. The single scenes were numbered and, for each consecutive scene, one or several judgments had to be given depending of the number of active discussants. Which discussants were regarded as verbally active during a sequence had been agreed upon by one of the authors and a student. Thus, each judgment referred to the verbal activity of a specified actor during a specified period. In the simple classification task, the judges were asked, "Which of the following categories is most appropriate to classify the behavior of Frank?" or "Which of the following categories is most appropriate to classify the behavior of Peter?" and so forth. Each of these questions was followed by a list of the 16 behavior categories plus the residual category, *no judgment possible.* The residual category was included because it was sometimes impossible to classify the behavior of someone having spoken for an extremely short period. Altogether, each of the two judges classified 3,696 activities in this way during a period of 3 weeks.

Prototypicality ratings. Two student judges (1 female, 1 male), unacquainted with the discussants and the hypothesis of the study, provided the prototypicality ratings. The videotapes they viewed were the same as those presented to the raters who performed the simple classifications. However, they had to view the tapes 16 times each. The two judges were administered the 128 combinations of categories and discussion groups in different random orders.

For each presentation of a discussion, the judges received a booklet in which the sequences were numbered, and the actors whose behavior had to be judged were specified. The judges were instructed "to indicate how well the given behavior may be characterized by the category at issue" by using 7-point rating scales (3 = *very good example,* and −3 = *blatant counterexample*). Thus, in contrast to the scales applied by Buss and Craik, the scales used in the present study included the notion of counterexamples. We used such scales because of our experience with unipolar prototypicality ratings. When such scales were used in earlier studies, our subjects complained about problems with distinguishing counterexamples from unrelated activities. Accordingly, in the present study, the scale digits were explained as follows.

3 = The behavior of the discussant is a very good example for the category at issue.
2 = The behavior of the discussant is an example for the category at issue.
1 = The behavior of the discussant is more an example than a counterexample for the category at issue.
0 = The behavior of the discussant is neither an example nor a counterexample for the category at issue.
−1 = The behavior of the discussant is more a counterexample than an example for the category at issue.
−2 = The behavior of the discussant is a counterexample for the category at issue.
−3 = The behavior of the discussant is a blatant counterexample for the category at issue.

Altogether, each rater made 16 judgments for each of the 3,696 activities, that is, 59,136 judgments altogether. This task was performed during a period of about 6 months.

Table 1
Means and Standard Deviations of the Retrospective-Frequency Estimates and Reliabilities of the Frequency Estimates and Rankings

Behavior class	*M* frequency estimate	*SD*	Reliability of frequency estimates	Reliability of frequency rankings
Supports	3.09	1.70	.60	.67
Takes up a contribution	4.86	2.34	.71	.79
Jokes	2.02	1.94	.59	.84
Mediates	2.30	1.79	.59	.78
Seeks arrangements	2.51	1.72	.61	.74
Agrees	3.88	1.79	.60	.60
Proposes	4.64	2.23	.67	.86
Directs the discussion	3.42	3.01	.84	.90
Criticizes	3.55	2.35	.74	.80
Informs	4.28	2.10	.72	.90
Explains	4.49	2.41	.66	.88
Changes the subject	1.33	0.99	.55	.72
Asks opinions	2.06	1.44	.58	.82
Contradicts	3.29	2.29	.64	.84
Disapproves	2.96	2.16	.63	.82
Ridicules	1.12	1.11	.52	.83

Note. The reliability of the judgments was estimated by intraclass correlations (*ICC*[2, 5], according to Shrout & Fleiss, 1979). All statistics were first computed for each discussion group and then averaged across the eight groups.

Results

Overall Activity of Single Actors

The number of verbal activities differed markedly among the single discussants. Their average number per participant was 77 (*SD* = 37.12). The most active discussant made 159 contributions, whereas the least active discussant spoke only 9 times. Within each discussion group, the minimum difference between the most active and the least active actor was 74 verbal activities. Thus, one important difference among actors was their overall activity; some participants filled the discussion by talking at least twice each minute, whereas others were silent for considerable periods. Accordingly, mainly positive intercorrelations across actors of the act frequencies for the various behavior classes were expected.

Retrospective Ratings

The frequency rankings were scored such that the value of 6 indicated the person with the highest frequency, and the score of 1 indicated the person with the lowest frequency of the behavior at issue. The rankings and the frequency estimates were averaged across the 5 retrospective judges to increase the reliability of these scores. The reliabilities of the retrospective judgments were then estimated with the use of intraclass correlations (*ICC*[2, 5], according to the taxonomy by Shrout & Fleiss, 1979). These coefficients are reported, together with the means and standard deviations of the frequency judgments, in Table 1.

The absolute frequency estimates differed from the frequency rankings in that the latter did not reveal any systematic differences between the eight discussion groups. For the frequency

Table 2
Correlations Among Retrospective-Frequency Rankings (Above Diagonal) and Among Retrospective-Frequency Estimates (Below Diagonal)

Behavior class	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	—	.80	.65	.81	.83	.82	.70	.60	.34	.52	.58	.28	.67	.33	.15	.53
B	.91	—	.73	.72	.71	.69	.83	.82	.70	.87	.89	.53	.75	.71	.58	.74
C	.60	.74	—	.53	.55	.65	.50	.54	.40	.65	.60	.70	.46	.52	.49	.89
D	.80	.70	.55	—	.91	.68	.74	.71	.35	.46	.61	.07	.83	.31	.12	.47
E	.87	.72	.66	.94	—	.69	.71	.61	.30	.46	.66	−.05	.81	.17	−.07	.35
F	.88	.91	.78	.78	.78	—	.53	.39	.05	.41	.42	.28	.54	.28	.10	.47
G	.78	.92	.70	.62	.62	.81	—	.92	.71	.80	.86	.30	.81	.66	.54	.51
H	.74	.89	.71	.75	.72	.81	.97	—	.77	.90	.88	.50	.79	.82	.70	.65
I	.53	.84	.60	.28	.27	.62	.93	.88	—	.74	.75	.55	.44	.81	.84	.58
J	.72	.89	.72	.55	.53	.78	.96	.92	.89	—	.93	.59	.66	.82	.80	.68
K	.67	.86	.65	.56	.54	.77	.94	.92	.88	.95	—	.49	.78	.74	.68	.60
L	.33	.66	.67	.20	.26	.60	.60	.57	.63	.71	.73	—	.08	.55	.64	.80
M	.73	.74	.63	.89	.86	.68	.75	.82	.57	.68	.62	.25	—	.54	.29	.50
N	.55	.82	.58	.35	.40	.68	.85	.84	.94	.82	.88	.62	.54	—	.91	.74
O	.42	.78	.53	.22	.21	.54	.84	.83	.96	.82	.85	.64	.45	.94	—	.73
P	.54	.73	.84	.52	.44	.74	.79	.68	.77	.78	.75	.66	.55	.78	.73	—

Note. A = supports; B = takes up a contribution of another participant; C = jokes; D = mediates; E = seeks arrangements; F = agrees; G = proposes; H = directs the discussion; I = criticizes; J = informs; K = explains; L = changes the subject; M = asks opinions; N = contradicts; O = disapproves; P = ridicules. The retrospective ratings were averaged across the five judges and then intercorrelated.

rankings, the mean for each discussion group and each behavior category was obviously 3.5. However, after analyzing the absolute frequency estimates by a multivariate analysis of variance (MANOVA), it turned out that the retrospective judges perceived significant systematic differences among the eight discussion groups, $F(112, 171) = 3.30, p < .001$. These differences could not be revealed by the frequency rankings. Accordingly, frequency rankings for actors who had been members of different discussion groups could not be compared. If statistics had been computed across all 48 discussants, systematic differences between groups would have constituted a source of error variance. Therefore, all analyses that involved retrospective judgments were performed separately for each of the eight discussion groups, and the resulting statistics were averaged afterward. Thus, for example, each reliability reported in the last two columns of Table 1 was the mean of eight single coefficients. Fisher's Z transformation was used to calculate averages of correlation coefficients.[2] The average reliability of the frequency estimates for the 16 behavior classes was .65, and the average reliability of the frequency rankings was .81. Thus, the former were less reliable than the latter; the judges agreed less about the absolute frequencies of the single discussants performing the 16 types of behavior than about the rank order of the discussants with respect to these frequencies. Moreover, the estimated frequencies differed markedly for the 16 categories, the most frequent class of behaviors (i.e., takes up the contribution of another participant) being more than 4 times as frequent as the least frequent one (i.e., ridicules).

The intercorrelations among the retrospectively estimated frequencies of the 16 classes of behavior are reported in Table 2. With few exceptions, these correlations are positive in sign. This lack of negative correlations may accurately reflect the high variance in the overall activity of the single participants.

On-Line Behavior Counts

Because two parallel judgments were available for each type of on-line behavior coding, separate analyses were performed for each single judge to demonstrate the replicability of the results across two raters.

The reliability of the assignments of the 3,696 activities to the 16 categories plus the residual one (i.e., no judgment possible) in the simple classification task was estimated by using Cohen's κ. This reliability turned out to be .30. The reliabilities of the prototypicality ratings were estimated by using intraclass correlations (*ICC*[2, 1], according to Shrout & Fleiss, 1979). The resulting coefficients for the single categories are listed in the last column of Table 3. Averaged across the 16 categories, the reliability of the prototypicality ratings of a single judge amounted to .42, a result that was, by and large, in agreement with those obtained when verbally described acts were judged for their prototypicalities (Borkenau, 1986b; Buss & Craik, 1983).[3]

Table 3 also reports the number of activities assigned by each judge to the 16 behavior classes and, separately for the two judges, the means and the standard deviations of the prototypicality ratings across the 3,696 activities. As may be seen from Table 3, the two judges who assigned each activity to only one of the behavior categories agreed in choosing some categories more often than others; the correlation between the first two columns was .83. Moreover, for each of the two judges, the cate-

[2] Because the reported coefficients are the average of several *raw* coefficients, no conventional tests of statistical significance can be applied. A similar problem arises for the comparisons that involve correlations between correlation matrices. Because the single entries in the correlation matrices are mutually dependent, the mathematical assumptions of the statistical tests of significance are not met. For this reason, no significance tests for correlation coefficients are reported throughout this article. Instead, parallel indepeñdent analyses were carried out in order to show the replicability of the results.

[3] The mean of the prototypicality ratings of both judges was, of course, more reliable. Averaged across the 16 categories, its reliability was .60.

gory chosen most frequently (i.e., explains) was chosen about 25 times more often than the least frequent class (i.e., mediates). Thus, the 16 categories differed more among one another with respect to the frequencies with which they were chosen in the on-line simple assignments than they did in the retrospective frequency ratings. The latter were more evenly distributed; that is, the judges who did the on-line simple assignments differentiated more in this respect than did the retrospective judges. However, both rating tasks led to comparable results with respect to the rank order of the frequencies of the single categories; the correlations across categories between the first column of Table 1 and the first two columns of Table 3 equaled .66 and .75, respectively. Moreover, the higher the retrospectively estimated average frequency for a behavior class, the higher was the average prototypicality of the 3,696 activities for this category (for the female judge, r = .66, and for the male judge, r = .87). Thus, quite independent of the type of frequency estimate, it turned out that some types of behavior were perceived as occurring more often than others.

Validity of Retrospective-Frequency Estimates

How accurately do the retrospectively estimated behavior frequencies of the single actors reflect the on-line-recorded ones for the same behavior class? A definite answer to this question could be given if there were an agreed-upon definition of what an on-line-recorded behavior frequency is. However, there is no such generally agreed-upon prescriptive model. Therefore, three models for scoring the activities of the single actors were applied and compared. Model 1 used the assignments of the 3,696 activities to only one category. If one of the 48 actors had performed an activity assigned to the category at issue in this forced-choice task, this was counted as 1 point. Because this assignment task had been performed by two independent judges, two parallel frequency scores were calculated in this way for each actor and each class of behavior.

The other two models used the prototypicality ratings of the 3,696 activities, each judged for all 16 behavior classes. Model 2 weighted each activity of an actor according to the prototypicality ratings by the two respective judges. Once more, independent analyses were performed for each judge. Note that for the prototypicality ratings a scale had been used that allowed for positive as well as negative scores. Accordingly, negative-act-frequency summaries were possible and occurred, as may be inferred from the average prototypicality ratings reported in Table 3. Finally, a third model was introduced wherein the prototypicalities were transformed by adding the value 4. Accordingly, the smallest prototypicality score became 1 and the highest score became 7. The acts shown by the 48 actors were then weighted by the resulting prototypicality score for the category at issue and summarized over all activities shown by the respective actor. One aspect of this scoring model was that each verbal activity of an actor led to an increase of all of his 16 act-frequency scores, because each activity received a weight of at least 1 for each behavior category. This model was incorporated because it yielded the highest correlations with the retrospective frequency estimates and rankings. Once again, separate analyses were performed for each of the two judges who had performed the prototypicality ratings. This model is here referred to as Model 3.

Correlations with absolute-frequency estimates. The correlations between the absolute frequency estimates and the online behavior counts, scored according to the three models already mentioned, are reported in Table 4. Note that each reported correlation is the average of eight single coefficients, calculated for the eight single discussion groups.

As may be seen from Table 4, the average correlations were substantial for all three models. Furthermore, they were highest for Model 3, in which they approached their theoretical maximum given the limited reliability of the retrospective-frequency ratings (see Table 1). Model 2 differed from Model 1 not so much with respect to the average correlations as with respect to a higher variance of the coefficients; substantial *negative* correlations were encountered only when Model 2 was applied. One origin for this pattern of results becomes evident when one looks at the correlations of the retrospective-frequency ratings with the number of activities of the single actors. These correlations, which are reported in the last column of Table 4, were substantial for each single-behavior class. Thus, the general activity of an actor was a strong predictor of his retrospectively estimated behavior frequencies for each single category. Accordingly, the act-frequency summaries, computed according to the best predicting model (Model 3), primarily reflected the general activity of the single actors. The intercorrelations among these act frequencies ranged from .980 to .999. Factor analyses of the Model 3 scores led to overwhelming strong general factors, which accounted for 96.8% of the total variance, when the prototypicality ratings of the female judge were applied. When the ratings of the male judge were used as weights, the proportion of variance accounted for by the first factor was 97.7%. Thus, the scores calculated according to Model 3 showed such high *validities* because they revealed hardly anything except general activity. The retrospective judges seem to have been highly sensitive to differences between the actors' overall activity. Accordingly, the correlations between the retrospective-frequency judgments and the act-frequency summaries, calculated from Model 2, were higher for the behavior classes with the higher mean prototypicality ratings; the correlation across categories between the third column of Table 3 and the second column of Table 4 is .89. The respective correlation for the male judge is .88, as may be verified from a comparison of the fifth columns of both Table 3 and Table 4. Thus, it may be stated that the validity coefficients for the single act-frequency summaries, calculated according to Model 2, were highly influenced by the degree to which they reflected general activity.

Correlations with frequency rankings. Whereas Table 4 reported the validity coefficients for the retrospective absolute-frequency estimates, those for the frequency *rankings* are displayed in Table 5.

The correlations in Table 5 are somewhat lower than those in Table 4, despite the higher reliability of the rankings (see Table 1). Hence, the absolute-frequency estimates contained some valid information in addition to that included in the frequency rankings. In all other respects, however, the results for the absolute-frequency estimates and the frequency rankings were quite similar.

Structural Correspondences

One purpose of the present study was to replicate the relations reported by Borkenau (1986b) among the semantic sim-

Table 3
Assignments to Categories and Prototypicality Ratings for 3,696 Activities

			Prototypicality ratings				
	Assignments		Female judge		Male judge		
Behavior class	Female judge	Male judge	*M*	*SD*	*M*	*SD*	Rater agreement
Supports	96	232	−0.09	1.03	0.05	1.02	.47
Takes up a contribution	118	391	0.52	1.03	0.53	0.93	.30
Jokes	72	78	0.07	0.69	−0.36	0.81	.34
Mediates	22	30	0.07	0.91	0.05	0.79	.44
Seeks arrangements	52	38	0.09	1.11	−0.09	0.89	.42
Agrees	539	468	−0.20	1.16	0.13	0.98	.48
Proposes	503	512	0.50	1.02	0.50	1.02	.50
Directs the discussion	24	120	0.23	0.70	0.28	0.70	.48
Criticizes	484	254	0.39	1.05	0.17	1.02	.53
Informs	424	141	0.49	0.93	0.70	1.09	.48
Explains	764	758	0.40	0.81	0.74	1.00	.39
Changes the subject	33	97	−1.47	1.13	0.04	0.94	.16
Asks opinions	204	83	−0.20	0.97	−0.22	1.00	.39
Contradicts	197	243	0.43	1.29	0.00	1.01	.47
Disapproves	30	53	0.20	1.11	−0.06	1.06	.50
Ridicules	34	50	0.05	0.31	−0.48	0.88	.11
No judgment possible	100	148	—	—	—	—	—

Note. Rater agreement was estimated by using an intraclass correlation (*ICC*[2, 1], according to Shrout & Fleiss, 1979).

ilarities of trait-descriptive terms, their interchangeability (as indicated by the number of cross-classifications among judges), and the intercorrelations of prototypicality ratings among these traits across activities. Whereas the strong relations reported by Borkenau (1986b, Study 1) had been found for verbally described acts that were classified as traits, videotaped activities had to be assigned to behavior categories in the present study.

Semantic similarities. The reliability of the semantic-similarity judgments was estimated by using an intraclass coefficient (*ICC*[2, 20], according to Shrout & Fleiss, 1979). For the judgments averaged over the 20 judges, the reliability turned out to be .94. These averaged semantic-similarity judgments are reported in Table 6.

Cross-classifications. A proportion of cross-classifications

Table 4
Correlations of Act-Frequency Summaries With Retrospective-Frequency Estimates

	Female on-line judges			Male on-line judges			
Behavior class	M1	M2	M3	M1	M2	M3	General activity
Supports	.39	.15	.83	.61	.38	.82	.81
Takes up a contribution	.65	.77	.92	.85	.80	.93	.92
Jokes	.08	.59	.76	.46	−.46	.77	.75
Mediates	.23	.50	.58	.36	.52	.57	.55
Seeks arrangements	.43	.59	.62	.43	.16	.60	.57
Agrees	.75	.03	.84	.76	.62	.85	.84
Proposes	.72	.74	.94	.74	.89	.95	.94
Directs the discussion	.57	.86	.94	.86	.92	.92	.90
Criticizes	.73	.64	.87	.40	.41	.87	.86
Informs	.77	.84	.91	.67	.87	.91	.90
Explains	.82	.89	.94	.87	.96	.94	.93
Changes the subject	.48	−.44	.66	.42	.45	.66	.64
Asks opinions	.69	.04	.63	.50	.45	.64	.62
Contradicts	.65	.58	.87	.84	.02	.86	.85
Disapproves	.71	.49	.82	.45	.08	.83	.78
Ridicules	.43	.74	.76	.50	−.62	.73	.75
Average correlation	.60	.57	.84	.65	.53	.84	.82

Note. M1, M2, and M3 indicate Models 1, 2, and 3, respectively. All statistics were first computed for each discussion group and then averaged across the eight groups.

Table 5
Correlations of Act-Frequency Summaries With Retrospective-Frequency Rankings

	Female on-line judges			Male on-line judges			
Behavior class	M1	M2	M3	M1	M2	M3	General activity
Supports	.40	.27	.65	.54	.48	.65	.62
Takes up a contribution	.54	.72	.81	.69	.83	.82	.79
Jokes	.30	.64	.68	.45	−.40	.68	.66
Mediates	.32	.50	.65	.20	.58	.63	.60
Seeks arrangements	.36	.74	.51	.22	.47	.49	.45
Agrees	.59	.22	.48	.65	.39	.47	.45
Proposes	.76	.76	.83	.81	.79	.84	.83
Directs the discussion	.44	.75	.94	.73	.83	.94	.92
Criticizes	.60	.54	.72	.39	.44	.72	.72
Informs	.78	.83	.93	.65	.89	.92	.91
Explains	.67	.79	.91	.78	.84	.91	.90
Changes the subject	.42	−.37	.60	.41	.38	.59	.57
Asks opinions	.63	.37	.54	.53	.55	.56	.57
Contradicts	.67	.68	.78	.74	.13	.77	.75
Disapproves	.57	.40	.76	.58	−.03	.75	.75
Ridicules	.49	.62	.69	.55	−.47	.68	.68
Average correlation	.55	.57	.76	.58	.50	.75	.73

Note. M1, M2, and M3 indicate Modes 1, 2, and 3, respectively. All statistics were first computed for each discussion group and then averaged across the eight groups.

Table 6
Semantic-Similarity Relations Among the 16 Categories

Behavior class	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	—															
B	1.10	—														
C	−0.50	0.00	—													
D	1.20	1.50	0.25	—												
E	0.75	0.60	0.50	2.30	—											
F	2.25	0.95	−0.35	0.90	0.40	—										
G	0.55	0.55	−0.65	1.50	0.95	−0.15	—									
H	0.75	1.45	−0.65	2.05	1.35	−0.10	1.20	—								
I	−0.80	0.10	−0.65	−1.05	−1.10	−1.60	−0.20	1.30	—							
J	0.50	0.35	−1.00	1.30	0.70	−0.35	0.95	−0.45	−0.40	—						
K	1.15	0.60	−0.25	1.65	0.85	0.10	0.65	0.10	0.00	1.90	—					
L	−1.30	−1.45	0.60	−0.85	−0.55	−0.80	−1.00	−0.65	−0.40	−1.20	−1.20	—				
M	0.15	0.20	−0.20	0.65	0.75	−0.10	−0.35	−0.65	−0.50	−0.95	0.60	−1.10	—			
N	−2.15	−0.15	−0.60	−1.50	−1.90	−2.60	−0.45	0.25	1.65	−0.90	−0.75	−0.50	−1.30	—		
O	−2.70	−0.85	−0.40	−1.55	−1.50	−2.85	−1.00	1.05	1.55	−0.90	−1.10	−0.30	−1.10	0.65	—	
P	−1.70	−0.75	0.45	−1.75	−1.60	−1.35	−0.75	0.40	0.30	−1.35	−0.75	0.50	−1.25	−0.40	−1.35	—

Note. A = supports; B = takes up a contribution of another participant; C = jokes; D = mediates; E = seeks arrangements; F = agrees; G = proposes; H = directs the discussion; I = criticizes; J = informs; K = explains; L = changes the subject; M = asks opinions; N = contradicts; O = disapproves; P = ridicules. The entries in the table are the means of the judgments of 20 subjects.

was assessed to determine the degree of interchangeability of the 16 behavior-descriptive terms at issue. The simple assignments of the 3,696 activities to 1 of the 16 behavior classes were used for this purpose. Note that, in this task, the most frequent category had been chosen about 25 times more often than the least frequent one. Therefore, the marginals for the single categories had to be taken into consideration. For each of the 120 combinations of different categories *i* and *j*, the number of activities that had been assigned to class *i* by the male judge *and* to class *j* by the female judge was counted. This figure was then divided by the geometric mean of the numbers of activities assigned to category *i* by the male judge *or* to category *j* by the female judge. Moreover, the comparable score was calculated for the number of activities assigned to class *i* by the female judge and to class *j* by the male judge, and both scores were added. For 17 out of the 120 trait pairs, no cross-classifications were found. The highest proportion (.39) was found for the pair criticizes and explains, the second highest (.36) for the pair jokes and ridicules.

Prototypicality ratings. The prototypicality ratings were intercorrelated among the 16 classes of behavior across the 3,696 activities. Separate analyses were performed for the ratings of the female and the male judge in order to test the replicability of the results across judges. For the female judge, the highest correlation among prototypicality judgments was .65, which was found for the categories contradicts and disapproves. The correlation was lowest ($r = -.71$) for agrees versus contradicts. For the male judge, the highest intercorrelation emerged for supports and agrees ($r = .65$) and the lowest one ($r = -.65$) for supports versus contradicts. When the prototypicalities were first summed over the male and female judges and then intercorrelated across the 3,696 activities, the intercorrelations became even more extreme, ranging from −.79 to .75.

The structural correspondences between the semantic similarities of the behavior-descriptive terms, the proportion of cross-classifications, and the intercorrelations of the prototypicality ratings across activities are summarized in Table 7. Pearson as well as Spearman rank-order correlations are reported.

Table 7 reveals consistent correspondences among the three indices of meaning overlap: The higher the semantic similarity of two behavior-descriptive terms, the higher are the prototypicality ratings intercorrelated across activities. Moreover, the behavior-descriptive terms are then regarded as being more interchangeable, as evidenced by the higher proportion of cross-classifications among judges. Thus, the systematic overlap hypothesis, as outlined by Borkenau (1986b), was confirmed in the present study; the higher the semantic similarity of terms, which are descriptive of behavior, the more they referred to overlapping-act universes.

Structural Fidelity of the Retrospective-Frequency Estimates

Shweder and D'Andrade compared the intercorrelations of on-line-recorded act frequencies, retrospectively estimated act

Table 7
Structural Correspondences Among Semantic Similarities, Proportion of Cross-Classifications, and Intercorrelations of Prototypicalities Across the 120 Combinations of Behavior-Descriptive Terms

Variable	1	2	3	4
1. Semantic similarities	—	.40	.73	.61
2. Cross-classifications	.35	—	.37	.43
3. Prototypicalities: Female judge	.65	.27	—	.89
4. Prototypicalities: Male judge	.45	.20	.77	—

Note. Pearson correlations are given above and rank correlations below the diagonal.

Table 8
Correlations Among the On-Line Behavior Counts Scored According to Model 1

Behavior class	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	—	.08	.12	.09	−.09	.61	.12	−.09	.10	−.01	.11	.42	.06	.34	.22	.32
B	.61	—	−.01	.26	.73	.32	.64	.78	.15	.84	.64	.26	.73	.27	.52	.03
C	.33	.30	—	−.27	−.06	.32	−.40	−.03	−.04	−.05	.30	.75	−.08	−.14	.14	.66
D	.10	.31	.14	—	.21	.26	.20	.14	.07	.02	.07	.00	.17	.50	−.02	−.07
E	.26	.69	.16	.37	—	−.02	.18	.88	.00	.67	.30	.04	.54	.19	.67	.10
F	.84	.72	.32	.05	.36	—	.36	.07	.41	.09	.47	.40	.38	.63	.36	.20
G	.06	.45	−.10	.40	.30	.18	—	.21	.41	.47	.50	.05	.53	.43	−.03	−.07
H	.06	.63	.10	.42	.59	.29	.69	—	−.32	.54	.01	−.03	.67	.03	.53	−.02
I	.22	.38	.02	.14	.44	.14	.36	−.05	—	.26	.57	.28	−.04	.57	.54	.27
J	.10	.38	.00	−.07	.44	.35	.67	.65	−.07	—	.49	−.08	.45	.15	.54	.01
K	.40	.69	.18	.10	.43	.54	.44	.31	.38	.59	—	.31	.21	.51	.63	.35
L	.49	.52	.12	.22	.73	.58	.13	.12	.39	.21	.65	—	.16	.11	.27	.78
M	.01	.32	.14	.19	.03	.12	.42	.72	−.14	.40	.18	−.15	—	.44	.22	.04
N	.26	.79	.26	.19	.44	.58	.41	.57	.37	.47	.60	.19	.49	—	.47	.34
O	.26	.15	.13	−.08	.17	.48	−.20	−.10	.36	−.03	.36	.69	−.01	.32	—	.42
P	.31	.32	.69	.28	.09	.15	−.23	−.04	.05	−.22	.06	.14	.09	.32	.20	—

Note. A = supports; B = takes up a contribution of another participant; C = jokes; D = mediates; E = seeks arrangements; F = agrees; G = proposes; H = directs the discussion; I = criticizes; J = informs; K = explains; L = changes the subject; M = asks opinions; N = contradicts; O = disapproves; P = ridicules. Scores based on the assignments by the female or male judge are reported above or below the diagonal, respectively. Correlations were computed for each discussion group and then averaged.

frequencies, and the semantic-similarity relations among the category descriptors. The intercorrelations among the retrospective-frequency estimates, obtained in the present study, are reported in Table 2. The semantic similarities are reported in Table 6. The intercorrelations among the act frequencies, scored according to Models 1 and 2, are reported in Tables 8 and 9. Because the intercorrelations among all act frequencies, scored according to Model 3, were beyond .98, the appropriate correlation matrix is omitted.

However, all three models for calculating act frequencies were incorporated into a Shweder-type analysis. Separate analyses were performed for the male and female on-line judges. Moreover, Pearson as well as Spearman rank correlations were calculated. The resulting structural correspondences are reported in Table 10.

As may be seen from Table 10, the scoring of the activities according to Model 1 by and large replicated the results reported by Shweder and D'Andrade, who also found quite low structural correspondences for memory-based and on-line-recorded act frequencies. Moreover, when Model 1 was used, the relations between intercorrelations of act frequencies and semantic similarities were weak in the present study, as they were in Shweder and D'Andrade's data. However, when the activities were scored on-line, using Models 2 or 3, the structural correspondences increased substantially with respect to both the intercorrelations of memory-based ratings and the semantic-similarity relations. Model 2 fared somewhat better than Model 3 in these respects. The correspondences between retrospective-frequency ratings and semantic similarities (not mentioned in Table 10) were .60 for the frequency rankings and .68 for the absolute-frequency estimates. These figures are also quite similar to those reported by Shweder and D'Andrade.

Discussion

The results of the present study may be summarized as follows. First, the retrospective judges estimated very accurately the differences with respect to the base rates of the 16 classes of behavior. The more activities had been assigned to a category or the higher the mean of the prototypicality ratings had been for a class of behaviors, the higher was its retrospectively estimated absolute frequency. Second, the retrospective judges estimated the overall frequency of the most frequently displayed kind of behavior to be about 4 times that of the rarest sort of behavior. In contrast, the judges who assigned each activity online to one of the 16 classes of behavior used the most frequent category about 25 times more often than the least frequent one. Third, the on-line-recorded act frequencies correlated somewhat higher with the retrospectively estimated absolute frequencies than they did with the frequency rankings. This result emerged in spite of the somewhat higher reliability of the frequency rankings. Fourth, the retrospective judges were highly accurate in rank ordering the actors according to their general activity. The stronger the distinct act-frequency summaries reflected the general activity of the single actors, the higher was the correlation between on-line-recorded and retrospectively estimated act frequencies. Fifth, concerning the realm of structural fidelity, it turned out that the relations among semantic similarities, intercorrelations of prototypicality ratings, and probabilities of cross-classifications were substantial. The more that two behavior-descriptive terms were similar in meaning, the more they classified overlapping sets of activities; substantial correspondences were found between the semantic similarities of pairs of trait-descriptive terms and the intercorrelation of the prototypicality ratings with respect to these traits. Finally, it turned out that findings such as those reported by Shweder and D'Andrade were only encountered when a key was applied for the scoring of act frequencies that did not consider meaning overlap among the behavior categories. But if scoring keys were applied that took meaning overlap into account, the structure of the behavior frequencies, judged on-line, approached the semantic-similarity structure as well as the structure of the ret-

Table 9
Correlations Among the On-Line Behavior Counts, Scored According to Model 2

Behavior class	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	—	.16	.29	.78	.57	.81	−.02	−.30	−.55	−.05	−.12	.01	.54	−.76	−.84	−.19
B	.27	—	.52	.41	.44	.23	.56	.13	.24	.65	.65	−.73	−.28	.05	−.31	.40
C	−.22	−.68	—	.41	.47	.38	.06	.03	−.30	.24	−.02	−.17	.13	−.14	−.44	.54
D	.56	.37	.01	—	.91	.78	.53	.36	−.44	.50	.28	−.53	.32	−.60	−.66	−.10
E	.48	−.22	.22	.73	—	.52	.68	.51	−.34	.51	.32	−.62	.46	−.33	−.45	−.11
F	.89	.64	−.44	.56	.33	—	.04	−.33	−.72	−.09	−.18	.11	.60	−.89	−.87	−.08
G	.08	.68	−.80	.49	.13	.39	—	.66	.39	.71	.81	−.86	−.23	.07	.08	.05
H	.11	.48	−.52	.50	.25	.25	.78	—	.34	.77	.65	−.69	−.09	.43	.39	.36
I	−.79	.14	−.43	−.23	−.45	−.60	.41	.33	—	.64	.65	−.55	−.68	.75	.74	.31
J	.20	.79	−.89	.32	−.07	.49	.87	.61	.48	—	.91	−.82	−.29	.21	.09	.15
K	.21	.87	−.87	.29	−.24	.53	.88	.64	.48	.97	—	−.79	−.38	.29	.18	.18
L	−.32	−.03	.16	−.55	−.41	−.24	−.03	−.52	.27	.12	.09	—	.36	−.29	−.12	−.19
M	−.08	−.21	.47	.21	.19	−.34	−.40	.40	−.05	−.55	−.62	−.40	—	−.66	−.51	−.57
N	−.92	−.21	.28	−.44	−.49	−.84	−.15	−.26	.66	−.20	−.13	.27	.25	—	.81	.32
O	−.93	−.30	.27	−.45	−.45	−.93	−.14	−.10	.82	−.22	−.25	.27	.12	.90	—	.03
P	.06	−.72	.94	−.01	.43	−.23	−.80	−.54	−.60	−.88	−.91	.16	.21	−.01	−.01	—

Note. A = supports; B = takes up a contribution of another participant; C = jokes; D = mediates; E = seeks arrangements; F = agrees; G = proposes; H = directs the discussion; I = criticizes; J = informs; K = explains; L = changes the subject; M = asks opinions; N = contradicts; O = disapproves; P = ridicules. Scores based on the assignments by the female or male judge are reported above or below the diagonal, respectively. Correlations were computed for each discussion group and then averaged.

rospective act-frequency estimates. A more detailed discussion of all of these results follows.

Perception of Base Rates

The finding of substantial correspondences among the three coding schemes with regard to the different base rates of the 16 types of behavior revealed a high sensitivity of the retrospective judges to these differences. Furthermore, the finding that the average prototypicality of the 3,696 acts for a behavior category covaries with the retrospectively estimated average frequency for this type of behavior is reminiscent of Mischel and Peake's (1982) finding concerning the relation between the perception of consistency and the temporal stability of highly prototypical acts.[4] In Mischel and Peake's (1982) study, as well as in the present one, it was mainly those acts that were prototypical for a behavior class that were most strongly reflected in global judgments about this category. When act frequencies for behavior categories were estimated retrospectively in the present study, the judges obviously discriminated between more and less prototypical instances for the category at issue. Furthermore, the impression of different base rates was then based on the distinct portions, for the single categories, of highly prototypical acts.

Moreover, it is remarkable that the strongest differences in base rates were revealed by the on-line simple assignment-coding scheme. Whereas a ratio of up to 25:1 was encountered for the most frequent as compared with the least frequent category, this ratio was about 4:1 for the retrospective frequency estimates. Thus, the *big* categories (i.e., those with a high average prototypicality rating) were "favored" by the forced-choice response format. Let us assume that the judges usually assigned an activity to that category for which the act was most prototypical. It would then follow that quite small differences in the average prototypicality for the categories should result in substantial differences in the frequencies with which the several categories were chosen in the forced-choice assignment task. In contrast, according to the systematic overlap hypothesis (Borkenau, 1986b), retrospective-frequency estimates reflect the multiple assignment of acts to several categories. Accordingly, the *bigger* categories would be favored to a higher extent by a forced-choice assignment procedure as compared with the retrospective-frequency ratings. This is what happened in the present study. One may compare these effects to those of a simple-majority voting system as opposed to proportional representation in politics. The former system favors, to a higher extent, the party with the most votes. Loosely speaking, within retrospective-frequency estimates, a system of proportional

Table 10
Structural Correspondences Among On-Line-Recorded Behavior Frequencies, Retrospective-Frequency Estimates, and Semantic Similarities Across the 120 Category Pairs

Model of act-to-trait assignment	Judge's sex	Retrospective-frequency ranking		Retrospective-frequency estimate		Semantic-similarity judgments	
		r	r_S	r	r_S	r	r_S
1	Female	.30	.32	.27	.29	.14	.12
1	Male	.30	.29	.31	.26	.19	.16
2	Female	.73	.71	.65	.64	.72	.69
2	Male	.61	.54	.59	.54	.64	.62
3	Female	.66	.62	.51	.48	.45	.41
3	Male	.42	.33	.44	.36	.48	.39

Note. r_S = Spearman rank-order correlation. The structural correspondences were assessed by computing correlations across the 120 heterotrait–monomethod coefficients.

[4] We are indebted to an anonymous reviewer for bringing this point to our attention.

representation seems to be applied. In contrast, forced-choice assignments would resemble a simple-majority voting system.

Although the rater agreement concerning the on-line classification of single activities was moderate, the judges agreed to a high extent on differences in the base rates of the 16 categories. This pattern was most pronounced for the on-line simple classification task. Whereas the rater agreement was a moderate .30 (Cohen's κ) for the single forced-choice assignments, the two judges exhibited an agreement as high as .83 on differences in the base rates of the 16 classes of behavior. This is just another example of the uses of aggregation. On-line judgments are hardly to be distinguished from questionnaire responses in this respect; judgments about single activities are quite unreliable, but averaging many such unreliable judgments results in considerably more reliable averages.

Validity of Retrospective-Frequency Estimates

The retrospective judges' accuracy concerning their perception of base rates was revealed by their estimates of the absolute frequencies with which the 48 discussants had performed the 16 types of behavior. No such analyses could be performed for the frequency rankings. Thus, the frequency estimates allowed for a more thorough analysis of our data. Moreover, this type of frequency estimate revealed higher correlations with the on-line-recorded act frequencies for the single discussants than did the frequency rankings. Accordingly, the frequency estimates may be regarded as more valid than the frequency rankings (compare Table 4 and Table 5). Surprisingly, the frequency rankings were superior in reliability, but this did not pay off in terms of higher validity. Usually, in psychological research, the reliability of measures is known, but their validity is unknown. According to the present study, this can result in inappropriate decisions. We understand our results as an encouragement to let subjects estimate behavior frequencies rather than apply a scale. It seems that humans are good estimators of such frequencies (cf. Hasher & Zacks, 1984; Nisbett & Kunda, 1985).

The highest validity coefficients were obtained when the activities were scored according to Model 3, in which all act frequency summaries were intercorrelated beyond .98. These intercorrelations decreased considerably when Model 1 or Model 2 was applied. This decrease in the intercorrelations among on-line-recorded act frequencies was accompanied by a decrease in the correlations between on-line-recorded and retrospectively estimated act frequencies for the same behavior category (see Tables 4 and 5). For Model 2, even substantial negative correlations were encountered. Thus, Model 3 was clearly superior in terms of validity. However, Model 3 reflected hardly anything except individual differences in the overall activity of the single actors.

This finding might be explained in several ways. One might suggest an accurate reflection hypothesis. Thus, Model 3 would reveal the highest validity coefficients because it reflects the true relations most accurately. Accordingly, the very high intercorrelations among the 16 act-frequency summaries may have resulted because our 48 actors discussed their topics quite cooperatively in order to win the premium of $12. No gross differences among actors were observed in this respect. It follows that, for example, whether a discussant agreed with or disagreed with a position expressed before depended on the relation of this position to his own. Thus, the frequencies of agreement and contradiction were highly situationally determined, and the most important dispositional factor may indeed have been the different overall activity of the single actors. Good examples for *agrees* were judged as weak examples for *contradicts*, and vice versa, as evidenced by the highly negative correlation among the prototypicality ratings for these two categories. However, our data may reveal a relation like "the more doors I open, the more I tend to close" (Mischel & Peake, 1982, p. 733). Although the opening of a door might be regarded as the opposite of closing it, the frequencies of both activities should be highly positively intercorrelated across subjects.

We are, however, somewhat reluctant to recommend this accurate reflection hypothesis too strongly. Intercorrelations beyond .98 are too high to be taken as representative of the true relations. Moreover, the intercorrelations among the retrospectively estimated act frequencies are considerably lower, although, with a few exceptions, positive in sign. Thus, we doubt the complete appropriateness of the act-frequency summaries that were scored according to Model 3. One might also suggest a distortion hypothesis to account for the high-validity coefficients obtained for the Model 3 scores. Thus, it might be argued that the retrospective judges were asked to do too much when they were requested to form, from a 50-min videotaped discussion that they viewed only once, accurate impressions for 6 targets and 16 distinct types of behavior. The retrospective judges may have applied a main-effects model instead; they may have perceived, quite accurately, the different base rates of the 16 types of behavior as well as the differences in the overall activity of the single actors. However, they may have been unable to estimate *interactions* (i.e., idiosyncratic profiles of behavior for single discussants) with a comparable degree of accuracy. From such an assumption it would follow that those act-frequency summaries that reflect hardly anything else except general activity are predicted with the highest precision.

The hypothesis of a small number of independent dimensions in judgments about personality is supported by findings obtained from factor analytic studies of questionnaire and rating data. It is usually found that (a) the number of independent dimensions for personality judgments is about five (Amelang & Borkenau, 1982; Hogan, 1983; McCrae & Costa, 1987; Norman, 1963) and that (b) raters agree more about the position of subjects on broad dimensions than on narrow dimensions (Amelang & Borkenau, 1982; Koretzky, Kohn, & Jeger, 1978).

Our data are not sufficient to provide a definite answer to this problem. It should be emphasized, however, that if the aforementioned distortion hypothesis were valid, this would only imply that the act-frequency summaries, obtained from Model 3, would not be the most appropriate ones. The very high validity coefficients obtained for this model (see Tables 4 and 5) would then be inflated by common biases. Note, however, that the validity coefficients were also substantial for the other two scoring models. Thus, it may be concluded from the present study that the retrospective judges estimated the act frequencies of the single actors with substantial accuracy.

Structural Fidelity of Frequency Estimates

With respect to the structural fidelity of the retrospective-frequency judgments, it was found that the more two behavior-descriptive terms were similar in meaning, the higher were the intercorrelations among the prototypicality ratings for the two categories. Thus, the findings on the correspondence between semantic similarity and the prototypicality for trait-pairs, as found by Borkenau (1986b) at the trait-level, could be replicated at the level of behavior. Another difference between the two studies is that in the earlier study verbally described acts had to be judged for their prototypicalities. In contrast, in the present study videotaped activities had to be judged. Thus, it seems that the correspondence between semantic similarities and prototypicalities for category pairs is a highly robust phenomenon.

Similar assertions can be made for the correspondence between semantic similarities and the proportion of cross-classifications among judges in a forced-choice assignment task. Here, the present study also replicated the results of the earlier study by Borkenau (1986b). Overlap in meaning might also explain this phenomenon. If activities that are good examples for Trait A tend also to be good examples for Trait B, the assignment decision in a forced-choice task becomes highly arbitrary; hence, the increased probability of discrepant choices.

Furthermore, when act frequencies were scored for the single discussants and then intercorrelated, it turned out that a forced-choice assignment of activities to only one category led to results similar to those reported by Shweder (1975), D'Andrade (1974), and Shweder and D'Andrade (1980). In contrast, structural correspondences of about .65 were obtained among the semantic-similarity structure, the structure of the retrospective-frequency estimates and rankings, and the structure of act-frequency summaries in which the 3,696 activities were coded on-line according to Model 2.[5] These results provide convincing evidence for the systematic overlap hypothesis that was not provided by earlier studies. The studies by Romer and Revelle (1984) and Borkenau (1986b) demonstrated that meaning overlap may explain the correlational structure of retrospectively estimated act frequencies. The present study showed that such puzzling results (e.g., those reported by Shweder and D'Andrade) are encountered only if meaning overlap among the behavior categories is not considered by the on-line coding scheme.[6] Thus the systematic distortion hypothesis suggested by Shweder and D'Andrade turned out to be an artifact.

This does not imply the nonexistence of illusory correlations based on conceptual associations in personality ratings. Illusory correlations of this type have been demonstrated elsewhere (Berman & Kenny, 1976; Borkenau, 1986a; Chapman & Chapman, 1967, 1969). Shweder and D'Andrade, however, argued that the intercorrelations among memory-based ratings reflect hardly anything except conceptual associations. The present study showed that this hypothesis was based on data that may be explained by the neglect of act overlap within the on-line coding scheme. Shweder's (1982) question about whether the intercorrelations among memory-based ratings reflect the structure of language or the structure of behavior implies a misconception about the role of language in behavior observations. Each coding of an observed act implies semantics (Romer & Revelle, 1984). Thus, the correlations among the act frequencies for several types of behavior are predetermined by the meaning relations among the behavior-descriptive terms. The present findings do not suggest, however, that each correlation among retrospective ratings reflects only meaning relations. On the contrary, the structural correspondences of the act-frequency summaries, which were scored according to Model 1, contradict such a position because they are all positive in sign (see Table 10). However, given a specific correlation between two retrospective ratings, it remains unclear to which degree it reflects meaning relations. The correlation may reflect covariations among behaviors, thus pointing to basic dimensions of personality. However, it may also reflect the multiple assignment of acts to several behavior classes, thus pointing to basic dimensions of the language of personality. This situation makes the interpretation of correlations among act-frequency estimates a highly ambiguous affair.

[5] These correspondences were somewhat lower when Model 3 was applied as a scoring key. Note, however, that under Model 3 all intercorrelations among act-frequency summaries reached or exceeded .98. Accordingly, these act-frequency summaries reflected little else except general activity. This seems to have enhanced their validity (see Tables 4 and 5). However, little variance was left among the 120 correlations that might have related to anything.

[6] In the study reported by Shweder and D'Andrade (1980), the online scorers were instructed to "check on the list of 16 terms, the term *or terms* which characterized each act" (p. 44, italics added). Thus, the judges were not provided with a forced-choice task but could indicate meaning overlap among several categories. Despite this opportunity, they produced results quite similar to those found in our study when each activity had to be assigned to only one category. The most reasonable explanation for this pattern of findings is that the judges in Shweder and D'Andrade's (1980) study did not make much use of this opportunity; that is, that they usually used only one category for the scoring of an activity. Moreover, they may mainly have made multiple assignments to dissimilar categories when they were undecided about the meaning of an act in the course of an interpersonal interaction.

References

Alston, W. P. (1975). Traits, consistency, and conceptual alternatives for personality theory. *Journal for the Theory of Social Behavior, 5,* 17–48.

Amelang, M., & Borkenau, P. (1982). Über die faktorielle Struktur und externe Validität einiger Fragebogen-Skalen zur Erfassung von Dimensionen der Extraversion und emotionalen Labilität [On the factor structure and external validity of some questionnaire scales designed to assess extraversion and emotional lability]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 3,* 119–146.

Aries, E. J., Gold, C., & Weigel, R. H. (1983). Dispositional and situational influences on dominance behavior in small groups. *Journal of Personality and Social Psychology, 44,* 779–786.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41,* 258–290.

Berman, J. S., & Kenny, D. A. (1976). Correlational bias in observer ratings. *Journal of Personality and Social Psychology, 34,* 263–273.

Block, J., Weiss, D. S., & Thorne, A. (1979). How relevant is a semantic similarity interpretation of personality ratings? *Journal of Personality and Social Psychology, 37,* 1055–1074.

Borkenau, P. (1986a). Systematic distortions in the recognition of trait information. In A. Angleitner, A. Furnham, & G. van Heck (Eds.), *Personality psychology in Europe: Current trends and controversies* (pp. 177–191). Lisse, the Netherlands: Swets & Zeitlinger.

Borkenau, P. (1986b). Toward an understanding of trait interrelations: Acts as instances for several traits. *Journal of Personality and Social Psychology, 51,* 371–381.

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. (1981). The act frequency analysis of interpersonal dispositions: Aloofness, gregariousness, dominance, and submissiveness. *Journal of Personality, 49,* 175–192.

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

Buss, D. M., & Craik, K. H. (1985). Why *not* measure that trait? Alternative criteria for identifying important dimensions. *Journal of Personality and Social Psychology, 48,* 934–946.

Cantor, N., & Mischel, W. (1979). Prototypicality and personality: Effects on free recall and personality impressions. *Journal of Research in Personality, 13,* 187–205.

Cattell, R. B. (1946). *Description and measurement of personality.* New York: World Book.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72,* 193–204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271–280.

D'Andrade, R. G. (1974). Memory and the assessment of behavior. In H. M. Blalock (Ed.), *Measurement in the social sciences* (pp. 159–186). Chicago: Aldine-Atherton.

Eysenck, H. J., & Eysenck, S. B. G. (1969). *Personality structure and measurement.* London: Routledge & Kegan Paul.

Fishbein, M., & Ajzen, I. (1974). Attitudes toward objects as predictors of single and multiple behavioral criteria. *Psychological Review, 81,* 59–74.

Guilford, J. P. (1975). Factors and factors of personality. *Psychological Bulletin, 82,* 802–814.

Hamilton, D. L., Dugan, P. M., & Trolier, T. K. (1985). The formation of stereotypic beliefs: Further evidence for distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology, 48,* 5–17.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information. *American Psychologist, 39,* 1372–1388.

Herrmann, T. (1973). *Persönlichkeitsmerkmale: Bestimmung und Verwendung in der psychologischen Wissenschaft* [Personality traits: Their definition and use in scientific psychology]. Stuttgart, West Germany: Kohlhammer.

Hogan, R. (1983). A socioanalytic theory of personality. *Nebraska Symposium on Motivation, 1982.* Lincoln: University of Nebraska Press.

Jaccard, J. J. (1974). Predicting social behavior from personality traits. *Journal of Research in Personality, 7,* 358–367.

Koretzky, M. B., Kohn, M., & Jeger, A. M. (1978). Cross-situational consistency among problem adolescents: An application of the two-factor model. *Journal of Personality and Social Psychology, 36,* 1054–1059.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694.

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

McGowan, J., & Gormly, J. (1976). Validation of personality traits: A multicriteria approach. *Journal of Personality and Social Psychology, 34,* 791–795.

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

Newcomb, T. M. (1931). An experiment designed to test the validity of a rating technique. *The Journal of Educational Psychology, 22,* 279–289.

Nisbett, R. E., & Kunda, Z. (1985). Perception of social distributions. *Journal of Personality and Social Psychology, 48,* 297–311.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings.* Englewood Cliffs, NJ: Prentice-Hall.

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

Romer, D., & Revelle, W. (1984). Personality traits: Fact or fiction? A critique of the Shweder and D'Andrade systematic distortion hypothesis. *Journal of Personality and Social Psychology, 47,* 1028–1042.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology, 10,* 173–220.

Semin, G. R., & Greenslade, L. (1985). Differential contributions of linguistic factors to memory-based ratings: Systematizing the systematic distortion hypothesis. *Journal of Personality and Social Psychology, 49,* 1713–1723.

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

Shweder, R. A. (1975). How relevant is an individual difference theory of personality? *Journal of Personality, 43,* 455–484.

Shweder, R. A. (1982). Fact and artifact in trait perception: The systematic distortion hypothesis. In B. A. Maher (Ed.), *Progress in experimental personality research* (Vol. 11, pp. 65–100). New York: Academic Press.

Shweder, R. A., & D'Andrade, R. G. (1980). The systematic distortion hypothesis. In R. A. Shweder (Ed.), *New directions for methodology of social and behavioral science* (Vol. 4, pp. 37–58). San Francisco: Jossey-Bass.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25–29.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Weigel, R. H., & Newman, L. S. (1976). Increasing attitude–behavior correspondence by broadening the scope of the behavioral measure. *Journal of Personality and Social Psychology, 33,* 793–802.

Received January 24, 1986
Revision received July 11, 1986 ■