ELSEVIER

Acta Psychologica 92 (1996) 33–57

acta psychologica

Confidence and accuracy in trait inference: Judgment by similarity

Derek J. Koehler [a,*], Lyle A. Brenner [a], Varda Liberman [b], Amos Tversky [a]

[a] *Department of Psychology, Stanford University, Stanford, CA, USA*
[b] *The Open University of Israel, Tel-Aviv, Israel*

Received 27 June 1994; revised 12 November 1994; accepted 26 January 1995

Abstract

We examine the confidence and accuracy with which people make personality trait inferences and investigate some consequences of the hypothesis that such judgments are based on similarity or conceptual relatedness. Given information concerning a target person's standing on three global personality dimensions, American and Israeli subjects were asked to estimate the target's self-ratings of 50 trait adjectives and to express their confidence by setting a 90 percent uncertainty range around each estimate. The estimates were positively correlated with the actual ratings obtained from subjects who had evaluated themselves in terms of the 50 traits, but were far too extreme. Furthermore, confidence was negatively correlated with accuracy: People's estimates were most inaccurate and made with greatest certainty when the trait in question was highly similar to the information provided as a basis for judgment. We suggest that intuitive personality judgments overestimate the coherence of the structure underlying trait constructs.

PsycINFO classification: 2340

Keywords: Overconfidence; Nonregressive prediction; Representativeness; Trait inference

1. Introduction

The field of human judgment is concerned with the accuracy of people's judgments (e.g., Dawes et al., 1988) and with their psychological bases (e.g., Kahneman et al.,

* Correspondence address: Department of Psychology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. E-mail: dkoehler@watarts.uwaterloo.ca.

0001-6918/96/$15.00
SSDI 0001-6918(95)00005-4

1982). The study of intuitive personality judgments is a particularly attractive domain in which to examine how people make inferences, both because it is engaging and familiar to subjects and because the domain is itself of theoretical interest to social and personality psychologists. Research in this area asks two questions: How do people use their general knowledge about personality traits to make inferences about specific individuals, and what are the characteristics of the body of knowledge used in such inferences?

In general, intuitive predictions can be made on the basis of two sources. First, one might refer to the actual correlation observed in previous experience. This might be thought of as a data-driven or extensional source. Second, one might rely on the conceptual relatedness or similarity among the constructs under consideration. This might be thought of as a theory-driven or intensional source. As an example, consider the relationship between a person's handwriting and some personality trait such as dominance. One might decide that people with bold handwriting tend to be more dominant than those with tentative handwriting, either because one has noted a systematic relationship between the two variables in day-to-day experience, or because of the similarity between the concept of boldness as used to describe handwriting and the concept of boldness as used to describe behavior and the subset of people who regularly exhibit such behavior.

As this example suggests, similarity (i.e., semantic or conceptual relatedness) often overshadows previously-observed correlation in intuitive judgement. Several lines of research support this contention. The work by Chapman and Chapman (1967; see also Chapman and Chapman, 1969) on illusory correlation first drew attention to the notion of theory-driven assessment in the domain of psychiatric diagnosis. They argue that clinicians continue to rely on projective tests, despite their lack of empirical support, because the strong conceptual relationship between critical test features and the diagnosis blinds them to the lack of an objective statistical association. In the domain of intuitive prediction, Kahneman and Tversky (1973; also Kahneman and Tversky, 1972; Tversky and Kahneman, 1983) have demonstrated that people's probability judgments are often determined by the degree to which the event in question is representative of or similar to the relevant stereotype or mental model. Nisbett and Ross (1980; also Jennings et al., 1982) emphasize the distinction between data-driven and theory-driven judgment and showed that the latter often dominates the former.

In the domain of personality judgment, Shweder and D'Andrade (1980; also D'Andrade, 1965; Mulaik, 1964; Shweder, 1977) underscore the relationship between judgments of likeness and of likelihood, arguing that personality judgments are made on the basis of semantic rather than statistical association. In their view, 'propositions about language' are confused with 'propositions about the world' in such judgments, yielding a lay personality theory in which perceived trait relationships are completely determined by the conceptual similarity among the trait terms used to describe people. As one might expect, this 'systematic distortion hypothesis' has proved highly contentious, resulting in considerable debate and subsequent research in the literature (e.g., Block et al., 1979; Borkenau and Ostendorf, 1987; deSoto et al., 1985; Jackson et al., 1979; Jackson and Stricker, 1982; Mirels, 1976, 1982; Shweder and D'Andrade, 1979; Weiss and Mendelsohn, 1986).

The present study contributes to research on judgment by similarity by testing several specific hypotheses about the relationship between people's confidence and accuracy in intuitive trait inference. With the notable exception of Oskamp (1965), there has been relatively little research investigating the relationship between the accuracy of people's personality judgments and the confidence with which they are made. This is rather surprising as much of the initial interest in the debates over cross-situational consistency of behavior (Mischel, 1968) and over the systematic distortion hypothesis (Shweder and D'Andrade, 1980) arose from the apparent discrepancy between people's strongly-held intuitions about personality and the empirical links among traits. An obvious question is whether people are confident in their intuitive personality judgments, even when such judgments are inaccurate. To date, however, only one series of studies (Dunning et al., 1990; Vallone et al., 1990) has examined confidence in a personality judgment task in which accuracy could be reliably measured. These researchers found considerable overconfidence when subjects were asked to predict the behavior of their roommates or of a previously-interviewed target. One of the major goals of the current study was to investigate the determinants of confidence in personality trait inferences.

In the present study we attempted to elaborate the psychological mechanisms underlying confidence in personality judgments by systematically varying the information available as the basis of inference. We presented subjects with a few pieces of 'personality type' information provided by a target individual and asked them to make inferences about personality traits the target was likely to endorse as self-descriptive. Subjects estimated the target's percentile score on each of 50 trait adjectives and assessed their confidence in each estimate by specifying a range which they were 90 percent confident contained the target's score. To measure accuracy, these judgments were compared to self-rated personality information from a sample of over 400 undergraduates. We should point out that our purpose in this study was to compare the targets' perceptions of themselves regarding the traits under consideration with others' predictions of these self-perceptions given the general personality type the target claims to have. As such, the task is different than that of predicting actual behavior or of predicting the target's classification along some objectively-defined personality dimensions (cf. Funder and Colvin, 1988). Furthermore, it does not provide a direct test of the systematic distortion hypothesis, which is normally tested by comparing personality ratings (and similarity ratings) with a more objective measure of trait-consistent behavior. The study nevertheless may be of interest to social psychologists as it addresses the question of whether the target's self-perceptions (perhaps influenced by a lay theory of personality) are accurately reconstructed by subjects completing the prediction task, regardless of the objective status of the target with respect to the traits under consideration.

A number of predictions for the results of this study follow from the premise that intuitive trait inferences and the confidence with which they are expressed rely on conceptual relatedness or similarity: (1) To the extent that conceptual relatedness reflects statistical association, trait inferences will exhibit some validity; (2) Because the statistical association among trait constructs is typically rather weak, however, inferences based on similarity will generally exaggerate the links between traits. This suggests that, especially under conditions of high similarity, trait inferences are likely to

be too extreme; (3) If expressed confidence in a trait inference also reflects conceptual relatedness rather than statistical association, then a generally low correlation between confidence and accuracy is expected; and (4) Because confidence judgments are assumed to be based on conceptual relatedness, they are unlikely to respond appropriately to factors affecting accuracy but not conceptual relatedness, such as whether the target is a single person or a group of people. This factor was manipulated in the current study, and is expected to have a much larger effect on accuracy than on the extremeness of the trait inferences or the confidence with which they are made.

To directly examine the relationship between conceptual relatedness and people's confidence and accuracy in the trait inference task, a separate group of subjects was asked to rate the degree of similarity between the traits under consideration and the information used as a basis for inference. Three more predictions apply to these data: (5) If trait inferences in this task are based primarily on similarity, then the similarity ratings will better predict the inferences than will the actual self-ratings given by the targets; (6) Similarity will also predict the confidence expressed in the trait inferences, with higher similarity inducing a greater sense of confidence; and (7) Under conditions of high similarity, inferences will be extreme and confidence will be high. Because extreme inferences are especially likely to be in error, this suggests that confidence will be negatively correlated with accuracy.

Data were collected in the United States and in Israel in order to examine whether culturally-based conceptual differences in perceived trait meaning would lead to systematic differences in trait inferences and in the confidence with which they are expressed. Cross-cultural comparisons are also relevant for testing the generality of the basic results across substantially different populations.

2. Method

2.1. The target group

Over three academic quarters, 206 undergraduates (95 females, 111 males) enrolled in an introductory psychology course at Stanford University and 266 undergraduates (120 females, 146 males) enrolled in a sciences program at Tel-Aviv University served as subjects in the American and Israeli Target Groups, respectively. Each completed the personality inventory, which took approximately half an hour, as part of a course requirement. The questionnaire distributed to the Israeli subjects was translated from English into Hebrew.

The personality inventory consisted of 5 sections. The first asked for some basic background information from each subject: gender, number of siblings, age, and college major. With the exception of gender, none of this information was used or analyzed in the current study.

The second section asked subjects to rate themselves in terms of 3 global personality dimensions based loosely on those of the Myers-Briggs Lifetypes Inventory, a popular

personality classification system used widely in business and industry. The dimensions were described as follows:

EXTRAVERT – project energy outward; enjoy interaction with people.
INTROVERT – keep energy inside; enjoy solitude.

ANALYTIC – prefer to act through a logical, step-by-step process.
INTUITIVE – prefer to act on inspiration or imagination.

DECISIVE – seek to control my life, exerting my will on events.
ADAPTIVE – seek to adapt my life to changing circumstances.

Subjects were asked to check one of 6 boxes between the 2 poles indicating the extent to which one pole was a better description of their personality than was the other. A major advantage of these dimensions is that, for each, neither pole is socially undesirable.

Once all the data were collected, we classified each subject in terms of one pole of each dimension on the basis of whether his or her response fell above or below the mean self-rating on that dimension. [1] This information was provided to subjects in the Inference Group as the basis for their judgments. While it could be argued that subjects might prefer to know the actual box the target had checked, it seemed to us that the mean-split information was simpler to use in that subjects would not need to take into account any skewness in the distribution of responses. Given that the task of the inference subjects is to judge how the targets compared with the general population on a set of traits, the position of the target relative to the mean for each dimension seems an appropriate inferential cue.

The third section of the personality inventory asked target subjects to rate themselves in terms of 50 personality trait adjectives. Our main concerns in selecting the 50 traits used in the questionnaire were (a) that the traits should represent a reasonably broad selection of personality concepts, and (b) that the traits should be simple, commonly-used terms that virtually all college students would be familiar with. A list of these traits appears in Table 1.

To test whether the selected traits are generally representative of the different types of personality traits used to describe people, we examined how they were distributed among the widely accepted 'Big 5' factors (see, e.g., McCrae and Costa, 1987). First, the 50 traits were categorized as belonging to one (or, in some cases, two) of the Big Five factors (agreeableness; surgency/extraversion; conscientiousness; emotional stability/neuroticism; intellect/culture) on the basis of the classification provided by Goldberg (1990) and Goldberg (1991). Then, to determine whether this distribution was fairly representative, it was compared to two presumably representative trait distributions found in the recent personality literature: the Goldberg (1982) classification of the Norman (1963) list of 1541 trait terms, and the Peabody (1987) shorter set which was specifically constructed to be representative in terms of the Big 5. The analysis revealed

[1] With one exception: Due to particularly skewed distribution, Israeli target subjects were classified as Decisive or Adaptive on the basis of the midrange rather than the mean for that dimension.

no significant differences between our sample of traits and those of Goldberg, $\chi^2(4, N = 50) = 1.32$, n.s., or Peabody, $\chi^2(4, N = 50) = 2.22$, n.s.

For each trait, subjects were asked to rate how well the trait described them as opposed to other students at their university by estimating their percentile score. Subjects were instructed on the meaning of percentile scores, and were asked to give their estimated percentile score for each trait by circling one of the numbers on an 11-point scale running from 0 to 100 in increments of 10. They were told, for example, that if they believed they were more helpful than 80 percent of students at their university then they should circle 80 on the scale. As expected, these ratings were biased by social desirability: Mean estimated percentile scores were considerably greater than 50 for desirable traits and less than 50 for undesirable traits (cf. Dunning et al., 1989). Because we did not want the subjects in the Inference Group to have to compensate for this bias when making their judgments, we converted the self-ratings into objective percentile scores (relative to the entire group of target subjects), thereby imposing a grand mean of 50 for each trait in both samples. One might argue that subjects would prefer to estimate the actual numbers circled by the targets, but as with the dimension information there is an obvious tradeoff between giving the actual response of the subject and giving the position of the subject's response relative to the overall distribution of responses. It seemed to us, however, that the original scale would not be especially meaningful given the social desirability bias. The conversion into objective percentile scores was intended to make it easier for subjects to indicate their belief that the target's self-rating was above or below average for that trait.

The fourth section of the inventory included 50 two-alternative questions about personal habits and preferences. The final, fifth section asked subjects to make hypothetical choices between pairs of jobs. Results from these two sections are reported in a separate paper (Brenner et al., in press-a).

2.2. *Selecting target profiles*

Recall that each subject was classified as either Extravert or Introvert, either Analytic or Intuitive, and either Adaptive or Decisive. We will refer to each of the 8 classification categories that result by combining the 3 dimensions as different personality *profiles*. Subjects in the Inference Group were provided with a profile as the basis for their judgments. In the conclusion of this paper we discuss the issue of whether the current results can be generalized to inferences based on more complex information.

To simplify the design and analysis, we presented only 2 of the 8 possible profiles to the Inference Group. We chose to hold constant the Adaptive-Decisive dimension (by using only Decisive target subjects), in part because it was significantly correlated with the Analytic-Intuitive dimension and in part because it seemed the least informative of the 3 dimensions. The two profiles used were the Extravert, Intuitive, Decisive (EID) profile and the Introvert, Analytic, Decisive (IAD) profile. In the American sample, there were 37 target subjects classified as EID and 37 classified as IAD. In the Israeli sample, there were 65 subjects classified as EID and 35 classified as IAD. This difference could be due either to genuine cultural differences or to other factors such as the translation from English to Hebrew.

In the American sample, males were more likely to be classified as Introverts and females were more likely to be classified as Extraverts, $\chi^2(1, N = 206) = 11.2$, $p < 0.001$. As a result, the IAD profile consisted of 70% male subjects and the EID profile consisted of 62% female subjects. Because there were more males than females in the Israeli Target Group, the IAD profile consisted of 71% male subjects and the EID profile consisted of 60% male subjects.

Before moving on to the inference task, it may be informative to examine just what could be predicted from the personality profiles. Table 1 displays the mean responses of both profile (target) groups for each of the 50 traits. In both the American and Israeli samples, the number of significant differences between the two profiles greatly exceeds the number expected by chance, $p < 0.0001$ by binomial test for both samples. Furthermore, these differences generally seem to fall in the direction that one would expect intuitively on the basis of the profile information. For example, American subjects in the IAD profile group tended to rate themselves as more soft-spoken, less affectionate, and less creative than did subjects in the EID profile group. These results provide evidence that the classification system we used and the personality inventory we constructed are suited for present purposes, in that they distinguish among people and have some face validity as individual difference measures.

2.3. The inference group

We recruited 40 American subjects for the inference task through an advertisement run in the student newspaper and fliers posted around campus. Subjects were paid $7 for their participation in the experiment, which lasted approximately 45 minutes. Almost all of the subjects were students at Stanford University. Data from one subject were discarded because he failed to complete a large number of items.

Subjects were first shown the 3 six-point scales used by the Target Subjects, each pole of which was labeled with the appropriate term and definition. They were told that an initial group of university students had been classified as either Introvert or Extravert, Analytic or Intuitive, and Adaptive or Decisive on the basis of whether their self-ratings were above or below the group mean. The subjects were informed that their task would be to estimate how the target subjects responded to the rest of the personality inventory on the basis of the targets' classification on these 3 personality dimensions. Subjects estimated the responses to the entire inventory for two personality profiles (IAD and EID). The order in which the two profiles were considered was counterbalanced across subjects. To prevent confusion or forgetting, the personality profile currently under consideration appeared at the top of each page of the questionnaire.

Each subject served in one of two experimental conditions. In the *individual* condition, subjects ($n = 19$) were presented with a personality profile and were told it belonged to a single individual, specified by initials, who had been randomly selected from the target group. In the *group* condition, subjects ($n = 20$) were asked to estimate the mean responses of the entire group of target subjects falling into a given classification category. The instructions presented to subjects in the two conditions were identical except for certain necessary differences noted below.

Subjects were informed that their task was to estimate the target's (or target group's

mean) percentile score on each trait. Subjects were instructed on the meaning of a percentile score, using the same description presented to the target subjects in the original personality inventory. For each trait, they were presented with the same 11-point scale given to the target subjects, and were asked to circle their best guess (which we will refer to as a *trait estimate*) and to draw brackets creating an interval which they were 90 percent certain contained the correct percentile score. Subjects in the *individual* condition were instructed as follows:

> We will call your percentile score estimate your 'best guess'. You will also be asked to give a high and low estimate. Your low estimate should be a number that you are quite certain is lower than the person's actual percentile score. You should make this low estimate such that you believe there is only a 5 percent chance that the actual percentile score is lower than this estimate. Likewise, your high estimate should be made such that you believe there is only a 5 percent chance that the actual percentile score is higher than this estimate.

Subjects were then presented with an example of a completed scale, using the trait *helpful*. The instructions continued:

> The range between your low estimate and your high estimate is your 90% *confidence interval*. Because you have made your high and low estimates such that there is only a 5 percent chance that the actual percentile score will fall below your low estimate, and a 5 percent chance that this number will fall above your high estimate, this means that there should be a 90 percent chance that the actual percentile score will fall *between* your low and high estimates. If we were to look at a large number of such estimates, the actual percentile score should fall within your 90% confidence interval 90 percent of the time, and should fall outside the interval 10 percent of the time (5 percent greater than the high estimate, and 5 percent below the low estimate).

The subjects were then shown a completed rating scale with '90%' written over the bracketed interval and '5%' written above the ranges outside the interval on either side. They were further instructed that a narrower interval indicates greater confidence in the accuracy of the trait estimate than does a wider interval. Again, they were presented with examples.

Subjects in the *group* condition were given similar instructions except that they were phrased in terms of estimating the mean percentile score of the entire target group rather than a single individual's rating. They were asked to draw brackets such that the interval included the mean percentile score for the target group in 90 percent of the judgments they made. To avoid confusion with the statistical concept of a *confidence interval*, we will refer to the interval our subjects set as an *uncertainty range,* which corresponds to what is called a *credible interval* in Bayesian inference. While Peterson and Pitz (1988) have demonstrated some interesting differences between the use of uncertainty ranges and probability estimates as measures of confidence, it seemed to us that subjects would find the uncertainty range the easiest to use in the current context.

For the Israeli Inference Group, 86 undergraduates at Tel-Aviv University served as subjects. A reward equivalent to roughly $25 was promised to the 3 subjects who gave

the most accurate judgments. As this is a fairly large amount by the standards of Tel-Aviv undergraduates, it appeared to generate some excitement among the subjects and presumably enhanced their motivation to make accurate inferences. The experimental design was identical to that used with the American subjects, except for the following two points. First, because more subjects were available, each Israeli subject considered only one of the two profiles ($n = 42$ for the IAD profile and $n = 44$ for the EID profile) rather than both as in the American sample. Second, because we felt the American data would be sufficient to test the individual versus group manipulation, all Israeli subjects were assigned to the *individual* condition.

3. Results

3.1. Trait estimates

In the next four sections we focus on the trait estimates provided by subjects in the Inference Group. This is followed by discussion of the uncertainty ranges. Because the *individual* versus *group* condition manipulation used in the American sample had no significant effect on the trait estimates, $p > 0.40$ by binomial test, (nor would a difference be expected, normatively) that factor is ignored in the analyses of the trait estimates. [2] We will return to this manipulation when we discuss the uncertainty ranges.

3.1.1. Accuracy

Accuracy can be examined using either a correlational or absolute error metric. We begin with the correlational analysis.

Table 1 displays the mean target ratings and trait estimates for both the Israeli and the American data. One immediate observation is that the trait estimates appear to be positively correlated with the target means. In fact, for the American subjects, the correlation between the mean trait estimates and mean target ratings was 0.56 for the IAD profile and 0.83 for the EID profile, $z(47) = 2.69$, $p < 0.01$ for the difference between the two profiles. For the Israeli subjects, the correlation was 0.66 for both the IAD profile and the EID profile. Thus, at this highly aggregated level, the trait estimates predicted the mean target ratings quite well. The difference between the correlations for the American and Israeli samples is nonsignificant for the IAD profile and marginally significant for the EID profile, $z(47) = 0.78$ and 1.92, respectively.

We also computed the correlation of each individual subject's trait estimates with the mean target ratings. For the American subjects, the mean correlation between estimated and target ratings was 0.37 for the IAD profile and 0.55 for the EID profile. For the Israeli subjects, the mean correlation between estimated and target ratings was 0.40 for the IAD profile and 0.36 for the EID profile. Of course, the predictability of a single

[2] Of 100 possible tests, only 4 differences between the two conditions were statistically significant at $p < 0.05$, which is almost exactly the number expected by chance under the null hypothesis.

Table 1
Actual and estimated differences in target ratings between the two target profiles (EID and IAD) for the Israeli and American samples. (Only differences of 10 or greater are listed.)

Israeli sample						Trait	American sample					
Target			Estimate				Target			Estimate		
EID	IAD	diff	EID	IAD	diff		EID	IAD	diff	EID	IAD	diff
58	40	18	64	37	27	unpredictable	64	40	24	68	27	41
55	52		63	39	24	risky	60	44	16	66	32	34
60	44	16	76	54	22	active	58	38	20	75	43	32
59	39	20	66	38	28	cheerful	62	41	21	74	43	31
59	41	18	72	43	29	affectionate	63	37	26	67	38	29
54	48		67	42	25	unconventional	63	41	22	65	36	29
59	38	21	78	39	39	friendly	58	36	22	73	47	26
58	45	13	66	42	24	likable	54	38	16	68	42	26
47	48		59	42	17	athletic	52	48		64	39	25
52	50		71	47	24	creative	61	41	20	71	46	25
46	49		58	39	19	flexible	52	41	11	59	35	24
44	48		37	39		procrastinating	56	46	10	57	33	24
59	37	22	67	35	32	warm	63	39	24	64	41	23
54	48		62	52	10	assertive	64	41	23	72	52	20
56	48		65	43	22	happy	60	41	19	67	47	20
55	44	11	61	53		idealistic	54	47		64	44	20
58	44	14	69	47	22	optimistic	59	45	14	69	50	19
49	51		44	42		gullible	47	49		50	33	17
51	47		64	53	11	helpful	56	42	14	62	49	13
55	45	10	62	44	18	sympathetic	58	30	28	59	46	13
45	52		67	62		curious	57	52		66	55	11
55	52		71	66		competitive	43	55	−12	67	57	10
49	43		38	39		religious	50	46		57	49	
54	47		78	74		ambitious	49	50		67	62	
54	47		56	54		arrogant	49	51		55	50	
55	41	14	56	48		sensitive	56	41	15	56	51	
46	47		46	48		tolerant	58	44	14	53	49	
50	53		57	46	11	moody	44	48		55	52	
46	57	−11	65	71		perceptive	56	44	12	63	60	
51	48		59	52		considerate	55	40	15	54	53	
52	53		75	74		independent	55	43	12	69	69	
47	48		35	52	−17	anxious	47	54		52	53	
43	51		46	48		gentle	44	45		53	54	
55	48		53	57		conceited	51	53		54	56	
48	48		58	60		polite	49	47		55	60	
55	54		63	66		tough-minded	54	48		56	62	
52	45		59	62		loyal	52	42	10	54	61	
50	49		58	59		conscientious	43	50		51	65	−14
48	51		57	61		reliable	56	48		51	70	−19
46	50		53	57		cynical	44	52		40	62	−22
46	49		47	56		theoretical	50	49		44	66	−22
50	50		58	72	−14	punctual	50	52		48	75	−27
45	51		58	69	−11	secretive	43	53	−10	37	65	−28
51	57		59	75	−16	realistic	44	45		46	76	−30

Table 1 (continued)

Israeli sample						Trait	American sample					
Target			Estimate				Target			Estimate		
EID	IAD	diff	EID	IAD	diff		EID	IAD	diff	EID	IAD	diff
50	59		65	75	−10	practical	40	51	−11	44	76	−32
48	51		57	69	−12	tidy	40	51	−11	39	71	−32
52	51		62	50	12	soft-spoken	33	60	−27	29	62	−33
51	52		59	73	−14	organized	46	59	−13	42	76	−34
49	56		61	78	−17	solemn	40	59	−19	28	64	−36
43	55	−12	35	53	−18	shy	29	59	−30	21	64	−43

target individual within a profile (rather than the mean of all individuals in the target profile group) is considerably lower. For American subjects in the *individual* condition, the average correlation (over individual targets) between estimated and target ratings for the 50 traits was 0.10 for the IAD profile and 0.16 for the EID profile. For the Israeli subjects, the average correlation was 0.07 for the IAD profile and 0.06 for the EID profile.

We now move from the correlational analysis to an examination of the actual magnitude and direction of our subjects' inferential errors. Table 1 lists the mean differences in target ratings between the two profiles along with the estimated differences. For easier reading, only differences of at least 10 percentile points are displayed. The table shows that the estimated differences tend to be considerably larger than the actual differences between the two profiles. American subjects in the Inference Group expected that people who differed in their classification profile would also differ in how they responded to almost all of the traits in the personality inventory. Israeli subjects appeared to expect somewhat fewer differences, but because there were in fact fewer actual differences between the two profiles for the Israeli sample, they too overestimated the differences. Subjects in both groups appear to have been generally accurate in assessing the direction of the differences between the two profiles even though they overestimated their magnitude.

We computed the absolute deviation of the trait estimates from the mean target ratings, which yields a simple error measure scaled in the original units used in making the estimate. Each subject's mean absolute deviation measure was computed across the 50 traits separately for each target profile. For the American sample, the mean absolute deviation of the trait estimates was 14.3 ($SD = 4.5$) for the EID profile and 16.2 ($SD = 4.5$) for the IAD profile. For the Israeli sample, the mean absolute deviation was 17.8 ($SD = 5.0$) for the EID profile and 16.9 ($SD = 3.6$) for the IAD profile. The difference in accuracy between the two profiles is statistically significant for the American data, $t(34) = 3.72$, $p < 0.001$, but not for the Israeli data, $t(84) = 0.92$, n.s.

This accuracy measure is not particularly informative without some standard of comparison. One obvious candidate is the accuracy that could be achieved by responding with the grand mean (50 percent) for each item in the inventory. The average absolute deviation that would be obtained by making a trait estimate of 50 for every trait

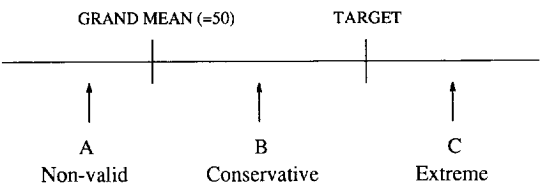

Fig. 1. Schematic diagram showing the three possible relationships between the trait inference, the grand mean, and the actual target rating.

for both profiles was 7.0 for the EID profile and 5.9 for the IAD profile in the American sample, and 4.2 for the EID profile and 3.8 for the IAD profile in the Israeli sample. Thus, somewhat surprisingly, these results show that subjects would have been considerably more accurate in terms of absolute deviation had they simply responded with the grand mean for each trait. In fact, only one of the 125 subjects achieved an accuracy score better than that which could have been obtained by using this strategy. Note, however, that such a strategy would have reduced the correlational accuracy measure to zero.

It is important to point out that these results do not imply that the profile information provided is entirely nondiagnostic. To the contrary, our earlier analyses showed that there were systematic, significant differences in the trait ratings that could be predicted from the profile information. Furthermore, subjects were generally successful in assessing the direction of the difference between the two profiles, showing that their estimates were not unrelated to the target ratings. Taken together, it seems that subjects correctly interpreted the relationship between the profile information and the trait ratings but greatly overestimated the strength of the relationship. We further investigate this idea in the next section.

3.1.2. Extremeness of trait estimates

Consider the diagram portrayed in Fig. 1, which shows three possible (ordinal) positions of a trait estimate relative to the target rating and the trait's grand mean.

We refer to the case labeled Estimate A in Fig. 1 as a *nonvalid* inference, because the profile information led the judge to depart in the incorrect direction from the grand mean. The remaining two cases represent *valid* inferences (i.e., the estimate departs from the grand mean in the direction of the mean target rating). We refer to the case illustrated by Estimate C as an *extreme* inference because in this case the judge would have been well-served by making his or her estimate more regressive with respect to the grand mean. Likewise, the case portrayed by Estimate B will be referred to as a *conservative* inference because the judge did not depart sufficiently from the grand mean in the direction implied by the profile information.

The mean percentage of valid inferences, collapsed across the 50 traits and the two personality profiles, was 74 percent for the American subjects and 71 percent for the Israeli subjects. If the estimates were unrelated to the target ratings, this proportion would be 50 percent. Instead, people's estimates were in the correct direction more often

than would be expected by chance, as measured by comparing the mean percentage of valid judgments per trait with 50, $t(99) = 10.7$ for the American sample and 10.1 for the Israeli sample, $p < 0.0001$ for both. For the valid inferences, an extremeness score was defined as the difference (target − estimate) if both were greater than 50, and as the difference (estimate − target) if both were less than 50. This score is positive if the estimate is too extreme, negative if the estimate is too conservative, and zero if the estimate is unbiased. Its magnitude corresponds to the distance (on the percentile scale) between the trait estimate and the target rating; it is therefore a measure of the estimate's accuracy.

The 100 mean extremeness scores (2 profiles × 50 traits) for each data set are overwhelmingly positive in sign, indicating a strong tendency for the estimates to be too extreme. First consider the American data: Of the 69 extremeness scores that are significantly different from zero (by a two-tailed t-test, $p < 0.05$), only 2 are negative in sign. The unweighted mean extremeness score is 7.5 points, which is significantly greater than zero, $t(99) = 11.9$, $p < 0.001$. Weighting the mean by the number of valid inferences for each trait yields an even greater extremeness score. The results are similar for the Israeli data. All 85 extremeness scores that are significantly different from zero are positive in sign. The (unweighted) mean extremeness score for the Israeli inference data is 12.0 points, which is significantly greater than that of the Americans, $t(99) = 5.94$, $p < 0.0001$. The Israeli estimates yield larger extremeness scores, $t(197) = 4.94$, $p < 0.001$, because the Israeli target ratings have means that are closer to the grand mean of 50 than those of the American targets.

In summary, the findings from the extremeness analysis indicate that subjects' estimates were valid in direction but too extreme in magnitude. Subjects appeared to follow a strategy Kahneman and Tversky (1973) call 'prediction by evaluation,' in which the role of regression is ignored as the extremeness of the impression determines the extremeness of the judgment. We test this interpretation below.

3.1.3. Similarity analysis

We hypothesize that people's trait inferences are based to a large extent on the semantic or conceptual relatedness between the trait under consideration and the information used as the basis of judgment. So, for example, a person might evaluate the likelihood that an extravert is cheerful on the basis of the similarity in meaning between the term *extravert* and the term *cheerful*. The preceding analysis is consistent with this notion since prediction by similarity or representativeness is generally nonregressive.

To obtain a more direct test of this hypothesis, we collected similarity ratings from a group of 232 students enrolled in an introductory psychology course at Stanford University and a group of 126 mathematics students at Tel-Aviv University. American subjects were presented with one of 6 concepts corresponding to the poles of the 3 personality dimensions used as profile information, and were asked to rate the similarity in meaning of all 50 traits to that concept. Israeli subjects considered 2 concepts, paired in the following way within a questionnaire: Introvert and Adaptive ($n = 40$), Decisive and Intuitive ($n = 44$), or Analytic and Extravert ($n = 42$). These subjects first rated the similarity of all 50 traits to one concept and then rated their similarity to the second concept.

The questionnaire was entitled 'Linguistic Similarity Judgments'. The following is an excerpt from the instructions:

> Consider the concept ANALYTIC:
> ANALYTIC: Prefer to act through a logical, step-by-step process.
> This concept is often contrasted with:
> INTUITIVE: Prefer to act on inspiration or imagination.
> Imagine that as part of the preparation of a thesaurus of personality trait terms, you are asked to judge the similarity of different terms. Below you will find a number of trait terms in italics (e.g. *soft-spoken, unconventional,* etc.). For each term, consider how similar it is in meaning to the concept ANALYTIC. Are the meanings similar, unrelated, or rather opposite?

Subjects were asked to rate each trait using a 7-point scale, ranging from −3 (rather opposite meaning) through 0 (unrelated meaning) to +3 (similar meaning).

To simplify the task, we asked subjects to rate the similarity of the trait terms to a single pole of one dimension (with the opposite pole provided as a contrast). We computed an overall similarity measure between a specific profile (IAD or EID) and each trait by taking the sum of the mean similarity ratings for each of the 3 component poles. [3]

For each profile, we computed the correlation between the mean similarity ratings and the mean trait estimates for the 50 traits. Table 2 shows the correlations for both the American and the Israeli data. The similarity ratings accounted for a large proportion (approximately 80 percent) of the variance in the trait estimates. The correlations between the mean target ratings and the mean trait estimates were also quite high (accounting for roughly 50 percent of the variance), but not as high as the correlations between the similarity ratings and the trait estimates. With the exception of the EID profile for the American data, similarity was a significantly better predictor of the trait estimates than were the target ratings, $p < 0.005$ for all three comparisons. The mean similarity ratings were also highly correlated with the mean target ratings for the 50

Table 2
Correlations among mean trait estimates (est), actual target self-ratings (act), and similarity ratings (sim), listed by sample and profile information. A dot (·) indicates the covariate in a semi-partial correlation

	American sample		Israeli sample	
	EID profile	IAD profile	EID profile	IAD profile
r(est, sim)	0.89	0.92	0.85	0.91
r(est, act)	0.83	0.56	0.66	0.66
r(sim, act)	0.74	0.53	0.70	0.62
r(est, sim·act)	0.41	0.74	0.55	0.63
r(est, act·sim)	0.25	0.09	0.09	0.12

[3] Other methods of combining the similarity ratings, such as taking the largest absolute deviation or the maximum similarity score across the three relevant poles, produce the same results.

traits, accounting for about 40 percent of the variance. Thus, in this task at least, conceptual relatedness is a reasonably good predictor of statistical association.

Table 2 also lists semi-partial correlations predicting the trait estimates when controlling for either similarity or for the target ratings. The table shows that the semi-partial correlation between similarity (controlling for the target rating) and the trait estimate is consistently greater than that between the target rating (controlling for similarity) and the trait estimate. In other words, not only is mean similarity an excellent predictor of the trait estimates, it is also a very good predictor of inferential errors. The fact that the correlation between trait estimates and target ratings is still nonzero when similarity is controlled for, however, shows that there is at least some systematic variance in the trait estimates not accounted for by the similarity judgments.

These results support the hypothesis that trait estimates are based largely on the similarity in meaning between the profile information and the trait under consideration. To the extent that there are cross-cultural differences in perceived similarity, such an interpretation implies that similarity ratings obtained from a given culture should correlate more highly with trait estimates from that culture than with trait estimates from another culture. Testing of this prediction is complicated by the fact that the American and Israeli similarity ratings are in fact highly correlated (in the aggregate), $r = 0.83$. To better separate the similarity ratings for the two samples, we computed semi-partial correlations to remove the variance attributable to the Israeli covariate from the American similarity ratings and vice versa. The American similarity ratings (controlling for Israeli similarity) were still strongly correlated ($r = 0.47$) with the American trait estimates but were only weakly correlated ($r = 0.10$) with the Israeli trait estimates. Likewise, the Israeli similarity ratings (controlling for American similarity) were still strongly correlated ($r = 0.42$) with the Israeli estimates but only weakly correlated ($r = 0.05$) with the American estimates. As expected, the similarity ratings from a given culture better accounted for the trait estimates obtained from that culture than they did the estimates obtained from a different culture. The same pattern was found for the actual target ratings, but with smaller magnitude correlations: The American similarity ratings (controlling for Israeli similarity) still correlated ($r = 0.32$) with the American target ratings but were only marginally correlated ($r = 0.09$) with the Israeli target ratings. Likewise, the Israeli similarity ratings (controlling for American similarity) still correlated ($r = 0.29$) with the Israeli target ratings but were essentially uncorrelated correlated ($r = 0.03$) with the American target ratings.

Finally, we re-analyzed the extremeness scores for the 50 traits (for both profiles) in terms of the similarity measure. Our hypothesis implies that people should give more extreme scores when conceptual relatedness (positive or negative) is high. We first computed the absolute values of the similarity measures between each profile and each of the 50 traits, such that high values indicate a strong degree of semantic relationship (i.e., the trait is rated as either very similar or quite opposite to the meaning of the profile components). Each trait-profile pair was then classified as strongly related (e.g., in the American data, *practical* and *unpredictable* were strongly related to the IAD profile) or weakly related (e.g., *tolerant* was weakly related to the IAD profile) on the basis of a median split of the absolute similarity scores. For the American subjects, the mean extremeness score for the strongly related trait-profile pairs ($M = 9.4$) was

significantly greater than the mean extremeness score for the weakly related trait-profile pairs ($M = 5.7$), $t(98) = 3.05$, $p < 0.005$. The same result held for the Israeli subjects: The mean extremeness score for the strongly related trait-profile pairs ($M = 14.3$) was significantly greater than the mean extremeness score for the weakly related trait-profile pairs ($M = 9.8$), $t(98) = 3.63$, $p < 0.001$. This result rules out the hypothesis that subjects gave estimates that were uniformly too extreme, perhaps because they were trying to be informative (i.e., giving inferences that were maximally different from the grand mean) at the expense of accuracy. It appears instead that similarity was used as the basis for inference, and that in cases of high similarity subjects were led to make extreme trait estimates.

An alternative interpretation of these results is that subjects in the Similarity Group based their ratings on the statistical association between the concept (i.e., the dimension pole) and the trait in question rather than on their similarity. The resulting high correlation would thus be due to the fact that subjects in the Inference Group and in the Similarity Group were judging the same thing. This interpretation seems unlikely, however, given the steps we took to keep the Similarity Group focused on semantic rather than statistical association. First, the questionnaire was described not as a prediction task, but rather as a psycholinguistic study of similarity in meaning being conducted as part of the compilation of a thesaurus of personality trait terms. Second, subjects were asked to consider only one pole of one dimension rather than a complete profile, which would be more likely to induce ratings in terms of statistical association. Third, subjects gave their ratings on a 7-point scale that neither implies nor is easily translated into a probability measure. While these preventative steps cannot eliminate the possibility that some subjects tried to rate statistical rather than semantic association, it seems more plausible to us that the high correlation between the similarity ratings and trait estimates arises because the trait estimates are similarity-based.

3.2. Uncertainty ranges

One of the main purposes of this research was to compare the accuracy of people's judgments in the trait inference task with their confidence in those inferences. Recall that subjects in the Inference Group were asked to set a 90 percent uncertainty range around their trait estimates. We have two main questions concerning these data. First, did subjects set well-calibrated uncertainty ranges in both the *individual* and *group* conditions? Second, how does confidence, as measured by the size of the uncertainty range, vary as a function of conceptual relatedness?

To understand the relationship between confidence and accuracy in the *individual* and *group* conditions, it is helpful to view the uncertainty involved in the task as arising from two independent sources. First, there is uncertainty about the relationship between the profile information and the trait being predicted. Second, there is uncertainty arising from individual variability within the profile group once the relationship between the information and the trait has been established. To the extent that subjects within a profile group vary in their trait ratings, subjects in the *individual* condition should set wider intervals than subjects in the *group* condition to compensate for the added uncertainty of this second component. If subjects make their inferences solely on the basis of

conceptual relatedness, however, then the number of targets under consideration should not have much effect on the uncertainty range, because the degree of similarity between a profile and a trait is unaffected by sample size (cf. Kahneman and Tversky, 1972).

In fact, American subjects asked to estimate a single target subject's ratings did set significantly wider uncertainty ranges ($M = 51.2$) than did subjects estimating the mean of the entire profile group ($M = 45.5$), $t(198) = 8.43$, $p < 0.0001$, indicating that the subjects were not insensitive to the influence of this factor on the accuracy of their inferences. The magnitude of the difference, however, is not nearly enough to compensate for the added uncertainty of predicting a single target subject, as is shown below.

If subjects were setting well-calibrated uncertainty ranges, then the target rating should fall inside their ranges about 90 percent of the time. Only 10 percent of the target ratings should lie outside the range. These occurrences are often referred to as 'surprises' (Alpert and Raiffa, 1982). To the extent that subjects are overconfident in their trait estimates, however, we should find a substantially greater surprise rate. In fact, subjects did tend to set uncertainty ranges that were too narrow. In the *group* condition for American subjects, the mean target rating for the profile fell outside the uncertainty range 26 percent of the time. This is significantly greater than the 10 percent surprise rate expected for well-calibrated ranges, $t(19) = 3.70$, $p < 0.005$, indicating overconfidence.

American subjects estimating a single target's ratings had a surprise rate of 44 percent, which is significantly greater than that of subjects in the *group* condition, $t(36) = 3.09$, $p < 0.01$. The surprise rate for the Israeli subjects, all of whom were asked to consider a single target subject, was 57 percent, which is considerably higher than that of their American counterparts. The difference in surprise rate between the two samples was caused primarily by the fact that the Israeli subjects set significantly narrower uncertainty ranges ($M = 39$ percentile points) than did the American subjects ($M = 48$ percentile points), $t(99) = 21.43$, $p < 0.0001$.

We interpret these results as suggesting that subjects in the *individual* condition based their uncertainty ranges almost entirely on their uncertainty regarding the relationship between the profile information and the trait under consideration, and essentially overlooked the added uncertainty stemming from individual variation within a given profile group. This result could be justified only if the subjects believed there is little or no variability among trait ratings provided by different subjects within a profile group. The following analysis excludes this interpretation.

We asked a separate group of Stanford University undergraduates ($N = 68$) to estimate the within-group variability of the trait ratings. In this task, subjects were provided with a profile description and that profile group's mean score on each trait. They were asked to set an interval around the mean which they believed to contain 90 percent of the ratings provided by individual subjects within the profile group. Half the subjects gave these intervals for the EID profile; the other half evaluated the IAD profile.

By combining these variability estimates with the uncertainty ranges provided by the Inference Group subjects in the *group* condition, we can estimate how wide the (American) subjects in the *individual* condition should have set their uncertainty ranges had they considered both sources of uncertainty. Variance estimates for the *individual*

condition should equal the sum of the variance of responses within a profile (as estimated by the variability ratings) and the variance of the estimated mean target rating for the profile (from the *group* condition subjects' uncertainty ranges). [4] Results from this analysis show that the American subjects in the *individual* condition should have set ranges that were nearly 50 percent wider than those set by the *group* condition subjects, given the estimated additional amount of uncertainty expected from predicting a single target subject rather than the entire group. Compared to this expected difference, the relatively small observed difference (*individual* condition subjects set ranges that were about 10 percent wider) reveals that subjects in the *individual* condition did not compensate adequately for the added uncertainty of predicting a single target subject.

To examine the relation between similarity and confidence, we calculated the correlation between the width of the uncertainty range set around a trait estimate and the rating of similarity between the trait in question and the profile information. The following analyses were all performed at the aggregate level, based, for each profile, on the mean absolute error, uncertainty range, and similarity rating of each trait. We again used the absolute value of the similarity ratings such that a greater score indicates stronger conceptual relatedness between the trait and the profile, regardless of its sign. For the American data, the similarity between the trait and the profile information was found to be a better predictor ($R^2 = 0.46$) of uncertainty than was the accuracy (as measured by absolute error) of the inference ($R^2 = 0.16$), accounting for almost three times as much variance. For the Israeli data, however, similarity proved to be only a marginally better predictor ($R^2 = 0.22$) of confidence than was accuracy ($R^2 = 0.18$). These results suggest that similarity affects confidence, though not as strongly as it affects the trait estimates themselves.

Finally, we turn to the direction of the relationship between confidence as measured by the uncertainty range and accuracy as measured by absolute error. The correlations were -0.40 for the American data and -0.42 for the Israeli data. The negative correlation indicates that as confidence increased, accuracy *decreased.* [5] This is perhaps the most remarkable consequence of judgment by similarity: People were most confident when they were least accurate. To understand this intriguing pattern, note that because there is only a moderate statistical association between the profile information and the traits under consideration, most of the target means fall between the 40th and 60th percentiles. As a result, any extreme estimates are likely to deviate considerably from the target mean. Because high absolute similarity ratings are associated both with more extreme inferences and with narrower uncertainty ranges, the inevitable result is that people will make inaccurate inferences and express a high degree of confidence under

[4] Variance estimates for the *group* condition subjects and for subjects in the variability estimation task were derived assuming a normal distribution.

[5] Because the scale is bounded (at 0 and 100), it could be the case that when people make trait estimates near either of the scale (as they tend to do when similarity is high), they appear more confident (set narrower ranges) because one end of the interval is necessarily bounded by the end of the scale. To rule out this possibility we computed a new measure of range width (which was set equal to the greater of either the high estimate minus the trait estimate or the trait estimate minus the low estimate) that would not encounter the boundedness problem, and obtained essentially the same results.

the same set of circumstances, namely when there is a strong conceptual relationship between the profile information and the trait being predicted. Note that these conclusions assume only that narrower uncertainty ranges are associated with greater confidence, and do not depend on the numerical interpretation of the uncertainty ranges as 90% credible intervals.

One might argue that subjects used their uncertainty ranges to assess correlational accuracy rather than absolute error (despite explicit instructions to the contrary). To examine this possibility, each subject's mean confidence (uncertainty range width), mean absolute error, and correlational accuracy was computed over the 50 traits. Note that this analysis tests mean-level differences between subjects across the set of 50 traits rather than a given subject's ability to discriminate accurate from inaccurate estimates within the set of 50 traits. Over subjects, mean confidence correlated more highly with the absolute error measure ($r = -0.225$ and -0.394 for the Israeli and American data, respectively) than it did with the correlational accuracy measure ($r = -0.030$ and 0.270 for the Israeli and American data, respectively). Thus this analysis reinforces that above and renders implausible the alternative that subjects were assessing correlational accuracy.

4. Discussion

In summary, the results indicate that subjects in the present study made trait estimates that: (a) were positively correlated with the target ratings and deviated in the correct direction from the grand mean considerably more often than would be expected by chance; (b) were generally too extreme, to such an extent that they would have been more accurate in terms of absolute deviation had subjects simply responded with the grand mean for each trait; (c) were more strongly correlated with the similarity ratings than with the actual target ratings; (d) were made with greater confidence than their accuracy warranted, as indicated by overly narrow uncertainty ranges; and (e) were less accurate when expressed with high confidence than when expressed with low confidence. These results support the hypothesis that subjects based their trait estimates on the conceptual similarity between the traits and the profile information.

Furthermore, the overall pattern of results was strikingly similar in the American and Israeli samples, in that (a) through (e) above held for both samples, despite considerable cultural and academic differences between the two groups and the imperfections of the translation from English to Hebrew. Furthermore, differences in the conceptual similarity ratings between the two samples predicted corresponding differences in the trait estimates. This suggests that the process by which intuitive trait inferences are derived from the conceptual structure of intuitive trait constructs may be similar across cultural variations in lay personality theory (cf. Church and Katigbak, 1989; Gidron et al., 1993). There were only three results that differed between the American and Israeli data. First, the mean trait ratings for the Israeli target subjects tended to fall nearer to the 50th percentile, resulting in trait inferences that tended to be more extreme than in the American sample. Second, the Israeli subjects tended to set narrower uncertainty ranges, resulting in a higher rate of 'surprises'. Third, the similarity ratings better predicted

uncertainty range width in the American data than in the Israeli data. While the cause of these differences is unclear, they are in any case overshadowed by the major findings common to both samples.

In this final section, we first address several questions regarding the generalizability of our results and then offer some speculations concerning how people go about making trait inferences. One possible limitation of the current study is that subjects were asked to estimate a target's self-rated standing on a list of traits, while in everyday life people might be more concerned with predicting behavior. Previous research (see, e.g., Ross and Nisbett, 1991, ch. 4; Shweder, 1975; Shweder and D'Andrade, 1980), however, suggests that self-ratings are more predictable than are behavioral measures, which in turn implies that asking subjects to predict a target's behavior rather than his or her self-ratings will only lower predictive accuracy. Predicting behavior will be more difficult because behavioral consistency across situations (e.g., stealing loose change, cheating on a test, lying to protect somebody) that would appear to be governed by a given trait concept (e.g., honesty) is typically quite low (e.g., Hartshorne and May, 1928; Mischel, 1968; Newcomb, 1929). Self-ratings, in contrast, are likely to draw to some extent on the same, culturally-shared theory of personality used in making trait inferences about others. As a result, the use of self-ratings as the criterion will generally overestimate the accuracy likely to be achieved in behavioral prediction.

A second issue involves the information available to our subjects as a basis for inference. Admittedly, this information is relatively impoverished compared to what is often available in everyday social judgment. An obvious criticism is that people are likely to have much more complex representations of others than the three pieces of information we provided to subjects in the Inference Group. We have three responses to this criticism. First is the methodological justification for our approach: Asking people to make judgments on the basis of highly structured and limited data allows straightforward conclusions about how the judgments were made. Second, although the information presented to subjects differs in form from that used in everyday judgment, it is not necessarily inferior in quality: People are quite willing to make judgments about others on the basis of presumably less diagnostic information such as appearance or mannerisms, not to mention astrological signs. Third, previous studies using highly complex input data have yielded results that are quite consistent with those reported here. Griffin, Ross, and colleagues (Dunning et al., 1990; Vallone et al., 1990), for example, found considerable overconfidence when they asked subjects to make predictions about their roommates or about people they had previously interviewed. These studies give no indication that the calibration of people's confidence assessments improves with increasingly complex target information.

A related objection is that, because subjects had nothing else to base their judgments on, it is not particularly surprising that their judgments were based on the conceptual relatedness between the profile information and the traits under consideration. The question addressed in this article, however, is not whether subjects will use conceptual relatedness as a cue, but how much weight they will give to these data in the context of a highly uncertain prediction task. Thus it is the degree rather than the direction of divergence from the baserate that is of primary interest. As an analogy, consider the task of predicting college grades from elementary school grades. Here, too, the direction of

the relationship is obvious – higher elementary school grades are generally associated with higher college grades – but the extent to which the extremeness of the prediction should be attenuated in the face of the low correlation between the two variables is much less obvious.

Finally, there are two potential methodological criticisms. First, one might argue that subjects in the *individual* condition did not believe the targets had been selected at random, and gave judgments on the assumption that the targets were especially typical or representative of their profile group (cf. Gigerenzer et al., 1988). Such an account, however, fails to explain the similar results obtained in the *group* condition, which involves no assumption of random selection. Second, previous researchers have suggested that the predominance of overconfidence reported in the literature is caused by the selection of particularly difficult or counterintuitive items (Gigerenzer et al., 1991; May, 1986). This objection, we argue, does not apply to the present study. Because our subjects were required to set uncertainty ranges rather than to assign a probability to the correctness of a given proposition, it is unclear how item difficulty should be defined in this task or how item selection will affect overconfidence.

We should note that conceptual relatedness is in many cases a reasonably predictive guide to statistical association. For example, one can be quite certain that a person who rates herself as *tidy* will also rate herself as *neat*, because the two terms have essentially the same meaning. On the other hand, the traits *tidy* and *punctual* have a relatively weak statistical association even though most people associate the two conceptually. But it is not the case that people fail to recognize that *tidy* is more closely related (conceptually or empirically) to *neat* than to *punctual*. Instead, we suggest, people generally underestimate how quickly the strength of empirical association degrades as a function of conceptual relatedness. While in reality only the strictest of synonyms show high statistical association (and sometimes not even then, see Goldberg and Kilkowski, 1985), people appear to expect statistical association to directly reflect conceptual relatedness, such that moderately related trait terms have a moderate correlation, and so on. We would suggest, then, that with a better understanding of the relation between perceived similarity and statistical association, and with greater attention given to statistical factors that affect accuracy but not similarity, trait inferences might be made more accurately and with a more appropriate degree of confidence. In the absence of corrective procedures, however, people's intuitive judgments are unlikely to be brought into line with the statistical regularities of their social world, despite the fact that such judgments are obviously frequent and important in everyday life.

As mentioned in the introduction, the current study was not intended as a direct test of the Shweder and D'Andrade (1980) systematic distortion hypothesis. Our results are in fact quite consistent with their argument, but can also be accommodated by alternative accounts of implicit personality theory. For example, Borkenau (1986; see also Block et al., 1979; Borkenau, 1992; Romer and Revelle, 1984) has proposed an 'overlap hypothesis' according to which the conceptual relatedness of two traits determines (and may be determined by) the set of behavioral acts that are instances of both traits. Thus, for example, an individual who consistently initiates conversations with others will be labeled as above average both in talkativeness and in sociability, and vice versa for someone who does not; the fact that many traits refer to overlapping sets of

behaviors will thereby lead to correspondence between trait ratings and similarity ratings. This approach differs from that of Shweder and D'Andrade (1980) in its predictions concerning the relationship between behavioral measures and trait and similarity ratings, but we did not obtain behavioral measures in the current study.

In this study, the confidence with which people made their inferences was negatively correlated with their accuracy, a finding that has been demonstrated in other research as well. Kahneman and Tversky (1973) noted that because the input information seems more coherent, predictions based on correlated cues are made with greater confidence than are predictions based on uncorrelated cues, even though the use of uncorrelated cues generally yields more accurate judgments. Brenner et al. (in press-b) found that subjects asked to predict the jury vote in a legal case were more confident but less accurate when given the arguments presented by one rather than both sides, presumably because the arguments from a single side presented a more coherent and less uncertain account of the case. Peterson and Pitz (1988) have also demonstrated that confidence decreases with the addition of inconsistent evidence, even as accuracy increases (cf. Ganzach, 1994). In such studies, evidence that seems consistent apparently instills a greater sense of certainty than does more informative but less consistent evidence. The results of the current study as well as these previous studies are compatible with the argument by Griffin and Tversky (1992) that confidence is determined primarily by the strength of the impression conveyed by the available evidence with insufficient regard given to the weight or credence of that evidence. (For other examples of negative confidence–accuracy relationships, see Sniezek and Buckley, 1993.)

Confidence will be high whenever a 'good fit' is established between the outcome variable and the information being used for judgment. The goodness of fit is determined by the way the problem domain is represented, which in the case of trait inference is heavily influenced by the culturally-shared lay theory of personality. The domain of trait psychology is highly complex and inherently unpredictable due to the generally weak statistical links among intuitive trait constructs. In such a domain, lay theories will almost certainly overestimate the coherence of the underlying structure, causing misperceptions of the predictability it affords. In particular, when strong conceptual relatedness leads to an extreme inference, confidence is likely to be high because of the good conceptual fit but accuracy is likely to be low. Even as the lay theory of personality provides what perhaps may be the only available basis for intuitive trait inference, it also establishes the conditions under which inferential errors are likely to be made with high confidence.

Acknowledgements

This research was supported in part by Grant MH53046 from the National Institutes of Health to Amos Tversky and Grant 92-00389 from the United States–Israel Binational Science Foundation to Varda Liberman and Amos Tversky. Lyle Brenner was supported by a National Science Foundation Graduate Fellowship and Derek Koehler by a National Defense Science and Engineering Graduate Fellowship during the course of

this research. We thank Hae-June Ahn, Mark Forehand, and Daniel Levitin for their assistance in data collection and analysis.

References

Alpert, M. and H. Raiffa, 1982. 'A progress report on the training of probability assessors'. In: D. Kahneman, P. Slovic and A. Tversky (eds.), Judgment under uncertainty: Heuristics and biases (pp. 294–305). Cambridge: Cambridge University Press.

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

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., 1992. Implicit personality theory and the five-factor model. Journal of Personality 60, 295–327.

Borkenau, P. and F. Ostendorf, 1987. Fact and fiction in implicit personality theory. Journal of Personality 55, 415–443.

Brenner, L.A., D.J. Koehler, V. Liberman and A. Tversky, in press-a. Overconfidence in probability and frequency judgments: A critical examination. Organizational Behavior and Human Decision Processes.

Brenner, L.A., D.J. Koehler and A. Tversky, in press-b. On the evaluation of one-sided evidence. Journal of Behavioral Decision Making.

Chapman, L.J. and J.P. Chapman, 1967. Genesis of popular but erroneous diagnostic observations. Journal of Abnormal Psychology 2, 193–204.

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

Church, A.T. and M.S. Katigbak, 1989. Internal, external, and self-report structure of personality in a non-Western culture: An investigation of cross-language and cross-cultural generalizability. Journal of Personality and Social Psychology 57, 857–872.

D'Andrade, R.G., 1965. Trait psychology and componential analysis. American Anthropologist 67, 215–228.

Dawes, R.M., D. Faust and P.E. Meehl, 1988. Clinical versus actuarial judgment. Science 243, 1668–1674.

deSoto, C.B., M.M. Hamilton and R.B. Taylor, 1985. Words, people, and implicit personality theory. Social Cognition 3, 369–382.

Dunning, D., D.W. Griffin, J.D. Milojkovic and L. Ross, 1990. The overconfidence effect in social prediction. Journal of Personality and Social Psychology 58, 568–581.

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

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

Ganzach, Y. 1994. Inconsistency and uncertainty in multiattribute judgment of human performance. Journal of Behavioral Decision Making 7, 193–211.

Gidron, D., D.J. Koehler and A. Tversky, 1993. Implicit quantification of personality traits. Personality and Social Psychology Bulletin 19, 594–604.

Gigerenzer, G., U. Hoffrage and H. Kleinbolting, 1991. Probabilistic mental models: A Brunswikean theory of confidence. Psychological Review 98, 506–528.

Gigerenzer, G., W. Hell and H. Blank, 1988. Presentation and content: The use of base rates as a continuous variable. Journal of Experimental Psychology: Human Perception and Performance 14, 513–525.

Goldberg, L.R., 1982. 'From Ace to Zombie: Some explorations in the language of personality'. In: C. D. Spielberger and J. N. Butcher (eds.), Advances in personality assessment (pp. 203–234). Hillsdale, NJ: Erlbaum.

Goldberg, L.R., 1990. An alternative 'description of personality': The Big-Five factor structure. Journal of Personality and Social Psychology 59, 1216–1229.

Goldberg, L.R., 1991 (November). The structure of personality traits: Vertical and horizontal aspects. Invited address to the Longitudinal Research Conference, Palm Springs, CA.

Goldberg, L.R. and J.M. Kilkowski, 1985. The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. Journal of Personality and Social Psychology 48, 82–98.

Griffin, D. and A. Tversky, 1992. The weighing of evidence and the determinants of confidence. Cognitive Psychology 24, 411–435.

Hartshorne, H. and M.A. May, 1928. Studies in the nature of character, I: Studies in deceit. New York: Macmillan.

Jackson, D.N., D.W Chan and L.J. Stricker, 1979. Implicit personality theory: Is it illusory? Journal of Personality 47, 1–10.

Jackson, D.N. and L.J. Stricker, 1982. Is implicit personality theory illusory? Armchair criticism versus replicated empirical research. Journal of Personality 50, 240–244.

Jennings, D.L., T.M. Amabile and L. Ross, 1982. 'Informal covariation assessment: Data-based versus theory-based judgments'. In: D. Kahneman, P. Slovic and A. Tversky (eds.), Judgment under uncertainty: Heuristics and biases (pp. 211–230). Cambridge: Cambridge University Press.

Kahneman, D., P. Slovic and A. Tversky, 1982. Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.

Kahneman, D. and A. Tversky, 1972. Subjective probability: A judgment of representativeness. Cognitive Psychology 3, 430–454.

Kahneman, D. and A. Tversky, 1973. On the psychology of prediction. Psychological Review 80, 237–251.

May, R.S., 1986. 'Inferences, subjective probability, and the frequency of correct answers: A cognitive approach to the overconfidence phenomenon'. In: B. Brehmer, H. Jungermann, P. Lourens and G. Sevo'n (eds.), New directions in research on decision making (pp. 175–189). Amsterdam: North-Holland.

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

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, 201–222.

Mischel, W., 1968. Personality and assessment. New York: Wiley.

Mulaik, S.A., 1964. Are personality factors raters' conceptual factors? Journal of Consulting and Clinical Psychology 28, 279–289.

Newcomb, T.M., 1929. The consistency of certain extrovert–introvert behavior patterns in 51 problem boys. New York: Columbia University, Teachers College, Bureau of Publications.

Nisbett, R.E. and L. Ross, 1980. Human inference: Strategies and shortcomings of social judgment. 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.

Oskamp, S., 1965. Overconfidence in case-study judgments. Journal of Consulting Psychology 29, 261–265.

Peabody, D., 1987. Selecting representative trait adjectives. Journal of Personality and Social Psychology 52, 59–71.

Peterson, D.K. and G.F. Pitz, 1988. Confidence, uncertainty, and the use of information. Journal of Experimental Psychology: Learning, Memory, and Cognition 14, 85–92.

Romer, D. and W. Revelle, 1984. Personality traits: Fact or fiction? Journal of Personality and Social Psychology 47, 1028–1042.

Ross, L. and R.E. Nisbett, 1991. The person and the situation: Perspectives of social psychology. New York: McGraw-Hill.

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

Shweder, R.A., 1977. Likeness and likelihood in everyday thought: Magical thinking in judgments about personality. Current Anthropology 18, 637–658.

Shweder, R.A. and R.G. D'Andrade, 1979. Accurate reflection or systematic distortion? A reply to Block, Weiss, and Thorne. Journal of Personality and Social Psychology 37, 1075–1084.

Shweder, R.A. and R.G. D'Andrade, 1980. 'The systematic distortion hypothesis'. In: R.A. Shweder (ed.), New directions for methodology of social and behavioral science, 4: Fallible judgment in behavioral research (pp. 37–58). San Francisco, CA: Jossey-Bass.

Sniezek, J.A. and T. Buckley, 1993. 'Becoming more or less uncertain'. In: N.J. Castellan Jr. (ed.), Individual and group decision making: Current issues (pp. 87–108). Hillsdale, NJ: Erlbaum.

Tversky, A. and D. Kahneman, 1983. Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review 91, 293–315.

Vallone, R.P., D.W. Griffin, S. Lin and L. Ross, 1990. Overconfident prediction of future actions and outcomes by self and others. Journal of Personality and Social Psychology 58, 582–592.

Weiss, D.S. and G.A. Mendelsohn, 1986. An empirical demonstration of the implausibility of the semantic similarity explanation of how trait ratings are made and what they mean. Journal of Personality and Social Psychology 50, 595–601.