Fact and Fiction in Implicit Personality Theory

Peter Borkenau and Fritz Ostendorf
Universitat Bielefeld

ABSTRACT This article reports two studies, where the accuracy of implicit personality theory (IPT) was investigated using on-line behavior counts as well as retrospective frequency estimates as standards of comparison Eight discussion groups, each comprising six members, were videotaped Their act frequencies with respect to 16 types of behavior were judged on-line using two coding schemes, each one being applied by two independent raters Five other judges estimated the act frequencies retrospectively Furthermore, judges revealed their IPT by estimating the conditional likelihood of these types of behavior It turned out that (*a*) retrospective judges perceive different base rates accurately, (*b*) the correlations among retrospectively estimated and among on-line recorded act frequencies show high correspondences, (*c*) IPT accurately mirrors the correlations among retrospectively estimated as well as among on-line recorded act frequencies, and (*d*) judges do not appropriately consider perceived base rates when estimating conditional probabilities It is concluded that IPT is considerably accurate in those respects that are important for the validity and structural fidelity of personality ratings

It is generally acknowledged that there exist common beliefs about the relationships among traits and behaviors that are usually referred to as implicit personality theory (IPT) If instructed appropriately, subjects express hypotheses and beliefs concerning the covariations among traits

The research reported in this paper was supported by a grant from Deutsche Forschungsgemeinschaft (Az Bo 774/1-1) to the first author We gratefully acknowledge the contributions of Ulrich Baumgartner, Kristina Herzig, Corinna Ruting, and Angelus Stulik who painstakingly performed the laborious on-line judgment task We thank Alois Angleitner, Rainer Riemann, and Robert A Wicklund for their helpful comments on earlier drafts of this article

Send correspondence to Peter Borkenau, Abteilung Psychologie, Universitat Bielefeld, Postfach 8640, D-4800 Bielefeld, Federal Republic of Germany

Journal of Personality 55 3, September 1987 Copyright © 1987 by Duke University Press CCC 0022-3506/87/$1 50

and among aspects of traıt-relevant behavıor that are largely ın agreement wıth the results obtaıned when real ratees are assessed (Jackson, Chan, & Stncker, 1979, Lay & Jackson, 1969, Mulaık, 1964, Stncker, Jacobs, & Kogan, 1974) Moreover, these belıefs may play a decısıve role ın shapıng traıt-attnbutıons about others Thus ıf subjects were requested to estımate acquaıntances' act frequencıes for behavıors they never had the opportunıty to observe, theır agreement was as hıgh as for observed behavıors (Newcomb 1931) Furthermore, when college freshmen who had never talked to one another were requested to judge each other on traıt-ratıng scales, the factor structure of these ratıngs was hıghly sımılar to that obtaıned from factor analyses of close acquaıntances (Passını & Norman, 1966) Therefore, the accuracy of IPT has ımportant ımplıcatıons for the "structural fidelıty" (Loevınger, 1957) of traıt-ratıngs ın general If the traıt relatıonshıps ın IPT mırror those ın actual behavıor, then a foundatıon would exıst for raters to use lımıted ınformatıon about others ın makıng extensıve judgments concernıng the target's personalıty (Jackson, Chan, & Stncker, 1979) If the nature of IPT ıs pnmanly ıllusory, however, ıt mıght ıntroduce bıas ınto personalıty judgments (Mırels, 1976, 1982)

In some earlıer studıes (Jackson, Chan, & Stncker, 1979, Mırels, 1976, 1982, Stncker, Jacobs, & Kogan, 1974), the accuracy of IPT was ınvestıgated by comparıng lay people's assumptıons about the covanatıons among questıonnaıre responses to the actual coendorsement frequencıes It turned out that subjects were quıte accurate ın predıctıng the relatıonshıps of a correlatıonal type If subjects deemed ıt hıghly probable that a person who endorses ıtem A wıll also endorse ıtem B, then ıtems A and B were usually more hıghly correlated ın self-reports than when thıs probabılıty was judged to be low (Jackson, Chan, & Stncker, 1979, Jackson & Stncker, 1982) The subjects were consıderably less accurate, however, wıth respect to theır absolute estımates concernıng the condıtıonal probabılıtıes Theır lıkelıhood estımates that a person wıll answer "True" to ıtem B ıf ıtem A had been endorsed, and vıce versa, revealed remarkable dıscrepancıes from the empırıcal relatıonshıps Subjects ın partıcular dıd not notıce the asymmetnes ınherent ın many of these condıtıonal probabılıtıes Whereas the condıtıonal probabılıty of a "True" answer to ıtem A, gıven an endorsement of ıtem B, may show a remarkable dıscrepancy from the reverse relatıonshıp, subjects estımated the respectıve condıtıonal probabılıtıes to be about the same (Mırels, 1982) The source of such asymmetnes, however, are dıstınct unconditional probabilities or base rates of item A and item B endorsements Thus it may be suspected that lay people are quite aware of the covariations among questionnaire responses They fail, however, when their task involves an additional consideration of base rates for deriving accurate conditional probability estimates

Does this imply that people are unaware of different base rates of distinct kinds of behavior? Or do their weaknesses lie more in combining knowledge about covariations with knowledge about base rates for deriving accurate conditional probability estimates? Whereas the former alternative would point to a severe lack of accuracy in judging personality characteristics, the latter alternative would point to insufficiencies in the application of formal mathematical concepts (cf Kahneman & Tversky, 1973, Nisbett & Ross, 1980, Tversky & Kahneman, 1974, 1983) According to the second alternative, even university students may be unaware of the fact that the consideration of base rates is necessary for deriving accurate conditional probability estimates They may be aware of different base rates, but still be unaware of their importance in the calculation of conditional probabilities Which of these two alternatives holds true is of crucial importance Whereas conditional probability estimates are quite unusual in personality research, judgments about act frequencies and trait positions are the major data source of personological studies Therefore, one purpose of the present study was to investigate the sensitivity of judges with respect to different base rates of distinct classes of behavior

Another problem is common in the studies by Mirels (1976, 1982) as well as in those by Jackson, Chan, & Stricker (1979) Both compared IPT to the coendorsement frequencies of questionnaire items This procedure, however, circumvents the crux of the matter with respect to the accuracy of IPT, because questionnaire responses themselves may be subject to systematic distortion processes (Mischel, 1968, Shweder, 1982) Thus the high correspondence between IPT and item-covariations (Jackson et al , 1979) may be explained by a common bias inherent in both data sources Subjects may estimate accurately the covariations among questionnaire responses but fail to predict the covariations among the classes of behavior referred to in the personality inventory Accordingly, the accuracy of IPT should be compared against a more objective standard, that is, on-line behavior counts Such on-line records are not free of semantics (Borkenau, 1986, Romer & Revelle, 1984) Semantics are involved in every decision about the meaning of an observed act, whether recorded on-line or retrospectively remembered Accordingly, the question as put by Shweder (1982), whether memory-based ratings reflect the structure of language or the structure of behavior, is misconceived (cf Borkenau, 1986) There may be other biases inherent in questionnaire responses, however, such as aspects of social desirability (Jackson, 1986), self-presentation (Johnson, 1981), and self-schemata (Markus, 1977), making questionnaire item coendorsements a poor standard of comparison regarding the accuracy of IPT

Therefore, in addition to memory based ratings, on-line behavior counts were incorporated into the present studies This made it feasible to investigate whether retrospective judges are (*a*) aware of different overall activities of single actors, (*b*) aware of different base rates of distinct classes of behavior, and (*c*) aware of the covariations among these classes of behavior, furthermore, the present studies investigated whether (*d*) IPT accurately mirrors the correlations among on-line recorded as well as retrospectively estimated behavior frequencies, and (*e*) if IPT takes account of different base rates of the types of behavior at issue

The problem then arises as to how to code observed activities on-line This topic has been one of considerable controversy in the recent past (Borkenau, 1986, Romer & Revelle, 1984, Semin & Greenslade, 1985) For instance, Romer and Revelle (1984) suggest that the results reported by D'Andrade (1974), Shweder (1975), and Shweder and D'Andrade (1980), regarding discrepant correlational structures within retrospectively estimated versus on-line recorded behavior frequencies, are due to the on-line coding scheme used by these authors According to Romer and Revelle, the correlational structure of on-line recorded act frequencies approaches that of retrospective frequency estimates if a scaling coding scheme instead of an identification coding scheme is used for the on-line codings A similar point is made by Borkenau (1986) who showed that prototypicality ratings for acts with respect to traits are positively intercorrelated across acts among semantically similar traits and are negatively intercorrelated for opposed traits Accordingly, the degree of semantic similarity is predictive of "act overlap" among dispositional terms Act overlap, however, influences the intercorrelations of the respective act frequency summaries across subjects Consequently, the intercorrelations among on-line recorded act frequencies are predetermined (but not completely determined) by the semantic similarity relationships among the categories used

Thus it was interesting to investigate the accuracy of IPT in comparison to two different on-line coding schemes one that incorporates overlapping activities among behavior categories in a systematic manner and one that does not In this way it was possible to pursue the following question To what degree is a common sensitivity to semantic relationships responsible for the correspondences between IPT and the intercorrelations among on-line recorded act frequencies?

Study 1

METHOD

Behavior Setting

Eight discussion groups, each comprising six male students, discussed controversial topics and were videotaped Grouping of the 48 students was done in a way that secured different attitudes toward the problem at issue within each of the eight groups The topics chosen (e g , speed limits on German highways) were controversial among the general public at the time when the study was conducted The discussants were seated at two sides of a square table such that the faces of all actors were videotaped during the entire session A name-card with a pseudonym was placed in front of each actor in order to allow observers an identification of the single discussants Each debate lasted about 50 minutes, after which it was interrupted by the experimenter The subjects were paid for their cooperation and a prize was promised for the group that provided the best debate

Coding of Behavior Sequences

On-line behavior counts excluding act overlap The aim of this procedure was (*a*) to keep the memory load low and (*b*) not to incorporate overlapping activities among the behavior categories Therefore, the eight discussions were first subdivided into 15-second units of observation Each 15-second sequence was followed by a 10-second still Two student judges (one female, one male), unacquainted with the discussants and the purpose of the study, and paid for their cooperation, viewed the eight discussions When a still appeared on the screen, they stopped the tape, answered a set of questions with respect to the last sequence, and restarted the video recorder

The ratings were done with respect to 16 categories and were made in booklets The single scenes had been numbered and, for each consecutive scene, one or several judgments had to be given, depending on the number of verbally active discussants Which discussant was regarded as verbally active during a sequence had been agreed upon by one of the authors together with a student Altogether, 3,696 activities were identified in this way Thus each judgment referred to the verbal activity of a specified actor

Overlapping activities among behavior categories were a priori excluded by presenting the rating task in a forced-choice format The two judges had to decide which of 16 behavior categories was most appropriate to classify a given activity Accordingly, the judges were asked to decide, for example "Which of the following categories is most appropriate to classify the behavior of Frank?" A list of the 16 categories was presented below each question The English translation of the behavior descriptors used is supports, takes up the contribution of another participant, jokes, mediates, seeks arrangements, agrees, proposes, directs the discussion, criticizes, informs, explains, changes the subject, asks opinions, contradicts, disapproves, and ridicules [1] A residual category "no judgment possible" was added The two judges indicated their decisions by choosing one and only one of the 16 categories (plus the residual one) as the best descriptor of each of the 3,696 activities This task was performed during a period of about three weeks

On-line behavior counts including act overlap These behavior counts were carried out in a way that allowed for (*a*) the multiple classification of acts to several of the behavior classes under study, and (*b*) consideration of the fact that membership of an act in a dispositional category seems to be a matter of degree (cf Buss & Craik, 1983) Finally, (*c*) an attempt was made to keep the memory load low

The first two requirements were met through the use of the following procedure Each discussion was displayed 16 times to each of the two judges The judges were instructed "to indicate how appropriately the given behavior may be characterized by the category at issue," using seven-point rating scales with endpoints + 3 (= very good example for the category at issue) and − 3 (= blatant counter-example for the category at issue) On these rating scales, each of the two judges made 16 × 3696, that is, 59,136 judgments, altogether This task was performed during a period of about six months The judges were paid, were unacquainted with the 48 discussants, and performed no other task in the course of the present project The third requirement was met by using the same tapes as for the on-line codings excluding act overlap

Retrospective frequency estimates Five student observers (three male, two female), unacquainted with the actors and paid for their cooperation, viewed

[1] The German terms used were *unterstutzt, greift Beitrage anderer auf, scherzt, vermittelt, sucht Ausgleich, stimmt zu, schlagt vor, leitet die Diskussion, kritisiert, informiert, erklart, schweift vom Thema ab, fragt nach Meinungen, widerspricht, lehnt ab,* and *macht lacherlich*

each of the eight discussions as a whole, in a different random order Before viewing the first tape, they were informed about the details of their rating task Then, after having viewed each 50-minute discussion, they were provided a booklet On the cover page of this booklet, they were instructed "to indicate how frequently the single discussants have acted in a way as described by the following categories " At the top of each of the 16 following pages, a different behavior category was written, followed by a sentence asking, "How frequently has each of the six participants shown corresponding behavior during the discussion?" The six pseudonyms were listed below The subjects wrote their frequency estimate behind each name The order of the 16 categories was randomized and different for each judge

Implicit Personality Theory

It is common experience that in most discussion groups some people talk more than others This fact is important for the accuracy of IPT since it increases the conditional probabilities of all behaviors That is, if a target person frequently shows the verbal behavior A, he or she probably is one of the more active participants Therefore, he or she is also likely to show almost any verbal behavior B more frequently than a less active group member Although different overall activities of discussants are thus relevant for conditional probability judgments, subjects may be unaware of this relationship For this reason, two sets of instructions were written, one of which explicitly directed the judges' attention to the different overall activity of the discussants In the other instructions, this information was omitted

Ten student judges (five female, five male), paid for their cooperation, were administered one of the following two instructions, respectively The first of these, where the information about distinct overall activities was omitted, read

> Please imagine the following situation There is a discussion group, comprising six male persons, who discuss about the topic "speed limits on highways " The participants hold contradictory attitudes toward this issue Imagine this situation as vividly as you can When you have done this, please answer the following questions

The second set of instructions, read to ten other judges, contained the above passage plus the addendum

> Please imagine additionally that the single members of this discussion group show very different amounts of overall activity Some members talk very frequently whereas others are silent most of the time

The questions, then administered to the two groups of judges, were identical The order of presentation, however, was randomized and different for each single judge The questions were worded, for example, "If a participant jokes frequently, how likely is it that he also informs frequently " Conditional probability estimates were made on seven-point scales, the endpoints of which were 1 ("very unlikely") and 7 ("very likely") Because asymmetries in the sense of *p* (A/B) being different from *p* (B/A) were of great interest, each pair of categories was presented "forward" and "backward " Accordingly, given 16 categories, 240 (i e , 16 × 15) conditional probabilities were estimated in this way by each judge

Semantic Similarity Judgments

Twenty subjects judged the semantic similarity relationships among the 16 terms descriptive of behavior This was done on a seven-point scale with the endpoints -3 (= antonyms) and $+3$ (synonyms or near synonyms) The 120 trait-pairs were presented on a video screen in a different random order for each judge

RESULTS

Overall Activity of Single Actors

As had been assumed, the number of verbal activities, as agreed upon by one of the authors and a student, differed markedly among the single discussants Their average number per participant was $M = 77$, the standard deviation being $s = 37 12$ The most active discussant made 159 contributions, whereas the least active one spoke only nine times Accordingly, mainly positive intercorrelations among the act frequencies for the various behavior classes should be expected across actors

Retrospective Frequency Estimates

Reliabilities The retrospective estimates by the five independent judges were averaged in order to increase the reliabilities of the scores The reliabilities were estimated using intraclass correlations [ICC (2,1) and ICC (2,5), according to the taxonomy by Shrout and Fleiss, 1979] The reliabilities of the single raters ranged from 18 to 52 for the 16 single categories Their average amounted to 30 For the mean frequency estimates of the five judges, the average reliability for the 16 categories was 65, the coefficients ranging between 52 and 84 for the 16 single categories

Mean frequency estimates for categories. The mean of the retrospective frequency estimates, averaged across the 48 discussants, differed markedly for the 16 single categories. It was lowest ($M = 1.12$) for the category "ridicules" and highest ($M = 4.86$) for the category "takes up the contribution of another participant." Accordingly, the conditional probabilities for many category-pairs should be asymmetrical, the conditional probability of B, given A, being different from the conditional probability of A, given B. Thus the precondition was met for investigating the accuracy of IPT according to Mirels' criteria.

Intercorrelations among retrospective frequency estimates. Estimates for the 16 behavior classes were first calculated for each discussion group and then averaged for the eight groups. This was done in order to reduce the influence of outliers and get more reliable correlation coefficients. As should be expected from the highly varying overall activity of the single actors (found also in each discussion group considered separately), these 120 coefficients were predominantly positive in sign (with only two exceptions). They ranged from $r = -.07$ to $r = .93$. This result might be interpreted as indicating a considerable sensitivity of the retrospective judges with respect to the highly varying overall activity of the single actors.

On-Line Behavior Counts

Reliabilities. Since two coding schemes had been applied to each of the 3,696 activities by two independent judges, separate analyses could be performed for each single judge to check the replicability of the results across judges. This procedure was extremely desirable in order to control for idiosyncrasies of the single judges. The reliability of single judges is at issue here. The reliability of the forced-choice assignments was estimated using Cohen's κ. This reliability turned out to be $\kappa = .30$. The reliability of the prototypicality ratings was estimated using intraclass correlations [ICC (2,1), according to Shrout & Fleiss, 1979]. For the single categories, the reliabilities of the single judges ranged from .11 to .53. The average coefficient for the 16 categories amounted to .42.

On-line recorded and retrospectively estimated base rates. With respect to the forced-choice, on-line coding scheme, the number of assignments to the 16 single categories was compared to the mean retrospective frequency estimates, averaged across the 48 discussants. This comparison was performed by the computation of correlation coefficients across the 16 categories. Separate analyses were performed for the assignments of the female and male on-line judges. The correlations amounted to .66 and .75, respectively.

With respect to the prototypicality on-line coding scheme, however, the rationale was somewhat more complicated. Because in this coding scheme, membership of an observed activity in a behavior class was regarded as a matter of degree, no definite on-line recorded behavior frequencies could be established. However, the mean prototypicality ratings, averaged across all 3,696 activities, showed remarkable discrepancies for the 16 single categories. This implies that some categories were judged to be better descriptors of the majority of observed activities than were others. If these would be predominantly those categories where the retrospective frequency estimates were high, this would point to a considerable accuracy of the retrospective judges with respect to different base rates of the classes of behavior under study. The correlations across categories between the mean of the retrospective frequency estimates and the mean of the prototypicality ratings amounted to .66 and .87 for the female and male on-line judges, respectively. When the prototypicality ratings were averaged for the female and male judges, the respective correlation increased to $r = .89$. Thus it may be concluded that the retrospective judges were highly sensitive to the different base rates of the 16 classes of behavior under study.

Intercorrelations among on-line recorded act frequencies. It was expected that the correlational structure of the on-line coded act frequencies should depend heavily upon the coding scheme used. Particularly, the structural correspondences between retrospectively estimated and on-line coded behavior frequencies should be more pronounced for the prototypicality coding scheme than for the forced-choice coding scheme.

For the forced-choice, on-line coding scheme, the frequencies of single discussants with respect to the single types of behavior were calculated by simply counting the respective "entries." These entries were then intercorrelated among categories across actors. Separate analyses were conducted for the assignments of the female and male on-line judges. The correspondences of these correlational structures with other types of correlational structure are reported in Table 1.

On-line recorded act frequencies of the single discussants, being based on the prototypicality on-line coding scheme, were calculated by

Table 1

Structural Correspondences of the Intercorrelations among Act Frequencies, of Conditional Probability Estimates, and of Semantic Similarities (Study 1)

Type of correlational structure	2	3	4	5	6	7
1 Intercorrelations among act frequencies, coded on-line with a forced-choice coding scheme (female judge)	44	16	26	27	33	14
2 Intercorrelations among act frequencies, coded on-line with a forced-choice coding scheme (male judge)	–	14	37	31	30	17
3 Intercorrelations among act frequencies, coded on-line with a prototypicality coding scheme (female judge)		–	58	65	62	72
4 Intercorrelations among act frequencies, coded on-line with a prototypicality coding scheme (male judge)			–	59	65	64
5 Intercorrelations among retrospectively estimated act frequencies				–	61	54
6 Conditional probability estimates (IPT)					–	74
7 Semantic similarities of the category pairs						–

Note Each correlation was calculated across the 120 pairs of categories

weighting each activity of an actor by its estimated prototypicality for the category at issue, and then summing over all activities of the respective actor Once again, separate analyses were performed for the male and female judges It was expected that the higher the intercorrelation of these act frequency summaries, the higher the intercorrelation of the retrospective frequency estimates should be The respective correlations across the 120 combinations of the 16 behavior classes are also reported in Table 1 [2]

[2] For more details of the results concerning the relationship between retrospective

Suffice it to say here that, contrary to the results reported by Shweder and D'Andrade (1980), intercorrelations among retrospective frequency estimates quite accurately mirrored those among on-line recorded ones This, however, was only achieved when meaning overlap among the behavior classes was considered In contrast, when a forced-choice, on-line coding scheme was applied, the results of the present study were comparable to those reported by Shweder and D'Andrade

Implicit Personality Theory

Reliabilities The reliability of the conditional probability estimates was assessed by intraclass correlations [ICC (2, 1) and (2, 10), according to Shrout and Fleiss, 1979] Separate computations were performed for the two groups of judges For the group that was not explicitly informed about the varying overall activity of the single discussants, the reliability of the single judges amounted to 59 (coefficient 2, 1) The average score of the ten judges had a reliability of 92 (coefficient 2, 10) For the other group of judges, having been explicitly informed about the varying overall activity of the discussants, the reliability of the conditional probability estimates was considerably lower, amounting to 29 (single judges) and 80 (average of the 10 judges)

Overall activity and conditional probability estimates This information alone is not surprising because it might be due to a restriction of range Restriction of range would have been expected, had the judges who were informed about differences in overall activity used only the higher digits of the rating scale This would have been an appropriate decision However, whereas restriction of range in the second as compared to the first instruction group was encountered and may be regarded as one reason for the lower obtained reliability, the means of the estimated conditional probabilities did not differ significantly between the two groups For the subjects not informed about differences in overall activity, the grand mean was 4 03 For the subjects who were informed about that fact, the grand mean was 4 00 The sign of the difference is contrary to expectations, but the difference is insignificant, $t(239) = 0\ 48, p > \ 05$

A further analysis checked whether single judges could be identified whose judgments tended toward the higher end of the scale and there

and on-line recorded act frequencies, the reader is referred to Borkenau and Ostendorf (1987) were none The systematic mean differences among the single judges were quite small In no case did the average of the 240 ratings of a single judge exceed 4 5, and the lowest average for a single judge was 3 62 Thus, none of the 20 judges concluded from the differences in overall activity of the discussants that the conditional probability estimates should tend toward the higher end of the scale Moreover, when explicitly informed about these different overall activities, the only effect was to lower the agreement of the judges

Symmetry in conditional probability estimates For two reasons, these further analyses will be reported only for the group uninformed about the different overall activities First, in earlier studies on the accuracy of IPT, there was also no information provided about different overall activities and, second, the judgments of the uninformed group were more reliable than those of the informed group [3] According to Mirels, judges estimate the conditional probabilities of A, given B, and B, given A, to be about the same We tried to replicate this finding by comparing the respective conditional probability estimates across the 120 pairs of categories at issue, using a correlation coefficient The replication of Mirels's results was completely successful in this respect, the correlation amounting to $r =$ 93 As this value is of a similar size as the reliability of the judgments, it may be concluded that the judges estimated the conditional probabilities to be completely symmetrical

This hypothesis was further pursued using the following rationale Given a certain correlation among two act-trends A and B, and different base rates, then if A occurs more frequently than B, the conditional probability p (A/B) must be higher than the reverse relationship p (B/A) Moreover, the higher the discrepancy of the two base rates, the higher will be the discrepancy of the two conditional probabilities Thus the difference of the base rates is positively correlated with the difference between the conditional probabilities Accordingly, when judges who estimate conditional probabilities consider base rates in an appropriate manner, a positive correlation is to be found between base rate differences and the differences between the corresponding conditional probability estimates But if no substantial correlation is found, it may be con-

[3] The results for the group of judges, having been informed about considerable differences in the overall activity of the single discussants, were highly similar to those uninformed thereabout In no case did the difference between the correlations, obtained for these two groups of judges, exceed $r =$ 10

cluded that judges do not incorporate base rate differences into their conditional probability estimates

In the present study differences of the base rates of the 16 categories could be estimated (*a*) from differences with respect to how frequently activities were assigned to the 16 single categories in the forced-choice, on-line coding task, (*b*) from differences in the means of the prototypicality ratings for the single categories, and (*c*) from differences of the mean retrospective frequency estimates The first two indices can be interpreted roughly as on-line behavior counts Accordingly, for each of the 120 pairs of categories, the difference between the base rates, A minus B, was calculated according to (*a*) the frequency of assignments to the two categories by the female judge, (*b*) the frequency of assignments to the two categories by the male judge, (*c*) the mean prototypicality ratings of the female judge, (*d*) the mean prototypicality ratings of the male judge, and (*e*) the mean retrospectively estimated frequencies These differences in base rates were then compared to the differences of the estimated conditional probabilities p (A/B) − p (B/A) The correlations among these six measures are reported in Table 2

The four judges who performed on-line behavior codings agreed to some extent that some sorts of behavior occurred more often than others The appropriate correlations, indicating agreement among the four on-line judges, are quite substantial, albeit somewhat lower for those comparisons involving the male judge who had estimated prototypicalities Despite this, all on-line judges exhibited considerable agreement with the retrospective judges with respect to different base rates of the 16 types of behavior at issue Thus, the retrospective judges considered the different base rates in an appropriate manner

However, a very different picture emerged when judges were instructed to estimate conditional probabilities Although all the five correlations in the last column of Table 2 have a positive sign, the highest of these five correlations is lower than any other correlation reported in Table 2 Thus, whereas the retrospective judges were highly accurate in their perception of distinct base rates, the judges who were required to estimate conditional probabilities hardly incorporated these different base rates into their judgments

Covariations among act trends and IPT Thus far, we corroborated Mirels's findings concerning the neglect of base rates in conditional probability estimates, using retrospective and on-line behavior ratings instead

Table 2
Accuracy of Perception of Different Base Rates and of Inferences with Respect to Asymmetries in Conditional Probabilities (Study 1)

Type of base rate difference	2	3	4	5	6
1 Difference of the frequency of assignments to the two categories (A-B) by the female judge	85	66	35	69	11
2 Difference of the frequency of assignments to the two categories (A-B) by the male judge	–	68	31	75	21
3 Difference of the mean prototypicality ratings (A-B), female judge		–	38	87	09
4 Difference of the mean prototypicality ratings (A-B), male judge			–	65	01
5 Difference of the mean retrospectively estimated act frequencies (A-B)				–	10
6 Difference of the estimated conditional probabilities *p* (A/B) – *p* (B/A)					–

Note Each correlation was calculated across the 120 pairs of categories

of questionnaire responses as a standard of comparison Another result of earlier studies was a strong relationship between estimated conditional probabilities and item covariations In the present study, IPT was compared to the intercorrelations among retrospectively estimated as well as on-line recorded act frequencies Once again, with respect to the on-line behavior counts, both coding schemes were incorporated into the analysis For IPT, the estimated conditional probabilities *p* (A/B) and *p* (B/A) were averaged for each pair of categories A and B The correspondence between the structure of on-line recorded act frequencies, that of retrospectively estimated act frequencies, and the conditional probability estimates was then assessed Furthermore, the semantic similarity relationships among the 16 behavior-descriptive terms were incorporated into this analysis Accordingly, correlations were calculated across the 120 category-pairs among six indices of behavior co-occurrences plus the semantic similarity relationships These coefficients are reported in Table 1 (see p 425)

This table has several remarkable features First, all correlations have a positive sign, independent of the on-line coding scheme applied, some correspondence is found between on-line, coded act frequencies, retrospectively estimated ones, conditional probability estimates, and semantic similarity relationships However, the second feature of Table 1 is that the lowest coefficient in the last four rows ($r =$ 54) is higher than the highest one in the upper two rows ($r =$ 44) This implies that the structural correspondences are considerably higher for the on-line coding scheme, taking account of meaning overlap, than for the forced-choice, on-line coding scheme This relationship holds true for all comparisons (i e , those with retrospective frequency estimates, those with conditional probability estimates, and those with the semantic similarities) Moreover, IPT turns out to reflect the semantic similarity relationships best, the intercorrelations stemming from the on-line prototypicality coding scheme second best, the intercorrelations among retrospectively estimated frequencies third best, the intercorrelations stemming from the forced-choice, on-line coding scheme considerably worse and the base rates of the 16 types of behavior worst (compare with Table 2 for the last information) This implies that the correspondence with the conditional probability estimates is higher the more that the single measures reflect semantic relationships Thus, for example, the influence of the base rates is negligible whereas the highest correspondence is found with the purely semantic relationships among the category descriptors

Study 2

The first study was ambiguous in one important respect The judges who retrospectively estimated the act frequencies of the 48 single targets viewed the videotapes of the eight discussions The judges, however, who estimated the conditional probabilities, did not Accordingly, it was demonstrated that knowledgeable informants are aware of different base rates of distinct classes of behavior But it was not demonstrated that exactly those judges who estimated the conditional probabilities were aware of these different unconditional probabilities Thus the symmetry of the conditional probability judgments may stem from several sources The respective judges may either have been uninformed about the differences in the base rates, or they may have been informed about them but unable or unwilling to incorporate these base rate differences into their conditional probability estimates In order to clarify this ambiguity, it was necessary to let one and the same group of judges estimate act frequencies and conditional probabilites This was the aim of the second study

METHOD

Subjects

In order to get highly knowledgeable raters, the retrospective judges of the first study were contacted a second time This happened about one year after their first participation Meanwhile, one female judge had left the university and could not participate again Thus four knowledgeable raters (three male, one female) remained for the judgment task, for which they were paid

Procedure

One of the eight discussions of the first study was displayed to the four judges a second time The same discussion group was independently shown to each of the four raters After having observed this discussion as a whole, the judges were first instructed to estimate the act frequencies of the single actors with respect to the 16 classes of behavior at issue For this purpose, they were given the same type of booklet that they had used previously in the first study After having thus estimated the act frequencies of six single targets, the booklet was collected and another one, comprising the conditional probability judgment task, was handed out The instructions on the cover page of the second booklet read

> Several types of behavior were displayed during the discussion that you have just observed We are interested in potential relationships among some of these types of behavior More precisely, you are requested to estimate conditional probabilities Accordingly, it is your task to judge how likely it was in the just-observed discussion that a participant exhibited a certain type of behavior, conditional upon his rate of displaying other kinds of behavior Thus the questions put to you are of the general format "If a participant showed behavior A frequently, how likely is it that he or she also showed behavior B frequently" Please base your judgments as far as possible on the discussion just observed

The material for the conditional probability estimates, which then followed, was the same as that applied for these estimates in Study 1 Accordingly, the main difference from Study 1 was that, in the second study, the judges were encouraged to base their conditional probability estimates on a just-observed discussion Furthermore, the design of Study 2 made it possible to check whether the very same judges, who had estimated the respective act frequencies a short time ago, would utilize different base rates when subsequently estimating the conditional probabilities

RESULTS

Retrospective Frequency Estimates

Reliabilities The reliabilities of the retrospective frequency estimates for the six discussants were estimated using intraclass correlations [ICC (2,1) and ICC (2, 4), according to Shrout and Fleiss, 1979] The coefficients for the single behavior classes ranged from 10 (single judgments) and 31 (mean judgment) to 59 (single judgments) and 85 (mean judgment) The average reliability for the 16 categories was 38 (single judgments) and 69 (mean judgment) These figures indicate that the reliability of these estimates was of a magnitude similar to that of Study 1

Mean frequency estimates for categories As in the first study, the mean of the retrospective frequency estimates (this time averaged across six discussants) differed markedly for the 16 classes of verbal activities It was lowest ($M = 0\ 42$) for the category "ridicules" and highest ($M = 4\ 63$) for the category "proposes " With respect to the relative prevalence of the 16 types of behavior, a comparison of the first and second study was performed by calculating the correlation across the 16 categories among the mean frequency estimates obtained in the two studies This correlation was $r = \ 88$, indicating a considerable stability of the judges in preferring certain categories Furthermore, this coefficient implies that the one out of the eight discussion groups selected for the second study was highly representative with respect to the types of behavior displayed

Intercorrelations among retrospective frequency estimates Once more, the six discussants were perceived as differing highly in their overall activities, the most active target was judged as having manifested five times as many verbal activities as the least active one Accordingly, of the 120 intercorrelations among the 16 kinds of behavior, each one computed across six targets, 88% were positive in sign

Implicit Personality Theory

Means and reliabilities The grand mean of all conditional probability estimates was $M = 3\ 74$ This time the four single judges differed markedly with respect to the averages of their conditional probability judgments that ranged from 2 41 to 4 37 The reliability of the conditional probability estimates was assessed via intraclass correlations The reliability of the single judgments [ICC (2, 1), according to Shrout and Fleiss] amounted to 25 and that of the averaged ratings of the four judges (ICC 2, 4) was 57 When comparing the respective reliabilities in the two studies, the highest rater agreement was obtained when the judges had the least information When the judges were informed about differences in overall activity (in Study 1), or had to estimate the conditional probabilities with respect to a specified discussion (in Study 2), the rater agreement decreased

Symmetry in conditional probability estimates The second study had been conducted primarily to answer the following question "When judges are aware of different base rates for distinct classes of behavior, do they consider these base rates when estimating conditional probabilities?" It has already been reported that the four judges under study perceived marked differences in the base rates for the 16 classes of behavior Consequently, if the judges relied on mathematical considerations when estimating conditional probabilities, marked asymmetries should be expected The degree of symmetry in the respective estimates was assessed once again by computing the correlation between *p* (A/B) and *p* (B/A) across the 120 combinations of the 16 categories This correlation was *r* = 83 Thus the judges estimated the conditional probabilities to be highly symmetrical The somewhat lower value of this correlation in the second compared to the first study (wherein this correlation was 93) might be explained by the lower reliability of the respective ratings in the second study

As in the first study, the hypothesis of the judges' ignorance of the importance of base rates for conditional probabilities was further pursued by comparing the differences in the base rates to asymmetries in the conditional probability estimates That is, for each of the 120 combinations of the 16 behavior classes, the difference in the base rates, A minus B, was computed The on-line recorded base rates for the 16 behavior classes had once again been estimated by counting the number of assignments to the 16 categories in the forced-choice task and by averaging the prototypicality ratings for the 16 single categories across all 3,696 activities The data of the first study were used here However, only the activities of the one discussion group, displayed in both studies, were consid-

Table 3

Accuracy of Perception of Different Base Rates and of Inferences with Respect to Asymmetries in Conditional Probabilities (Study 2)

Type of base rate difference	2	3	4	5	6
1 Difference of the frequency of assignments to the two categories (A-B) by the female judge	53	46	62	70	02
2 Difference of the frequency of assignments to the two categories (A-B) by the male judge	–	39	71	63	17
3 Difference of the mean prototypicality ratings (A-B), female judge		–	40	51	– 01
4 Difference of the mean prototypicality ratings (A-B), male judge			–	77	11
5 Difference of the mean retrospectively estimated act frequencies (A-B)				–	12
6 Difference of the estimated conditional probabilities *p* (A/B) – *p* (B/A)					–

Note Each correlation was calculated across the 120 pairs of categories

ered Separate analyses for the female and the male on-line judges were performed The retrospectively estimated base rates as well as the conditional probability estimates both stemmed from the second study Differences in the base rates, A minus B, calculated accordingly, were compared to the differences of the conditional probability estimates *p* (A/B) minus *p* (B/A) For this purpose, the respective correlations were computed across the 120 category-pairs If this correlation would be positive and substantial, it would indicate that the judges considered the different base rates when estimating conditional probabilities Note that in this second study the conditional probabilities were estimated by judges who had perceived the frequency of the most frequent type of behavior to be about ten times that of the rarest type Table 3 reports the resulting correlations

Table 3 may immediately be compared to Table 2 The results are highly similar, independent of whether or not the judges had observed and accurately estimated the different base rates for the distinct classes of behavior This means that in the second as well as in the first study, the retrospective judges considered the different base rates in an appropriate manner Furthermore, in both studies, base rates had hardly any influence on conditional probability estimates The results reported in Table 3, however, are more revealing than those reported in Table 2 Study 2 demonstrates that the neglect of base rates in conditional probability estimates does not stem from a lack of knowledge with respect to the distinct base rates Rather, whereas the retrospective judges were highly sensitive to the different base rates of the behaviors at issue, the very same judges did not incorporate these different base rates into their conditional probability estimates

Correlations among act trends and IPT As in the first study there was a check to see how accurately the conditional probability estimates reflected the correlations among act trends Intercorrelations among online recorded act trends relied on the data of the first study In contrast to the results reported in Table 1, however, only the data for the six discussants were considered, having also been displayed to the judges in the course of the second study The intercorrelations among the retrospective frequency judgments as well as the conditional probability estimates relied on the respective judgments performed during the second study The "forward" and "backward" conditional probability estimates were averaged Correlations were then computed among these indices of act-trend covariation across the 120 category-pairs Moreover, the semantic similarity relationships among the behavior categories were incorporated into the analysis All of these coefficients are reported in Table 4

Table 4 is essentially a replication of Table 1 Note, however, that in the second study the judges estimated frequencies and conditional probabilities with respect to exactly the same discussion that they had just observed Accordingly, the finding of correlation coefficients of a similar size in both studies is far from trivial This finding means that the conditional probability estimates by judges, who never observed the discussions, were as accurate as those by judges who were knowledgeable to a high degree Accordingly, IPT seems to be both highly persistent and highly accurate with respect to the intercorrelations among act trends Furthermore, Table 4 as well as Table 1 suggests an explanation of why this turns out In both studies IPT exhibits stronger correspondences to the degree that single measures of act trend covariation reflect semantic relationships

Table 4
Structural Correspondences of the Intercorrelations among Act Frequencies, of Conditional Probability Estimates, and of Semantic Similarities (Study 2)

Type of correlational structure	2	3	4	5	6	7
1 Intercorrelations among act frequencies, coded on-line with a forced-choice coding scheme (female judge)	25	12	24	03	05	05
2 Intercorrelations among act frequencies, coded on-line with a forced-choice coding scheme (male judge)	–	23	26	14	13	05
3 Intercorrelations among act frequencies, coded on-line with a prototypicality coding scheme (female judge)		–	67	56	53	53
4 Intercorrelations among act frequencies, coded on-line with a prototypicality coding scheme (male judge)			–	43	57	53
5 Intercorrelations among retrospectively estimated act frequencies			–		47	32
6 Conditional probability estimates (IPT)					–	75
7 Semantic similarities of the category pairs						–

Note Each correlation was calculated across the 120 pairs of categories

DISCUSSION

The results of the present studies may be summarized as follows First, highly significant correspondences between conditional probability estimates (IPT) and the intercorrelations among the respective act trends were found This result is in agreement with previous studies (Jackson, Chan, & Stricker, 1979, Jackson & Stricker, 1982) In the present study, however, on-line behavior counts were incorporated in addition to retrospectively estimated act frequencies In contrast to the earlier studies, the present results may not be explained by undetected biases inherent in both implicit personality theory and the intercorrelations among questionnaire responses (cf Shweder, 1975) This means that the present investigation provides more dependable evidence with respect to the validity of IPT

Second, in agreement with the results reported by Mirels (1976), the conditional probability estimates were highly symmetrical, even when the base rates of the behaviors involved were quite different This relationship was found independently of whether the base rates stemmed from on-line recorded or retrospectively estimated act frequencies It may then be concluded that the judges' neglect of base rates is not limited to the situation where conditional probabilities of questionnaire responses are to be assessed This may have been an especially difficult task, because lay people lack experience with respect to the endorsement frequencies of questionnaire items

Third, the memory based frequency ratings reflected the different base rates of the sixteen types of activities and the highly varying overall activity of the single actors, as indicated by the on-line behavior counts This pattern of results suggests that retrospective judges are highly aware of different base rates and incorporate them into their frequency judgments (Hasher & Zacks, 1984, Nisbett & Kunda, 1985) Judges seem to fail, however, when requested to derive asymmetrical conditional probability estimates from these discrepant base rates Especially revealing in this respect were the results of the second study, where the very same judges who perceived gross differences in the base rates did not derive asymmetrical conditional probabilities from these differences The judges remained widely ignorant of the fact that, given two activities with distinct base rates, the conditional probability of the more frequent activity should be higher

Why did the perceived distinct base rates not become incorporated into the conditional probability estimates? One reason may be that the judges were not aware that base rate information is a prerequisite for an accurate estimation of conditional probabilities It might then be possible that the judges estimated the conditional probabilities mainly on semantic grounds When two behavior-descriptive terms were similar in meaning, the judges may have estimated the respective conditional probabilities to be high, and vice versa Such an interpretation would be in agreement with results from other studies, supporting the concept of a representativeness heuristic, violating the rules of formal and mathematical reasoning (cf Kahneman & Tversky, 1973, Tversky & Kahneman, 1974, 1983) Moreover, it would be compatible with the results reported in Table 1 and Table 4, because in both studies the highest correspondences were found with respect to the semantic similarity judgments. Furthermore, IPT showed considerable correspondences with on-line recorded act frequencies when the latter took the meaning relationships among the category descriptors into account, that is, when a prototypicality coding scheme was applied. According to the systematic overlap hypothesis (Borkenau, 1986), covariations among act-trends are partially determined by the meaning relationships among the behavior-descriptive terms. If this explanation were valid, it would follow that IPT mirrors the intercorrelations among act frequencies in an accurate way, because both are largely determined by the meaning similarity relationships among the categories applied (cf. Romer & Revelle, 1984). IPT would, therefore, not bias the intercorrelations among act frequency estimates to a significant extent. IPT would be insensitive, however, to asymmetries in conditional probabilities that do not immediately follow from the meaning relationships among the behavior categories.

There exists another possible explanation for the present findings. We used items such as "If a participant jokes frequently, how likely is it that he or she also informs frequently," when asking for conditional probability estimates. The meaning of "frequently" remained thus unspecified. If the judges used this term in the sense of deviation from the mean, that is, "more frequently than the average frequency of pertinent behavior in this discussion group," the influence of different base rates of the 16 classes of behavior should have been eliminated. Highly symmetrical conditional estimates for *p* (A/B) and *p* (B/A) would then have been appropriate. In this case, however, the considerable differences in the overall activity of the single discussants should have exerted their influence. If the judges did indeed apply the concept "frequently" in the sense of "more than average," their conditional probability estimates should have tended towards the higher end of the seven-point rating scale due to the perceived or informed about differences in overall activity of the single discussants. This, however, did not turn out. The conditional probability estimates remained uninfluenced even by explicit information about differences in the overall activity of the single discussants.

The problem of the generalizability of the present findings is at issue here, namely, generalizability across targets, across situations, across raters, and across behavior categories. Among these, the last problem is the most important. Semin and Greenslade (1985) argued that semantic similarity relationships are predictive of co-occurrences for mediate, that is trait-descriptive terms, but not for immediate terms that is, terms that focus on specific behaviors. In the present study, however, immediate terms were used and their co-occurrences were accurately predicted by the semantic similarity relationships. The notable exception was the intercorrelations among on-line codings based on a forced-choice assignment task. But Semin and Greenslade's (1985) co-occurrence measure resembled our IPT-measure highly, implying that the present study failed to replicate their findings. Unfortunately, Semin and Greenslade (1985) neither report the variance nor the reliability of their semantic similarity judgments, although these are of crucial importance for the interpretation of their findings. We only know from their study that a set of correlation coefficients, that is, those involving semantic similarity relationships among immediate terms, was unexpectedly low. Such a finding may have been due to restriction of range or to unreliability of the measures that were compared. Without any information about these statistics, their findings are not directly interpretable.

It would have been possible to select behavior-descriptive terms with a narrower meaning than those that were used in the present study, for example, "smokes," "writes," "reads," etc. Possibly the degree of meaning overlap among such terms would have been lower and the correspondences between semantic similarities and act frequency co-occurrences would have been diminished. The present study, however, was designed to investigate the systematic distortion hypothesis vis-à-vis a systematic overlap hypothesis. Thus we chose categories at a level of inclusiveness similar to that used by Shweder and D'Andrade (1980). D'Andrade (1974) reanalyzed studies using categories that stemmed from Bales's (1950) coding scheme. So did we. However, some modifications were implemented for the present study because we felt that some types of behavior, being quite frequent in leaderless discussion groups, were insufficiently represented by Bales's categories. If anything, the categories became more immediate as a result of this modification (e.g., jokes), thereby working against our hypothesis.

The study by Shweder and D'Andrade (1980) also used categories that were similar to the present study in level of inclusiveness. For example, they used the terms "explains," "informs," and "criticizes," which were also applied here. Meaning overlap among categories at this level of inclusiveness, however, is easy to demonstrate (cf. Borkenau & Ostendorf, 1987).

There remains Shweder's (1975) reanalysis of the Newcomb study. The behavior categories used there were indeed considerably more immediate than those in the present study (e g , "spends more than an hour of the day alone") However, multiple assignments of behaviors to several categories were appropriate even in this study, other categories being "reads a half hour or more during the day" and "continues on a single activity for the whole morning " Accordingly, act overlap among behavior categories seems also to exist for this study Given the assumption, however, that it was less pronounced in this study than in those by D'Andrade (1974) and Shweder and D'Andrade (1980), it is interesting to recognize that the structural correspondences between on-line coded and retrospectively estimated behavior frequencies turned out to be highest in Newcomb's data One might speculate that puzzling results like those reported by Shweder and D'Andrade turn out most clearly in studies wherein (*a*) a forced-choice, on-line coding scheme is applied and (*b*) meaning overlap among behavior categories is high

Obviously, we could have chosen more traditional categories, such as "extraversion," "dominance," and "agreeableness," that is, more mediate terms But we did not, for several reasons First, we chose categories at a similar level of inclusiveness as Shweder and D'Andrade because the systematic distortion hypothesis is based on this empirical evidence Second, in order to investigate the impact of act overlap, it was necessary to select categories that were not mutually orthogonal like many traditional personality factors Third, it was more convenient to use categories being specifically designed to code the behavior in leaderless discussion groups Otherwise the coding task would have been more difficult and the rater agreement might have been lower Finally, meaning overlap among traditional trait-categories has already been demonstrated in the study by Borkenau (1986), where verbally described acts had to be assigned to the categories "aloofness," "dominance," "submissiveness," "quarrelsomeness," and "agreeableness " Given this background, we intended in the present study to demonstrate the generalizability of the findings across categories of a dissimilar level of inclusiveness

IMPLICATIONS

The question of the accuracy of retrospective reports about act frequencies is of critical importance for personality psychology because they may provide an economical shortcut for getting information about persons In contrast, on-line codings are expensive, tremendously time consuming, and extremely limiting in that observation is a prerequisite for such codings Accordingly, since only public events can be observed, these are the only ones amenable to this kind of analysis The use of retrospective reports, however, requires their validation Furthermore, the problem of validity may be divided into two subproblems, external validity and structural fidelity (Loevinger, 1957) This article is about structural fidelity It has been argued in this respect that IPT may bias the intercorrelations among retrospective ratings because it deviates considerably from the empirical relationships (Mirels, 1976, 1982, Shweder and D'Andrade, 1980) What, then, may be concluded from the present study?

The implications of the present study are somewhat puzzling in this respect because it suggests that IPT reflects mainly the structure of language but does not distort the correlations among retrospective ratings The reason is that the correlations among act frequencies are themselves predetermined by the meaning relationships among the behavior-descriptive terms Thus the present study favors an optimistic view with respect to the accuracy of personality impressions Implicit personality theory quite accurately reflects the intercorrelations among act frequencies recorded on-line Accordingly, as far as our trait-attributions are shaped by IPT (Newcomb, 1931, Passini & Norman, 1966), their intercorrelations seem not to be very much distorted

However, there are also other facts about behavior that may not be predicted from meaning relationships, for example, different base rates for distinct types of behavior Retrospective judges seem to be highly sensitive to base rates but they do not incorporate this knowledge into conditional probability estimates of an IPT-type In this respect, Mirels's (1976) findings were corroborated in the present study, thus demonstrating illusory aspects of IPT The finding, however, that judges are very poor in accurately estimating conditional probabilities need not bother personologists very much Whereas act frequency and trait ratings are very common in personality research, and the correlations calculated from these data are of crucial importance for personality theory (Shweder, 1975), no study comes to our minds where any important personological assumption would have relied upon conditional probability estimates In the correlational sense, where the accuracy of IPT is crucial for personality research, IPT seems to be highly accurate With respect to conditional probability estimates, however, which seem to be seriously flawed, the importance of IPT for the accuracy of personality impressions may be regarded as negligible

REFERENCES

Bales, R F (1950) *Interaction process analysis* Chicago University of Chicago Press

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

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

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

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

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

Jackson, D N (1986) The process of responding in personality assessment In A Angleitner & J S Wiggins (Eds), *Personality assessment via questionnaires Current issues in theory and measurement* (pp 123–142) Berlin Springer

Jackson, D N, Chan, D W, & Stricker, L J (1979) Implicit personality theory Is it illusory? *Journal of Personality,* **47**, 1–10

Jackson, D N & Stricker, L J (1982) Is implicit personality theory illusory? Armchair criticism vs replicated empirical research *Journal of Personality,* **50**, 240–244

Johnson, J A (1981) The "self-disclosure" and "self-presentation" views of item response and personality scale validity *Journal of Personality and Social Psychology,* **40**, 761–769

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

Lay, C H, & Jackson, D N (1969) Analysis of the generality of trait-inferential relationships *Journal of Personality and Social Psychology,* **12**, 12–21

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

Markus, H (1977) Self-schemata and processing information about the self *Journal of Personality and Social Psychology,* **37**, 63–78

Mirels, H L (1976) Implicit personality theory and inferential illusions *Journal of Personality,* **44**, 467–487

Mirels, H L (1982) The illusory nature of implicit personality theory Logical and empirical considerations *Journal of Personality,* **50**, 203–222

Mischel, W (1968) *Personality and assessment* New York Wiley

Mulaik, S A (1964) Are personality factors raters' conceptual factors? *Journal of Consulting Psychology,* **28**, 506–511

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

Passini, F T, & Norman, W T (1966) A universal conception of personality structure? *Journal of Personality and Social Psychology,* **4**, 44–49

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

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 *Progress in Experimental Personality Research*, **11**, 65–100

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

Stricker, L J , Jacobs, P I , & Kogan, N (1974) Trait interrelations in implicit personality theories and questionnaire data *Journal of Personality and Social Psychology*, **30**, 198–207

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

Tversky, A , & Kahneman, D (1983) Extensional versus intuitive reasoning The conjunction fallacy in probability judgment *Psychological Review*, **90**, 293–315

Manuscript received March 17 1986 revised August 22 1986