Judging Social Behavior Using "Thin Slices"

Nalini Ambady and Robert Rosenthal

In our daily lives we make judgments regarding other people intuitively, unwittingly, and rapidly. A fleeting glimpse or a mere glance lead to evaluative judgments of others. Such judgments have important consequences. They influence whom we sit next to on the subway, whom we hire for a job, and sometimes even whom we marry. Thus, as employers, within a minute into a job interview we decide that we have found the ideal candidate. As employees, within a minute of meeting a future boss, we decide that we should start perusing the employment advertisements. How accurate are these judgments? Is the candidate really ideal and is the boss really an ogre?

Social psychologists have long been interested in issues relating to judgmental accuracy. And for a long time social psychologists concluded, based on innovative and intriguing research, that human beings are not very good judges. Our judgments are prone to a depressing assortment of errors and biases. To name just a few, we tend to be insensitive to the base rates of a condition in the population, to ignore information regarding the representativeness of samples, and to rely on judgmental heuristics that can be shown to lead to uninformed and poor decisions.

Preceding and parallel to the work on judgment and decision making, another line of research suggested that people (and even animals) can be quite accurate in making certain types of judgments. For example, personality and social psychologists found that people are quite good at judging certain aspects of others' personality, and these judgments can be accurate at predicting certain variables such as extroversion even when there is absolutely no interaction between the targets and raters—when impressions are based on five-minute video clips of the targets. Furthermore, research on expectancy effects documented very clearly that people and even animals accurately can sense the subtle, unstated expectations that other people have of them and behave according to those expectations. Systematic studies showed that people can accurately pick up and interpret cues from the nonverbal behavior of others. In the light of this research, it seemed surprising and paradoxical that social psychologists were lamenting the inaccuracy of social and perceptual judgments. The difference between the two streams of work, of course, was the domain being examined—judgments regarding abstract information and judgments regarding other people.

In this article we report on a research program attempting to evaluate the accuracy of judgments of others from minimal information. To date, this program suggests that people are remarkably accurate in the consensual judgments that they make based on very little information—mere 10–30-second silent video clips.

The first step in our research program was to identify a satisfactory criterion variable. We were not very happy with using self-report measures on questionnaires, the criterion most often used by personality psychologists. Ample evidence showing that people are not very accurate at judging themselves suggests that self-report measures can often be biased and inaccurate. We, therefore, searched for a more satisfactory criterion variable that had pragmatic utility. We were fortunate to find one that was ecologically valid. We selected a real-life naturalistic sample of teachers in the classroom and a criterion that seemed to be both ecologically valid and of considerable consequence in the real world—

teacher effectiveness usually measured by any one or some combination of student ratings, peer ratings, and supervisor ratings. Such evaluations are important because of their practical significance: They influence salary, promotion, and tenure decisions. Two studies examining whether teacher effectiveness can accurately be gauged from very brief observations of behavior are described here.

Study 1: Can Teaching Effectiveness Be Identified From Minimal Information?

In the first study we had judges rate 13 different teachers on the basis of three 10-second silent video clips of each teacher on several different personality dimensions. Permission was obtained to use videotapes of classes taught by 13 teachers at the Teaching Laboratory at Harvard University. The teachers (six female, seven male) were all graduate student Teaching Fellows who were videotaped while teaching sections for undergraduate courses at the Teaching Laboratory, in order to receive feedback about their teaching from specialists at the Center. Each section lasted for approximately one hour. The sections covered diverse areas of the curriculum: eight of the sections were in the area of the humanities and social sciences, three in the natural sciences, and two dealt with languages. The number of students in the sections varied from eight to over 20. The sample of teachers was selected by members of the Teaching Center from their library of videotapes. The criterion used for selection was that the sample should represent a fairly wide range in their ratings on the teaching effectiveness criterion measure. The experimenters were blind to the effectiveness ratings of the teachers until the final stage of data analysis.

We used end-of-the-semester ratings by students as a measure of teaching effectiveness. Although student achievement (adjusted for student ability) might be the best possible criterion of effective teaching, it is very difficult to obtain such data. Teacher effectiveness in the real world is often evaluated solely on the basis of ratings by supervisors and/or students. Considerable evidence supports the validity of student evaluations of teacher effectiveness: Student ratings seem to be consistent over time and across raters; correlate positively with expert, colleague, and administrator ratings; be independent of extraneous characteristics or characteristics of the students themselves; correlate significantly with how much students actually learn; and, last, not change appreciably with greater age and reflection. Ratings, compiled by the Committee on Undergraduate Education (CUE), were obtained by averaging the responses to the following items: "Rate the quality of the section overall" and "Rate section leader's performance overall." We were provided with the average of the two ratings for each teacher in our sample but not with the two separate ratings. Because some teachers were rated on a five-point scale and others on a seven-point scale, these ratings were converted into percentages for this analysis, which varied from 41.67% to 91.67%. Only after all the videotape-based data had been collected did we receive the ratings for the sample of teachers in this study.

One master videotape was derived from the 13 individual videotapes using the following method: For each teacher, 10 seconds from the first 10 minutes of the class focusing on the teacher alone (without any of the students), 10 seconds from the middle of the class with the teacher alone, and 10 seconds from the last 10 minutes of the class with the teacher alone were assembled and rerecorded onto one videotape. The final stimulus tape contained 39 ten-second clips—3 clips for each of 13 teachers.

Nine female undergraduates were paid to rate the 39 clips of the teachers individually. They were told that they would see short segments of teachers teaching a class and would be asked to rate the nonverbal behavior of the teacher in each segment on 15 dimensions on a scale running from 1 (not at all) to 9 (very). Besides these instructions, the judges received no other training. Previous research has demonstrated that naive subjects are able reliably to make such ratings without training. The dimensions rated were not accepting/accepting, not active/active, not anxious/anxious (this dimension was reverse scored), not attentive/attentive, not competent/competent, not confident/confident, not dominant/dominant, not empathic/empathic, not enthusiastic/enthusiastic, not honest/honest, not lik-

Table 1—Correlations of Molar Nonverbal Behaviors With College Teacher Effectiveness Ratings (student ratings)

Variable	*r*	95% confidence interval	
		From	To
Accepting	.50	−.07	.82
Active	.77	.38	.93
Attentive	.48	−.09	.82
Competent	.56	.01	.85
Confident	.82	.49	.94
Dominant	.79	.42	.93
Empathic	.45	−.13	.80
Enthusiastic	.76	.36	.92
Honest	.32	−.28	.74
Likable	.73	.30	.91
(Not) anxious	.26	−.34	.71
Optimistic	.84	.54	.95
Professional	.53	−.03	.84
Supportive	.55	−.00	.84
Warm	.67	.19	.89
GLOBAL variable	.76	.36	.92

Note: $N = 13$ teachers.

able/likable, not optimistic/optimistic, not professional/professional, not supportive/supportive, and not warm/warm. Each segment was played once with the audio turned down completely, and the judges were given enough time to complete their ratings before the next segment was shown.

The reliabilities of the judges' ratings of the individual molar nonverbal behaviors were computed by means of intraclass correlations. The effective reliabilities of the mean of nine judges' ratings ranged from .60 to .89, indicating that variation between teachers is much larger than the variance of judgments for a given teacher (see sidebar on reliability).

The mean of the nine judges' ratings of each of the 15 molar dimensions was computed for each of the 39 clips. These means were intercorrelated in a 15 × 15 correlation matrix, which was subjected to a principal components analysis. Principal-components analysis is often used with high (in this case 15)—dimensional data to identify composite variables. In this case a single composite variable was identified as a sort of "global" measure. This variable was created by computing the mean of the 15 variables (The variable of anxiety was reverse scored as not anxious because it was negatively related to all the other molar variables).

On the whole, teachers with higher CUE ratings were judged more favorably on the average of 15 dimensions of nonverbal behavior. The correlation coefficient was .76, which is significantly different from 0. An interesting statistical methodology point is that the effective sample size for assessing significance is the number of teachers (13) and not the number of clips (13 × 3 = 39). Table 1 reports the correlations of the individual variables. All are positively and substantially correlated with teacher ratings. Specifically, teachers who were rated higher by their students were judged to be significantly more optimistic, confident, dominant, active, enthusiastic, likable, warm, competent, and supportive on the basis of their nonverbal behavior.

The overall result of this study was quite remarkable: Based on observations of video clips just half a minute in length, complete strangers were able to predict quite accurately the ratings of teachers by students who had interacted with them over the course of a whole semester! Quite surprised by these results, we decided to replicate the study with a different sample of teachers and a slightly different criterion variable.

Table 2—Correlations of Molar Behaviors with High School Teachers' Effectiveness Ratings (Principal's Rating)

Variable	*r*	95% confidence interval	
		From	To
Accepting	.64	.14	.88
Active	.41	−.18	.78
Attentive	.64	.14	.88
Competent	.47	−.11	.81
Confident	.35	−.25	.76
Dominant	.27	−.33	.71
Empathic	.53	−.02	.84
Enthusiastic	.62	.10	.87
Honest	.42	−.17	.79
Likable	.64	.14	.88
Optimistic	.58	.04	.86
Professional	.35	−.25	.76
Supportive	.74	.32	.92
Warm	.63	.12	.88
GLOBAL Variable	.68	.21	.90
(Not) anxious	−.12	.63	.46

The variable *not anxious* was dropped from the global variable because of low reliability. Note: $N = 13$ teachers.

Study 2: Does This Effect Replicate?

For this study we employed a sample of high school teachers. We used essentially the same methodology as for the previous study. Thirteen high school teachers (eight female, five male) were videotaped in the classroom at a private high school in a large urban city. Each class lasted 50 minutes. Teachers were told about the study, and they volunteered to be videotaped. The criterion measure used to assess the effectiveness of the teachers was the evaluation by the principal of the high school on a five-point scale in response to the request "Rate the overall effectiveness of P as a teacher." All of the teachers received ratings of 4 or 5, however, possibly because the school was very selective and did not renew the contracts of teachers thought to be unsatisfactory. A master videotape of "thin slices" was derived from the 13 videotapes using the same method as in the previous study. The final stimulus tape was composed of 39 clips, 3 clips for each of 13 teachers. Eight female undergraduates were paid to rate the 39 video clips of the teachers individually on the molar dimensions described in Study 1. Once again, the judge's ratings were found to be reliable (except for ratings about instructor anxiety), and a principal-components analysis suggested the creation of a composite variable (the mean of the ratings on all the 14 variables excluding anxiety).

On the whole, teachers rated more highly by the principal were judged more favorably on the composite, $r = .68$, which is significantly different from 0. Specifically, teachers who were rated higher were judged to be significantly more supportive, likable, accepting, attentive, enthusiastic, warm, and optimistic on the basis of their nonverbal behavior. These results are shown in Table 2. Thus, the results of this study

supported the results of the first study. Judgments regarding teacher effectiveness can accurately be made from thin slices of behavior.

Other Factors and Explanations

To determine whether specific behaviors of teachers were associated with more favorable evaluations, we had two undergraduates code the frequencies per clip of specific molecular nonverbal behaviors. For each clip from both Study 1 and Study 2, the raters counted the number of head nods; head shakes; smiles; laughs; yawns; frowns; biting of the lips; downward gazes; self touches; fidgeting by shaking hands, legs, or by fiddling with an object; emphatic gestures (like pointing, clapping etc.); weak gestures (using hands while talking); the position of the hands (symmetrical or asymmetrical, folded or open); position of legs (crossed or open); and position of the torso (leaning forward or backward). They also coded whether the teacher was sitting or standing. The results were inconsistent across both studies. In the first study, teachers who fidgeted more with their hands or fiddled with an object (like chalk or a pen) received significantly lower ratings from their students. In the second study, none of the behaviors examined reached conventional levels of statistical significance. Thus, it appears that in contrast to the more global, impressionistic variables employed in the previous studies, specific behaviors are not as accurate predictors of teaching effectiveness.

Social psychologists have found that physical attractiveness is associated with a ubiquitous halo effect: Physically attractive people tend to be rated more positively on several different social and interpersonal variables. It is possible that both the students and the raters might have been influenced by teachers' physical attractiveness and thus more attractive teachers received higher student ratings.

In Study 1, the correlation between ratings of teachers' physical attractiveness and the criterion variable was $r = .32$, suggesting that students' ratings of teachers may have been affected, although not significantly so, by the physical appearance of the teachers.

Interestingly, in Study 2, the correlation between ratings of teachers' physical attractiveness and the criterion variable was $r = -.18$, suggesting that the principal was probably not positively influenced by the physical appearance of teachers in making her ratings. Any correlation obtained between the judges' ratings and the criterion variable in this study could, therefore, be attributed to factors other than the appearance of the teachers. These results regarding the negligible average (i.e., $r = .07$) relationship of physical attractiveness with teacher evaluations suggest that the famous social psychologist Gordon Allport was right when he wrote that "teaching can proceed quite successfully no matter how unfavored the teacher is by nature" (Allport, 1953, p. 857).

Minimal Information: How Low Can You Go?

The two studies so far suggest that short (10-second) video clips can be used to accurately predict longer-term or professional judgments. We also investigated whether thin-slice ratings would still predict the criterion variables if we "thinned" the slices even more. We reduced the original 10-second clips from the studies to 5-second and 2-second-long clips by randomly selecting portions from the longer clips. Thus, four new videotapes were created—2 for each of the two sets of teachers, one with three 5-second clips for each teacher and the other with three 2-second clips for each teacher. Thirty-two female undergraduates were paid to rate the four videotapes; eight different judges rated each tape and no judge saw more than one tape.

As would be expected, the mean reliabilities of the judge's ratings for the two-second clips were slightly lower than the reliabilities of the five-second clips. The means of the ratings for each dimension were summed (as in the previous studies) to create a composite variable. The correlation of the composite based on the 5-second clips with the criterion variable was slightly higher than the correlation of the composite based on the 2-second clips for the high school teachers. For the college sample, however, there was an unexpected reversal—the composite based on the 2-second clips correlated unexpectedly highly with the criterion variable. The 5-second and 2-second clips did not correlate as highly with the criterion variable as did the 10-second clips. As will be indicated, however, the effect sizes for each set of clips and sample of teachers did not differ significantly.

To summarize the results of all the studies we combined the effect sizes (i.e., correlations) for ratings on the "global" variable (composed of the individual dimensions) in Table 3. The mean overall effect size from all six substudies was $r = .59$, which is extremely significant. There were no significant differences in the accuracy of judgments based on video clips 10 second, 5-second, and 2-second in length. In addition, there were no significant differences between the accuracy of judgments for the high school and college teachers.

Our results were striking. First, we found that the ratings of complete strangers based on very thin slices of teachers' nonverbal behavior (video clips from 2-seconds to 10-seconds long) predicted with surprising accura-

Table 3—Combined Effect Sizes for Clip Length and Teacher Samples

	High school	College	Mean	95% confidence interval for mean r	
Clip length	r with crit.	r with crit.	r	From	To
10-s	.68	.76	.72	.44	.87
5-s	.47	.44	.46	.06	.73
2-s	.31	.71	.54	.16	.78
Mean r	.50	.66	.59	.23	.81

cy the ratings of the same teachers by people who had substantial interactions with those teachers, such as their students or their supervisors! These findings demonstrate the wealth of information conveyed in thin slices of behavior. Second, our results demonstrated the value of using ecologically valid criteria and behavior in actual everyday situations in assessing personality and affective variables.

These results have important implications for education. They highlight the considerable potential influence of very subtle affective nonverbal behaviors on the teaching process and reveal that these subtle influences are identifiable from thin slices of behavior. These results also suggest that potentially good teachers can be accurately identified from ratings of their molar nonverbal behavior. It should be noted, however, that as yet there is no unequivocal evidence about how much training in nonverbal skills might improve teaching effectiveness.

Using Meta-Analysis to Combine Information From Other Domains

To permit clearer and more general conclusions regarding the accuracy of predictions based on thin slices of behavior, a meta-analysis was conducted of studies using such slices to predict various social and clinical outcomes. A meta-analysis refers essentially to an "analysis of analyses," a means of summarizing the literature concerned with a particular scientific hypothesis. In this meta-analysis we examined studies that had (a) experimentally or objectively defined behavioral criteria, or (b) criteria that were ecologically valid and commonly used in everyday decisions about people. An example of an experimentally-defined criterion would be whether a subject was actually lying in a deception-detection experiment. An ecologically valid criterion would be the use of supervisors' ratings to evaluate therapeutic effectiveness because it is one of the primary methods of making decisions about performance and promotion. For the same reason, ratings by students would be a satisfactory criterion to evaluate college teacher effectiveness.

In the typical paradigm for research using such criteria, raters are asked to rate short samples of targets' behavior on various affective or personality dimensions. If the ratings are satisfactorily reliable, they can be used to post-dict, predict, or paridict the criterion variable. For example, judges might hear short samples of therapists talking to their patients, and they might be asked to rate the therapists on a series of dimensions such as anxiety, competence, and warmth. Correlations are computed between these ratings and the criterion variable, which might be the patient's prognosis judged by someone other than the therapist, such as a supervisor or a team of caretakers, or the supervisors' ratings of the therapist. Accuracy, in this context, refers to the correspondence between the average judgments of the group of judges and the criterion variable.

Four methods were used to locate the relevant studies. First, an initial computer search of Psychological Abstracts was conducted to retrieve documents containing the terms *accuracy, deception*, and *nonverbal behavior*. Because our main criterion for inclusion of studies in this analysis was a methodological one, however, this method did not prove to be very useful. The second method was a manual search of volumes of clinical and social psychology journals covering a time span ranging from 1970 to 1990. Journals were selected on the basis of relevance to social and clinical psychology and citation in other relevant articles and books. Several journals were manually searched for articles. Third, publications (especially those published before 1970) were located by searching reference lists of relevant articles and books. Last, our own files provided preprints and unpublished manuscripts pertinent to this investigation. The last three methods yielded nearly 100 studies. To be included in this review, a study had to meet the following criteria: (a) Ratings or judgments should have been based on no more than 300 seconds of behavioral observations of a subject. If no information was provided on the length of observations and if length could not be assessed from other information provided, the study was not included. (b) Short behavioral ratings had to be related to some clearly defined external, objective, behavioral criterion or to the criterion of ratings by experts, such as the existence of deception or supervisor ratings of effectiveness. (c) The study had to contain enough information to permit estimation of the significance level and effect size. Forty-four studies met all the criteria for inclusion. While all the studies were being coded, it became clear that a few studies were correlated (because they had used the same subjects). Results from these were combined and the final sample consisted of 38 independent results drawn from 44 studies.

The variables regarding subjects and raters or judges coded for each study were: (a) number of subjects overall; (b) proportion of female subjects; (c) whether the subjects were college students, noncollege students, or children; (d) whether behavior of the subjects was naturally occurring or experimentally manipulated; (e) the number of raters (or judges); and (f) the proportion of female raters.

General information about the study recorded included: (a) whether it was a field or a laboratory study (this category overlapped considerably but not completely with the previously mentioned category regarding naturally occurring or manipulated behavior), (b) whether the outcome variables could be defined as related to clinical psychology, social psychology, or the psychology of deception (although studies on deception fall under the general rubric of social psychology, they were considered a separate category because of previous work on deception accuracy as a separate category); (c) the year of publication of the study; and (d) the publication outlet.

Specific information coded for each study included: (a) the number of clips rated for each subject, (b) the number of behavioral clips rated by each judge, (c) the length of each clip, (d) the length of the total observation time for each subject, (e) whether nonverbal behavior alone or whether both verbal and nonverbal behaviors were rated, and (f) the behavioral channels rated (face, body, speech, tone of voice, transcripts, or combinations of the preceding).

Information about the results also included the magnitude (the effect size) and the significance level of the relationship between the criterion var-

Table 4—Stem-and-Leaf Plot of Mean Effect Sizes (*r*) and Statistical Summary of 38 Results (45 studies)

Stem	Leaf
1.0	
.9	
.8	7
.7	3, 4
.6	3, 8
.5	0, 2, 2, 3, 4, 4
.4	0, 0, 0, 1, 7
.3	1, 3, 5
.2	1, 1, 1, 2, 3, 3, 4, 5, 6, 6, 7, 8, 9
.1	0, 0, 4, 5, 6, 6
0	

NOTE: 95% Confidence intervals (*r*) for the combined correlation ($k = 38$) from .08 to .63

able and the behavioral dimensions rated. For each study, only one effect size and one level of significance were recorded to meet the requirement of independence of effect sizes and significance levels. Also coded were (a) the direction of the effect, and (b) the effect sizes and significance levels separately for each behavioral channel, if they were reported separately in the studies. The meta-analytic procedures used to analyze these data are those described by Rosenthal (1991).

A stem-and-leaf display of the effect sizes of the studies in the meta-analysis is presented in Table 4.

All the results show a positive effect size for accuracy in predictions, where 50% would be expected under the null. The mean effect size for accuracy of judgment from all segments of behavior under 300 seconds was .39. This significant and substantial effect size indicates that people can predict outcomes quite accurately from very small segments of behavior. Interpreting meta-analytic results can be complicated because results that fail to reach statistical significance are less likely to have been published. A "file drawer analysis" computes the number of nonsignificant studies sitting in researchers' files that would be required to make this significant result become nonsignificant. There would have to be 7,110 studies with zero effect size languishing in file drawers to overturn the significant effect of these 44 studies.

Studies were classified and compared along various categories of interest. These categories and their respective results are displayed in Table 5 and are now summarized briefly.

A central issue in this review was whether increasing the length of the audio or video clips rated would increase the accuracy in predicting criterion variables. One of our most important findings in terms of its methodological implications was that effect sizes for studies with varying length of exposures (less than 30 seconds, 30-60 seconds, 60-120 seconds, 120-180 seconds, 180-240 seconds, or 240-300 seconds) were very close, ranging from .24 to .45. A linear contrast examining the effect of exposure length on mean effect sizes was not significant, indicating that accuracy does not noticeably increase with longer exposures. Thus, longer exposures are not associated with greater accuracy.

Classification of studies by the type of outcome predicted revealed an average effect size of .41 for studies with clinical outcomes. Studies with general outcomes in the area of social psychology had an average effect size of .47, and the effect size for those with outcomes specifically related to the accuracy of detecting deception was .31. On the whole, it appears that the type of outcome predicted is not appreciably or significantly related to the accuracy of predictions.

Combined effect sizes obtained for studies in which subjects' behavior had been experimentally manipulated in the laboratory were compared with those obtained from nonexperimental studies in which subjects' behavior was allowed to vary naturally. The degree of control employed in the study was not significantly related to the accuracy of predictions. Note that most of the experimental studies were in the area of detecting deception, and most of the nonexperimental studies related to the prediction of clinical criterion variables. As far as they go, then, these results do not overwhelmingly support the proposal that observations from natural situations should be more accurate than observations from laboratory situations, although it is interesting that the effect size for field studies was higher than the effect size for laboratory studies.

For the final analysis, results were categorized according to whether only nonverbal behavior was rated, or whether verbal behavior was also included (usually speech or transcripts). Again there were no appreciable or significant differences between the two categories.

Table 5—Results of Studies Classified by Length of Exposure, Type of Outcome, Type of Study and Type of Behaviors Rated in the Study

Category	*n*	*r*
Exposure length		
0-30 seconds	7	.33
30-60 seconds	7	.44
60-120 seconds	16	.39
120-180 seconds	1	.24
180-240 seconds	3	.39
240-300 seconds	4	.45
Type of outcome		
Clinical	11	.41
Deception	16	.31
Social	11	.47
Type of study		
Field	17	.47
Laboratory	21	.32
Behaviors rated		
Nonverbal (NV)	15	.45
Verbal + NV	23	.35

Measuring Reliability

In studies like the one reported here, several judges (students or experts here) are used to rate a set of stimuli (teachers in this case). Another example would be using a set of test items to rate a set of examinees. Reliability refers to the consistency or stability of judgments or measurements made by judges or test items. There are two principal types of reliability important to psychological research. Test–retest reliability refers to the stability of judges or test items over time. It can be defined operationally by measuring a set of stimuli twice and using the correlation coefficient as a measure of reliability. A second measure of reliability is the internal consistency of a set of judges or test items. In this instance each pair of judges or tests can be summarized by a correlation coefficient. The reliability of the ensemble of judges or test items is a function of the number of judges and the mean pairwise correlation. Averaging the opinions of a greater number of judges leads to greater reliability.

Here we are primarily interested in the internal consistency reliability of our groups of judges. Rather than compute pairwise correlations for each judge, we describe how to compute reliability using the analysis of variance. We can take the rating of the ith ($i = 1, \ldots, I$) person by the jth judge ($j = 1, \ldots, J$) to be $Y_{ij} = \mu + a_i + b_j + e_{ij}$ with person effects $a_i \sim N(\sigma, \sigma^2_a)$, judge effects to account for possibly different mean ratings from each judge, and errors $e_{ij} \sim N(\sigma, \sigma^2_e)$. Then the intraclass correlation

$$\rho = \frac{\sigma^2_a}{\sigma^2_a + \sigma^2_e}$$

provides an estimate of the reliability of a single judge and can be estimated as

$$\hat{\gamma} = \frac{\mathrm{MS}_a - \mathrm{MS}_e}{\mathrm{MS}_a + (J-1)\mathrm{MS}_e}$$

The reliability of the mean of all J judge ratings would be

$$R = \frac{\dfrac{\sigma^2_a}{\sigma^2_a + \sigma^2_e}}{J} = \frac{n\hat{\gamma}}{1 + (n-1)\hat{\gamma}}$$

which can be estimated as

$$\frac{\mathrm{MS}_a - \mathrm{MS}_e}{\mathrm{MS}_a}$$

As an example consider the following from Rosenthal & Rosnow (1991, p. 56). There are three judges and five people being rated.

	Judge		
Person	A	B	C
a	5	6	7
b	3	6	4
c	3	4	6
d	2	2	3
e	1	4	4

The MS(persons) is 6.00 and the MS(Error) = .85 with r = .669 and R = .858. For this small example we can compute pairwise correlations (r_{ab} = .645, r_{bc} = .582, r_{ac} = .800) that have a mean .676, quite close to that obtained by the analysis of variance.

Summary of Findings

To return to the question that we posed at the beginning of this article, how accurate are the judgments that we make about others from minimal information? Is the job candidate really ideal? We believe that our results suggest that our judgments can be accurate provided that these judgments are based on relevant information. We would probably have to observe the job candidate performing the task. Thus, it is not clear whether effective teachers could be identified from a casual conversation. In addition, judgments are more accurate when the judgments of several individuals are pooled. The present research employs the aggregated judgments of individuals to predict outcomes. Judgments of single individuals will probably not be quite as accurate due to skill and personality differences among individuals.

What types of behaviors or dimensions can be judged accurately by using ratings of thin slices of behavior? It would be unrealistic to suggest that brief observations can predict most clinical and social outcomes. Brief observations may be most appropriate in predicting criterion variables characterized by (a) greater observability and (b) greater affectivity. First, the behaviors or traits to be judged should be observable to permit greater reliability in ratings. Thus, traits such as extroversion and conscientiousness seem to be judged more accurately than traits such as emotional stability. Second, the dimensions or traits to be judged should include a substantial *affective* or *interpersonally* oriented component, such as those found in positive or negative teachers' expectations or patient satisfaction with doctors. Although dimensions such as anxiety, dominance, shyness, or warmth might be revealed in brief observations, less interpersonal but more personal qualities such as conscientiousness, intelligence, or persistence are probably more difficult to judge in this way. These affective, observable dimensions seem to be the ones that need to be judged quickly for survival and adaptation to the environment.

These findings have several practical implications. First, researchers can save

continued on page 51

as to whether the ability to measure something says anything about its existence. Does theory propel you toward making certain measurements, or does the theory emerge naturally from an unforced pattern of data?

As a nonspecialist I conclude that IQ is a slippery and complex concept, not at all well defined even to those working in the area. Eysenck stated blatantly that "questions of definition usually worry the scientist much less than the interested layman; the former realizes, as the latter does not, that a proper definition comes at the end, not at the beginning, of a scientific quest." Well, that would explain a lot of the despairing arguments I have had with researchers who consult a statistician only after they have collected their data and "want to know what it means."

Because IQ's are created from raw scores of tests by reference to tables or formulas, they would appear to have the status of standard normal deviates, or z-scores. It should be mathematically valid to do anything that you could do with a continuous normal deviate; you can convert scores to percentiles, perform ANOVA, and so forth. But you cannot assume that either the Normal distribution or the distribution of the raw scores describes an underlying latent variable. That is the danger.

The straight answer to my poll question is that a large majority of people probably do assume that underlying general intelligence is randomly distributed and, for a large population, the scores would follow a normal or bell-shaped curve. More sophisticated arguers would realize that the bell curve is a mathematical artifact of the tests employed, but they would at least allow that these might map an underlying rank order in performance. The idea of the quotient still lurks in the background, especially in a climate in which standard age-related tests are administered to monitor both the pupils and their teachers. Social engineers (of any persuasion) will argue that the tests are meaningless, or socially and culturally determined, or evidence of meaningful differences between groups. Specialists in education or psychology will argue that a naive intervention by a meddling nonspecialist has ignored all the recent work in refining ability measures to avoid the outdated notion of 'general intelligence.' Specialists hate being questioned.

References and Further Reading

Cronbach, L. J. (1984), *Essentials of Psychological Testing*, New York: Harper & Row.

Eysenck, H. J. (1973), *The Measurement of Intelligence* (selected readings), Lancaster, U.K.: Medical and Technical Publishing Co.

Kaplan, J. (1997), "A Statistical Error in the Bell Curve," *Chance*, 10(1), pp. 20–21.

Using "Thin Slices"

continued from page 18

time (their own and that of their raters) and money by using thin slices of behavior to evaluate important affective variables, without sacrificing accuracy. Second, ratings of thin slices of behavior can be used to predict important criterion variables, particularly those that are interpersonally oriented. For example, these ratings might be used to identify biased teachers, assess aspects of the therapeutic process, or gauge the expectancies of various targets such as newscasters. Third, ratings of thin slices of behavior might be very useful in the selecting, training, and evaluation of people who need strong interpersonal skills, such as managers, salespersons, teachers, and therapists. Our evaluative judgments of other people based on minimal perceptual information are sometimes surprisingly accurate.

Author Notes

Work on this manuscript was supported by grants from the Bayer Institute for Health Care Communication and the National Science Foundation.

Correspondence regarding this article should be sent to Nalini Ambady, Department of Psychology, 1102 William James Hall, Harvard University, Cambridge, MA 02138 U.S.A. E-mail can be sent to na@wjh.harvard.edu.

References and Further Reading

Allport, G. W. (1953), "The Teaching-Learning Situation," Public Health Reports, p. 857.

Ambady, N., and Rosenthal, R. (1992), "Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis," *Psychological Bulletin*, 111, 256–274.

——— (1993), "Half a Minute: Predicting Teacher Evaluations From Thin Slices of Behavior and Physical Attractiveness," *Journal of Personality and Social Psychology*, 64, 431–441.

Feldman, K. A. (1989a), "Instructional Effectiveness of College Teachers as Judged by Teachers Themselves, Current and Former Students, Colleagues, Administrators, and External (Neutral Observers)," *Research in Higher Education*, 30, 137–194.

——— (1989b), "The Association Between Student Ratings of Specific Instructional Dimensions and Student Achievement: Refining and Extending the Synthesis of Data from Multisection Validity Studies," *Research in Higher Education*, 30, 583–645.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982), *Judgment Under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press.

Rosenthal, R. (1987), *Judgment Studies: Design, Analysis and Meta-analysis*, Cambridge: Cambridge University Press.

——— (1991), *Meta-Analytic Procedures for Social Research*, Beverly Hills, CA: Sage.

Rosenthal, R., and Rosnow, R. (1991), *Essentials of Behavioral Research*, New York: McGraw-Hill.