CHAPTER 5

CONDUCTING JUDGMENT STUDIES: SOME METHODOLOGICAL ISSUES

ROBERT ROSENTHAL

Although the focus of this chapter is on nonverbal behavior in the affective sciences, the basic principles described apply to any context in which judgment studies are conducted. The term 'judgment studies' refers most generally to those studies in which behaviors, persons, objects, or concepts are evaluated by one or more judges, raters, coders, or categorizers, referred to collectively as 'judges'. These judges may be general experts, specialist-experts, members of the general public, college students and the like; the judgments they are asked to make run the gamut from the degree of warmth shown by a psychotherapist to the affect or creativity shown in a work of art. In this chapter we consider some of the fundamental methodological issues that contemporary researchers will want to consider when they conduct judgment studies including issues of the nature of judgment studies, the reliability of judgments, the selection of judges, the formation of composite variables, and some related topics.

Introduction

Research in nonverbal communication very often requires the use of observers, coders, raters, decoders, or judges. Although distinctions among these classes of human (or at least animate) responders are possible, we shall not distinguish among them here but, rather, use these terms more or less interchangeably.

Judgment studies may focus on nonverbal behaviors considered as independent variables; for example, when the corners of the mouth rise, do judges rate subjects as being happier? Judgment studies may also focus on nonverbal behaviors considered as *dependent* variables; for example, when subjects are made happier, are the corners of their mouths judged as having risen more?

Judgment studies may employ a variety of metrics, from physical units of measurement to psychological units of measurement. For example, the movement of the corner of the mouth can be given in millimeters, while judges' ratings of happiness may be given on a scale (perhaps of seven points) ranging from 'not at all happy' to 'very happy'.

The judgments employed in a judgment study may vary dramatically in their reliability. Thus, judgments based on physical units of measurement are often more reliable than are judgments based on psychological units of measurement, although, for some purposes, the latter may be higher in validity despite their being lower in reliability (Rosenthal 1966). This may be due to the lower degree of social meaning

Table 5.1 Dimensions tending to distinguish various types of judgment studies

Dimensions	Examples
Type of variable	Dependent vs. independent variables
Measurement units	Physical vs. psychological units
Reliability	Lower vs. higher levels
Social meaning	Lower vs. higher levels

inherent in the more molecular physical units of measurement compared to the more molar psychological units of measurement. Table 5.1 shows some of the dimensions upon which it is possible to classify various judgment studies.

The judgment study model

The underlying model of a basic judgment study is shown in Fig. 5.1. One or more encoders characterized by one or more attributes (e.g. traits, states) (A) are observed by one or more decoders who make one or more judgments (C) about the encoders on the basis of selectively presented nonverbal behavior (B). The AB arrow refers to the relationship between the encoder's actual attribute (e.g. state) and the encoder's nonverbal behavior. The AB arrow reflects the primary interest of the investigator who wishes to employ the nonverbal behavior as the dependent variable. The BC arrow reflects the primary interest of the investigator who wishes to employ the nonverbal behavior as the independent variable. The AC arrow reflects the primary interest of the investigator interested in the relationship between the encoder's attribute and the decoders' judgment (e.g. the decoders' accuracy).

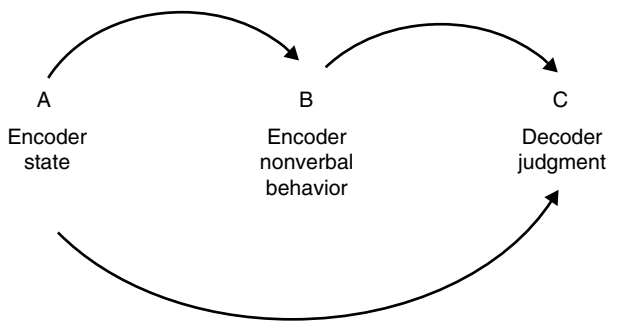

Figure 5.1 A simple model of judgment studies

The nonverbal behavior (B) presented to the decoders tends to be highly selected as part of the research design. Investigators interested in facial expressions might present still photographs of the face (e.g. Ekman 1973; Ekman *et al.* 1987), while investigators interested in tone of voice might present speech that is content-standard (Davitz 1964), randomized-spliced (Scherer 1971), or content-filtered (Rogers *et al.* 1971). Investigators interested in comparing the relative efficiency of cues carried in various channels of

nonverbal communication might provide access to different channels of nonverbal cues (e.g. face, body, tone of voice) (Rosenthal *et al.* 1979; Scherer *et al.* 1977).

To summarize the simple judgment model then, we have encoder attributes (e.g. states) (A), manifested behaviorally (B), and decoded by judges (C). The states then are antecedents of both the nonverbal behaviors and the decoders' judgments.

A more complex judgment study model based on Brunswik's (1956) lens model has been described by Scherer (1978).

The purposes of judgment studies

Judgment studies serve many purposes. In terms of our simple model of judgment studies (Fig. 5.1), the focus of a judgment study may be on the encoder state or other attribute (A), the encoder's nonverbal behavior (B), the decoder's judgment itself (C), the AB, AC, and BC arrows, or the ABC chain.

Encoder state

Suppose we wanted to develop a system for the diagnosis of anxiety in college students from various nonverbal cues (e.g. Harrigan *et al.* 1996). Suppose further that we had available film clips of 30 students being interviewed. Before we could correlate various nonverbal behaviors with the degree of anxiety of the students we would have to ascertain their 'actual' anxiety level. One way of defining this might be to show the 30 film clips to a sample of experienced clinical psychologists or other experts on anxiety, and obtain ratings of the degree of anxiety shown by each college student.[1] The mean rating of anxiety of each stimulus person (encoder) becomes the operational definition of the true state of the encoder. Note that our emphasis here is on defining the encoder state, *not* on specifying the cues that might have led the expert judges to decide on what ratings they would give. In addition, note that this particular judgment study, done for the purpose of estimating parameters (mean anxiety) rather than establishing relationships, was a kind of preliminary study to be followed up by a study linking the state of anxiety to the nonverbal concomitants (an AB arrow) (Rosenthal and Rosnow 1975*a*, 1991).

Encoder nonverbal behavior

Suppose we wanted to study the mediation of teacher expectancy effects (Rosenthal 1966, 1969, 1974, 1976, 1985, 2002*a*,*b*, 2003; Rosenthal & Jacobson 1968; Rosenthal & Rubin 1978). One of our hypotheses might be that teachers who expect more from their students treat them more warmly. Furthermore, we may believe that this warmth will be expressed, in part, through tone of voice. Before we can examine the relationship between teachers' expectations and teachers' warmth in tone of voice, however, we must be able to define tonal 'warmth'. One way of defining warmth would be to ask

[1] An alternative way of defining 'actual' level of anxiety in terms of test scores is described in a subsequent paragraph on *AC arrows*.

judges to make ratings of the degree of warmth shown in the content-filtered voices of teachers talking to their students. The mean rating of warmth obtained for each stimulus teacher's content-filtered voice becomes the definition of the warmth of the nonverbal behavior.

This particular judgment study, like the one described just above, was conducted for the purpose of estimating parameters (mean warmth) rather than establishing relationships. As such, it might serve as a kind of preliminary study that would be followed up by a study relating the nonverbal behavior to a teacher state or some other type of variable. Such studies have been conducted focusing on the tone of voice shown by, for example, teachers, psychotherapists, counselors, physicians, and mothers (Ambady *et al.* 2002; Ambady & Rosenthal 1992, 1993; Babad 1992; Blanck & Rosenthal 1984; Blanck *et al.* 1990; Eden 1990; Halverson *et al.* 1997; Harris & Rosenthal 1985, 1986; Milmoe *et al.* 1967, 1968; Rosenthal *et al.* 1984). (Though a number of sources have just been listed, it should be noted that they are listed only as illustrations, with no attempt to provide a review of any aspect of the literature of nonverbal communication.)

Decoder judgment

In the case of the two purposes of judgment studies described so far, judges' ratings were employed to provide the definitions of encoder states and encoder nonverbal behavior, usually in the context of a preliminary study or a simple descriptive study; for example, what proportion of experimenters smile at their research subjects? (Rosenthal 1967) Sometimes, however, it is the judgments themselves we want to study. The interpretation of nonverbal cues may depend heavily on personal characteristics of the judges. Thus, we might not be surprised to find that aggressive, delinquent boys will tend to interpret nonverbal cues as more aggressive than would less aggressive boys (Nasby *et al.* 1980). Or we might be interested to learn that blind children may be more sensitive to tone of voice cues (content-filtered and randomized-spliced) than are sighted children (Rosenthal *et al.* 1979). One of the earliest uses of decoders' judgments was to help establish that nonverbal behavior could, in fact, be decoded accurately (Allport 1924; Ekman 1965, 1973).

AB arrows

If we record therapists' expectations for their patients and observe their nonverbal behaviors as they interact with their patients, we have the ingredients of an AB arrow study. Therapists' nonverbal behaviors could be defined in terms of muscle movements in millimeters, or voice changes in hertz, or in terms of warmth, pride, esteem, and expectation, as rated on nine-point scales. In any case, we regard the nonverbal behaviors as the dependent variable, and the therapists' expectations as the independent variable (Blanck *et al.* 1986).

BC arrows

A common type of BC arrow judgment study might experimentally manipulate various encoder nonverbal cues and observe the effects on decoders' ratings of various encoder

characteristics (e.g. Friedman 1976, 1978, 1979*a*; Harrigan & Rosenthal 1983). Questions addressed might include:

- Are smiling faces rated as more friendly?
- Are voices with greater pitch range judged more pleasant?
- Are louder voices judged more extraverted? (Scherer 1970, 1978, 1979*a*,*b*, 1982; Scherer *et al.* 1972; Scherer & Oshinsky 1977)

AC arrows

AC arrow judgment studies are common in the general research domains of clinical diagnosis and person perception. The general paradigm is to ask decoders to assess the encoders' true attributes (e.g. diagnosis, anxiety level, adjustment level) and to correlate decoders' judgments with independently determined definitions of encoders' true traits or states. Thus, for example, clinicians' ratings of adjustment and anxiety might be correlated with encoders' scores on various subscales of such tests as the MMPI, or scores based on the Rorschach, TAT, or life history data.

When AC arrow judgment studies are employed in research on nonverbal communication it is often in the context of 'accuracy' studies. Encoders might, for example, show a variety of posed or spontaneous affects, and judgments of these affects made by decoders are evaluated for accuracy. Sometimes these accuracy studies are conducted to learn the degree to which judges show better than chance accuracy (Allport 1924; Ekman 1965, 1973). At other times, these accuracy studies are conducted to establish individual differences among the judges in degree of accuracy shown. These individual differences in decoding accuracy may then be correlated with a variety of personal attributes of the judges (e.g. gender, age, ethnicity, psychopathology, cognitive attributes, personality attributes) (Rosenthal *et al.* 1979). It should be noted that such individual difference studies can be meaningfully conducted even when the mean level of accuracy shown by the entire sample of judges does not exceed the chance expectation level as in comparisons of people scoring above chance with those scoring below chance.

ABC chains

Sometimes we are simultaneously interested in the AB arrow and the BC arrow; such studies can be viewed as studies of the ABC chain. Suppose we want to study the mediation of teacher expectancy effects. We begin with a sample of teachers known to vary in their experimentally created expectations for their pupils' intellectual performance (i.e. known to vary in encoder states (A)). These teachers are observed interacting with pupils for whom they hold higher or lower expectations and a sample of judges rates the teachers' behavior on degree of smiling, forward lean, and eye contact (i.e. encoder nonverbal behaviors (B)). Finally, a sample of judges rates the nonverbal behavior of the teachers for degree of overall warmth and favorableness of expectancy (i.e. makes decoder judgments (C) of a fairly molar type). We would now be in a position to examine the effects of experimentally induced teacher expectation on teacher nonverbal behavior and the role of these nonverbal behaviors in

predicting outcomes of social consequence, all in the same study (Harris & Rosenthal 1985, 1986).

Designing judgment studies

The particular purpose of any judgment study should determine the particular procedures of any judgment study. Given the diversity of purposes of judgment studies we have discussed above, it is not possible to prescribe the detailed procedures that should be employed for any particular judgment study. However, because judgment studies do have certain communalities, it is possible to discuss methodological issues likely to be confronted in many judgment studies. In the following pages we address three of these issues in some detail:

1. the reliabilities of judgments made;
2. the selection of judges;
3. the combining of judgments to form composite variables.

Issues of reliability

How many judges shall we employ in a judgment study in which our primary interest is in the encoders rather than the judges, and who should they be? The major factors determining the answers to these questions are:

1. the average reliability coefficient (r) between pairs of judges chosen at random from a specified population;
2. the nature of the population of judges to which we want to generalize our results.

Effective reliability

Suppose our goal were to establish the definition of the encoder's state (A) or of some encoder nonverbal behavior (B). We might decide to employ judges' ratings for our definition. As we shall see shortly, if the reliability coefficient (any product moment correlation such as r, point biserial r, or phi) were very low, we would require more judges than if the reliability coefficient were very high. Just how many judges to employ is a question for which some useful guidelines can be presented (Rosenthal 1973, 1982, 1987; Li *et al.* 1996).

If we had a sample of teachers whose nonverbal warmth we wanted to establish, we might begin by having two judges rate each teacher's warmth based on the videotaped behavior of each teacher. The correlation coefficient reflecting the reliability of the two judges' ratings would be computed to give us our best (and only) estimate of the correlation likely to be obtained between any two judges drawn from the same population of judges. This correlation coefficient, then, is clearly useful; it is not, however, a very good estimate of the reliability of our variable, which is not the rating of warmth made by a single judge but rather the mean of two judges' ratings. Suppose, for example, that the correlation between our two judges' ratings of warmth were 0.50; the reliability of the mean of the two judges' ratings (the 'effective' reliability) would

then be 0.67, not 0.50. Intuition suggests that we should gain in reliability in adding the ratings of a second judge because the second judge's random errors should tend to cancel the first judge's random errors. Intuition suggests further that adding more judges, all of whom agree with one another to about the same degree, defined by a mean inter-judge correlation coefficient of 0.50 (for this example), should further increase our 'effective' reliability. Our intuition would be supported by a very old and well-known result reported independently and simultaneously by Charles Spearman (1910) and William Brown (1910). With notation altered to suit our current purpose, the well-known Spearman–Brown equation is:

$$R_{SB} = \frac{nr}{1 + (n-1)r} \tag{1}$$

where R_{SB} = 'effective' reliability; n = number of judges; r = mean reliability among all n judges (that is, mean of $\frac{n(n-1)}{2}$ correlations).

Use of this formula depends on two assumptions:

1. a comparable group of judges would show comparable 'mean' reliability among themselves and with the actual group of judges available to us;
2. that all judges have essentially the same variance of their ratings of the same sample.

It should be noted that the 'effective' reliability also can be obtained computationally by means of the Kuder–Richardson '20 equation' or by means of Cronbach's coefficient alpha (Guilford 1954).

When the assumptions underlying the use of the Spearman–Brown equation are not met, as when we can think of two or more different subgroups of judges, adjustments to the equation are available and are described in detail in Li *et al.* (1996).

As an aid to investigators employing these and related methods, Table 5.2 has been prepared employing the Spearman–Brown equation. The table gives the effective reliability, R_{SB}, for each of several values of n (the number of judges making the observations) and r (the mean reliability among the judges). It provides quick, approximate answers to each of the following questions:

1. Given an obtained or estimated mean reliability, r, and a sample of n judges, what is the approximate effective reliability, R_{SB}, of the mean of the judges' ratings? The value of R_{SB} is read from the table at the intersection of the appropriate row (n) and column (r).
2. Given the value of the obtained or desired effective reliability, R_{SB}, and the number, n, of judges available, what will be the approximate value of the required mean reliability, r? The table is entered in the row corresponding to the n of judges available, and is read across until the value of R_{SB} closest to the one desired is reached; the value of r is then read as the corresponding column heading.
3. Given an obtained or estimated mean reliability, r, and the obtained or desired effective reliability, R_{SB}, what is the approximate number (n) of judges required? The table is entered in the column corresponding to the mean reliability, r, and is read down until the value of R_{SB} closest to the one desired is reached; the value of n is then read as the corresponding row title.

Product moment correlations

It should be noted that the mean reliability (r) of Table 5.2 is to be a product moment correlation coefficient such as Pearson's r, the point biserial r, or the phi coefficient. It is often not appropriate to employ such indices of 'reliability' as percentage agreement or multidegree of freedom indices of interjudge agreement.

Some risks in not using Pearson's *r*-based indices of reliability

Percentage agreement

It has long been common practice for some researchers to index the reliability of judges' categorizations using percentage agreement defined as:

$$\left(\frac{A}{A+D}\right)100 \tag{2}$$

where A represents the number of agreements and D represents the number of disagreements (Rosenthal and Rosnow 1991).

Table 5.2 Effective reliability (R_{SB}) of the mean of judges' ratings

	Mean reliability (r)																				
No. of judges (n)	.01	.03	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
1	01	03	05	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
2	02	06	10	18	26	33	40	46	52	57	62	67	71	75	79	82	86	89	92	95	97
3	03	08	14	25	35	43	50	56	62	67	71	75	79	82	85	88	90	92	94	96	98
4	04	11	17	31	41	50	57	63	68	73	77	80	83	86	88	90	92	94	96	97	*
5	05	13	21	36	47	56	62	68	73	77	80	83	86	88	90	92	94	95	97	98	*
6	06	16	24	40	51	60	67	72	76	80	83	86	88	90	92	93	95	96	97	98	*
7	07	18	27	44	55	64	70	75	79	82	85	88	90	91	93	94	95	97	98	98	*
8	07	20	30	47	59	67	73	77	81	84	87	89	91	92	94	95	96	97	98	*	*
9	08	22	32	50	61	69	75	79	83	86	88	90	92	93	94	95	96	97	98	*	*
10	09	24	34	53	64	71	77	81	84	87	89	91	92	94	95	96	97	98	98	*	*
12	11	27	39	57	68	75	80	84	87	89	91	92	94	95	96	97	97	98	*	*	**
14	12	30	42	61	71	78	82	86	88	90	92	93	94	95	96	97	98	98	*	*	**
16	14	33	46	64	74	80	84	87	90	91	93	94	95	96	97	97	98	98	*	*	**
18	15	36	49	67	76	82	86	89	91	92	94	95	96	96	97	98	98	*	*	*	**
20	17	38	51	69	78	83	87	90	92	93	94	95	96	97	97	98	98	*	*	*	**
24	20	43	56	73	81	86	89	91	93	94	95	96	97	97	98	98	*	*	*	**	**
28	22	46	60	76	83	88	90	92	94	95	96	97	97	98	98	98	*	*	*	**	**
32	24	50	63	78	85	89	91	93	95	96	96	97	98	98	98	*	*	*	*	**	**
36	27	53	65	80	86	90	92	94	95	96	97	97	98	98	*	*	*	*	**	**	**
40	29	55	68	82	88	91	93	94	96	96	97	98	98	98	*	*	*	*	**	**	**
50	34	61	72	85	90	93	94	96	96	97	98	98	98	*	*	*	*	**	**	**	**
60	38	65	76	87	91	94	95	96	97	98	98	98	*	*	*	*	*	**	**	**	**
80	45	71	81	90	93	95	96	97	98	98	98	*	*	*	*	*	**	**	**	**	**
100	50	76	84	92	95	96	97	98	98	*	*	*	*	*	*	**	**	**	**	**	**

Note: Decimal points omitted
* Approximately 0.99
**Approximately 1.00

Table 5.3 shows how percentage agreement can be a very misleading indicator of interjudge reliability. In Part A of Table 5.3 we find that two researchers, Smith and Jones, each had two judges evaluate a series of 100 film clips of children for the presence or absence of frowning behavior. Both Smith and Jones found their judges to show 98% agreement, but Smith's 98% agreement was a hollow victory indeed. The correlation between judges A and B was actually slightly negative, $r = -.01$, ($\chi^2_{(1)} = 0.01$). Jones' 98% agreement, on the other hand, was associated with an r of + .96, ($\chi^2_{(1)} = 92.16$).

Part B of Table 5.3 shows two additional cases of percentage agreement obtained by researchers North and West. This time, the two investigators have both obtained an apparently chance level of agreement (i.e. 50%). Both results, however, are very far from reflecting chance agreement, both with $p = 0.0009$. Most surprising, perhaps, is that North obtained a substantial negative reliability ($r = -0.33$) while West obtained a substantial positive reliability ($r = +0.33$); another illustration that percentage agreement is not a very informative index of reliability.

Multi-*df* interjudge reliability

Among the first psychologists to appreciate the problems of percentage agreement as an index of reliability was Jacob Cohen (1960, 1968). He developed an index, kappa, that solved the problem of the percentage agreement index by adjusting for any agreement based simply on lack of variability (e.g. the lack of variability found in Part A of Table 5.3 where both of Smith's judges found 98% of the film clips to show frowning behavior).

Table 5.3 Examples of percentage agreement

A. Two cases of 98% agreement

Smith's results			**Jones's results**		
	Judge A			Judge C	
Judge B	Frown	No frown	Judge D	Frown	No frown
Frown	98	1	Frown	49	1
No frown	1	0	No frown	1	49
Agreement = 98% but $r_{AB} = -.01$; $\chi^2_{(1)} = 0.01$			Agreement = 98% but $r_{CD} = +.96$; $\chi^2_{(1)} = 92.16$		

B. Two cases of 50% agreement

North's results			**West's results**		
	Judge E			Judge G	
Judge F	Frown	No frown	Judge H	Frown	No frown
Frown	50	25	Frown	25	50
No frown	25	0	No frown	0	25
Agreement = 50% but $r_{EF} = -.33$; $\chi^2_{(1)} = 11.11$			Agreement = 50%, but $r_{GH} = +.33$; $\chi^2_{(1)} = 11.11$		

Table 5.4 Results of two diagnosticians' classification of 100 persons into one of four categories

		Judge 1				
		A Schizophrenic	B Neurotic	C Normal	D Brain-damaged	Σ
Judge 2	A Schizophrenic	13	0	0	12	25
	B Neurotic	0	12	13	0	25
	C Normal	0	13	12	0	25
	D Brain-damaged	12	0	0	13	25
	Σ	25	25	25	25	100

$kappa(df = 9) = \frac{O-E}{N-E} = \frac{50-25}{100-25} = .333$

Table 5.4 gives an example of the type of situation in which kappa is often employed. Two clinical diagnosticians have examined the silent videotapes of 100 people in a clinical interview where the interviewer is not shown. Each clinician was asked to assign each interviewee to one of four classifications: schizophrenic, neurotic, normal, and brain damaged. Only three quantities are required to compute kappa:

1. $O =$ observed number on which the two judges have agreed (i.e. the number on the diagonal of agreement). In this example, the observed number is: $13 + 12 + 12 + 13 = 50$.
2. $E =$ expected number under the hypothesis of only chance agreement for the cells on the diagonal of agreement. For each cell, the expected number is the product of the row total and the column total divided by the total number of cases. In this example, the expected number is: $(25 \times 25)/100 + (25 \times 25)/100 + (25 \times 25)/100 + (25 \times 25)/100 = 6.25 + 6.25 + 6.25 + 6.25 = 25$.
3. $N =$ total number of cases classified. In this example, $N = 100$.

Kappa is computed from:

$$kappa = \frac{O - E}{N - E} = \frac{50 - 25}{100 - 25} = .333 \tag{3}$$

in the present example.

Although kappa is clearly an improvement over percentage agreement as an index of reliability, it does raise some serious questions. When kappa is based on tables larger than a 2×2 (e.g. a 3×3, a 4×4 (as in Table 5.4), or larger), as it often is, it suffers from the same problem as does any statistic on $df > 1$. That problem—of diffuse or omnibus procedures—is that for most values of kappa we cannot tell which focused or specific judgments are made reliably and which are made unreliably. Only when kappa approaches unity is the actual interpretation of a value of kappa straightforward (i.e. essentially all judgments are made reliably) (Rosenthal 1991). We illustrate the difficulty in interpreting kappa by returning to Table 5.4.

The 4×4 table we see, based on 9 *df,* can be decomposed into a series of six pairwise 2×2 tables, each based on a single *df,* and addressing a very specific, conceptually clear question of the reliability of dichotomous judgments—A vs. B, A vs. C, A vs. D, B vs. C, B vs. D, and C vs. D. Table 5.5 shows the results of computing kappa separately for each of these six 2×2 tables.

Table 5.5 Breakdown of the 9 *df* omnibus table of counts of Table 5.4 into six specific (focused) reliabilities of $df = 1$ each

	A Schiz	B Neurotic	Σ		A Schiz	C Normal	Σ
A Schiz	13	0	13	**A Schiz**	13	0	13
B Neurotic	0	12	12	**C Normal**	0	12	12
Σ	13	12	25	**Σ**	13	12	25
	kappa = 1.00				*kappa* = 1.00		
	A Schiz	D Brain	**Σ**		**B Neurotic**	C Normal	**Σ**
A Schiz	13	12	25	**B Neurotic**	12	13	25
D Brain	12	13	25	**C Normal**	13	12	25
Σ	25	25	50	**Σ**	25	25	50
	kappa = 0.04				*kappa* = −0.04		
	B Neurotic	D Brain	**Σ**		**C Normal**	D Brain	**Σ**
B Neurotic	12	0	12	**C Normal**	12	0	12
D Brain	0	13	13	**D Brain**	0	13	13
Σ	12	13	25	**Σ**	12	13	25
	kappa = 1.00				*kappa* = 1.00		

Schiz = schizophrenic; brain = brain-damaged

Of the six focused or specific reliabilities computed, four are kappas of 1.00, and two are kappas near zero (0.04 and −0.04). The mean of the six 1 *df* kappas is 0.667, and the median is 1.00; neither value being predictable from the omnibus 9 *df* kappa value of 0.33. Tables 5.6 and 5.7 show, even more clearly, how little relation there is between the omnibus values of kappa and the associated 1 *df* kappas (i.e. the focused reliability kappas). Table 5.6 shows an omnibus 9 *df* kappa value of 0.33—exactly the same value as that shown in Table 5.4.

Table 5.7 shows the six focused reliabilities of $df = 1$ associated with the omnibus value of kappa (0.33) of Table 5.6. We see that of these six focused kappas, four are kappas of 0.00, one is a kappa of +1.00, and one is a kappa of −1.00. The mean and median-focused kappa both show a value of 0.00. We can summarize the two omnibus kappas of Tables 5.4 and 5.6, and their associated focused kappas, as follows:

	Example 1	Example 2
Omnibus kappa	0.33	0.33
Mean-focused kappa	0.67	0.00
Median-focused kappa	1.00	0.00

Thus, we have two identical kappas: one made up primarily of perfect reliabilities, the other made up primarily of zero reliabilities.

Although the greatest limitations on kappa occur when kappa is based on $df > 1$, there are some problems with kappa even when it is based on a 2×2 table of counts where $df = 1$. The basic problem under these conditions is that very often kappa is not equivalent to the

Table 5.6 Alternative results of two diagnosticians' classification of 100 persons into one of four categories

		Judge 1				
		A	**B**	**C**	**D**	**Σ**
Judge 2	**A**	25	0	0	0	25
	B	0	0	25	0	25
	C	0	25	0	0	25
	D	0	0	0	25	25
	Σ	25	25	25	25	100

$kappa(df = 9) = \frac{O-E}{N-E} = \frac{50-25}{100-25} = .333$

Table 5.7 Breakdown of the 9 *df* omnibus table of counts of Table 5.6 into six specific (focused) reliabilities of $df = 1$ each

	A	B	**Σ**		**A**	C	**Σ**
A	25	0	25	**A**	25	0	25
B	0	0	0	**C**	0	0	0
Σ	25	0	25	**Σ**	25	0	25
	$kappa = 0.00$				$kappa = 0.00$		
	A	D	**Σ**		**B**	C	**Σ**
A	25	0	25	**B**	0	25	25
D	0	25	25	**C**	25	0	25
Σ	25	25	50	**Σ**	25	25	50
	$kappa = 1.00$				$kappa = -1.00$		
	B	D	**Σ**		**C**	D	**Σ**
B	0	0	0	**C**	0	0	0
D	0	25	25	**D**	0	25	25
Σ	0	25	25	**Σ**	0	25	25
	$kappa = 0.00$				$kappa = 0.00$		

product moment correlation computed from exactly the same 2×2 table of counts. This is certainly not a criticism of kappa, since it never pretended to be a product moment correlation. The limitation, however, is that we cannot apply various interpretive procedures or displays to kappa that we can apply to product moment correlations. Examples include the use of the *coefficient of determination*, r^2, (Guilford 1954) and the *binomial effect size display* (Rosenthal & Rubin 1982; Rosenthal & Rosnow 1991)

Here, we need only indicate the conditions under which a 1 *df* kappa is or is not equivalent to a product moment correlation (referred to as a Pearson *r* in the general case and sometimes referred to as phi (or ϕ) in the case of a 2×2 table of counts). Kappa and *r* are equivalent when the row totals for levels A and B are identical to the column totals for levels A and B, respectively. Consider the following example:

		Judge 1		
		A	B	Σ
Judge 2	A	70	10	80
	B	10	10	20
	S	80	20	100

For these data, the marginal totals for level A are identical for Judges 1 and 2 (i.e. 80),

$$kappa(df = 1) = \frac{O - E}{N - E} = \frac{80 - 68}{100 - 68} = .375,$$

and r (or equivalently, phi) yields the identical value of 0.375. Therefore, we could meaningfully compute a coefficient of determination or a binomial effect size display for this particular kappa because it is equivalent to a Pearson r or phi (ϕ).

Now consider the following example in which we have the same four cell entries and the same marginal totals as in the preceding example. The only thing that has changed is the location of the cell with the largest count (70) so that the marginal totals for level A differ for Judges 1 and 2 (20 versus 80).

		Judge 1		
		A	B	Σ
Judge 2	A	10	70	80
	B	10	10	20
	S	20	80	100

In this example,

$$kappa(df = 1) = \frac{O - E}{N - E} = \frac{20 - 32}{100 - 32} = -.176,$$

but r (or ϕ) yields a markedly different value of -0.375. We can, therefore, compute a meaningful coefficient of determination or a binomial effect size display for r, but we cannot do so for kappa.

Other approaches to reliability

Reliability and analysis of variance

When there are only two judges whose reliability is to be evaluated, it is hard to beat the convenience of a product moment correlation coefficient for an appropriate index of reliability. As the number of judges grows larger, however, working with correlation coefficients can become inconvenient. For example, suppose we employed 40 judges and wanted to compute both their mean reliability (r) and their effective reliability (R_{SB}). Table 5.2 could get us R_{SB} from knowing r, but to get r we would have to compute $(40 \times 39)/2 = 780$ correlation coefficients. That is not hard work for computers, but averaging the 780 coefficients to get r can be hard work for investigators or their programmers. There is an easier way, and it involves the analysis of variance.

Table 5.8 Judges' ratings of nonverbal behavior

	Judges			
	A	**B**	**C**	**Σ**
Encoders				
1	5	6	7	18
2	3	6	4	13
3	3	4	6	13
4	2	2	3	7
5	1	4	4	9
Σ	14	22	24	60

Table 5.9 Analysis of variance of judges' ratings

Source	**SS**	**df**	**MS**
Encoders	24.0	4	6.00
Judges	11.2	2	5.60
Residual	6.8	8	0.85

Table 5.8 shows a simple example of three judges rating the nonverbal behavior of five encoders on a scale of 1 to 7, and Table 5.9 shows the analysis of variance of these data.[2] Our computations require only the use of the last column, the column of mean squares (Guilford 1954). Examination of the computational formulas given below shows how well the judges can discriminate among the sampling units (e.g. people) minus the judges' disagreement after controlling for judges' rating bias or main effects (e.g. *MS* encoders – *MS* residuals), divided by a standardizing quantity.

Our estimate of R_{anova} (the effective reliability of the sum or the mean of all of the ratings of the judges) is given by:

$$R_{anova} = \frac{MS\ encoders - MS\ residual}{MS\ encoders} \tag{4}$$

Our estimate of r (the mean reliability or the reliability of a *single* average judge) is given by:

$$r_{anova} = \frac{MS\ encoders - MS\ residual}{MS\ encoders + (n-1)\ MS\ residual} \tag{5}$$

where n is the number of judges as before (equation 5 is known as the intraclass correlation).

[2] In our own research, we typically use seven or nine-point rating scales (1–7 or 1–9) with unipolar rather than bipolar scales (i.e. one scale of 'behaves warmly' and a second scale of 'behaves coldly' rather than a single scale of 'warm–cold'.) Our unipolar scales usually run from 1 (not at all warm) to 7 or 9 (extremely warm). For details on response formats see Rosenthal 1987, Chapter 4.

For our example of Tables 5.8 and 5.9 we have:

$$R_{anova} = \frac{6.00 - 0.85}{6.00} = .858$$

and

$$r_{anova} = \frac{6.00 - 0.85}{6.00 + (3 - 1)0.85} = .669$$

In the present example, it will be easy to compare the results of the analysis of variance approach with the more cumbersome correlational approach. Thus, the correlations (r) between pairs of judges (r_{AB}, r_{BC}, and r_{AC}) are 0.645, 0.582, and 0.800 respectively, and the mean intercorrelation is 0.676, which differs by only 0.007 from the estimate (0.669) obtained by means of the analysis of variance approach.

If we were employing only the correlational approach, we would apply the Spearman–Brown equation (1) to our mean reliability of 0.676 to find R_{SB}, the effective reliability. That result is:

$$R_{SB} = \frac{(3)\ (.676)}{1 + (3 - 1)(.676)} = .862$$

which differs by only 0.004 from the estimate (0.858) obtained by means of the analysis of variance approach. In general, the differences obtained between the correlational approach and the analysis of variance approach are quite small (Guilford 1954).

It should be noted that, in our present simple example, the correlational approach was not an onerous one to employ, with only three correlations to compute. As the number of judges increases, however, we would find ourselves more and more grateful for the analysis of variance approach or for such related procedures as the Kuder–Richardson equations, Cronbach's alpha, or similar methods available in commonly used data analytic packages. Because of its widespread use in software packages, we briefly describe and illustrate the use of Cronbach's alpha (α).

Cronbach's alpha

More than half a century ago, Cronbach (1951) proposed *coefficient alpha* (α) that gives the reliability of a group of judges considered as a set. To compute Cronbach's α we need only three ingredients: n, the number of judges in the set contributing to the 'score' for each encoder rated; S^2_{judge}, the variance of the scores generated by an individual judge; and S^2_{total}, the variance of the total of scores given by the judges to each individual encoder. Cronbach's α is obtained from:

$$\alpha = \left(\frac{n}{n-1}\right)\left(\frac{S^2_{total} - \sum S^2_{judge}}{S^2_{total}}\right) \tag{6}$$

Table 5.10 shows the raw data of Table 5.8 but with the addition of the variance for each judge (S^2_{judge})) and the variance of the sum of the n judges' ratings (S^2_{total}). We find:

Table 5.10 Variances of individual judges and of the sum of judges' ratings of five encoders

	Judges				
Encoders	**A**	**B**	**C**		**Σ**
1	5	6	7		18
2	3	6	4		13
3	3	4	6		13
4	2	2	3		7
5	1	4	4		9
M	2.8	4.4	4.8		12.0
S^2_{judge}	2.2	2.8	2.7	S^2_{total}	18.0

$$\alpha = \left(\frac{3}{3-1}\right)\left(\frac{18-(2.2+2.8+2.7)}{18}\right) = .858$$

—a value that is the same as that reported earlier for R_{anova} (equation 4) and very close to the value (0.862) reported earlier for R_{SB} and obtained from equation 1. All three of these results (R_{anova}, R_{SB}, and α) tend to be quite similar and more so as the judges are more homogenous in their variances (that is, S^2_{judge}) and in their correlations with other judges.

Reliability and principal components

In situations where the ratings made by all judges have been intercorrelated, and a principal components analysis is readily available, another very efficient alternative for estimating the reliability of the total set of judges is available. Armor (1974) has developed an index, *theta*, that is based on the unrotated first principal component (where a principal component is a factor extracted from a correlation matrix employing unity (1.00) in the diagonal of the correlation matrix). The equation for theta is:

$$theta = \frac{n}{n-1}\left(\frac{L-1}{L}\right) \tag{7}$$

where n is the number of judges and L is the latent root or eigenvalue of the first unrotated principal component. The latent root is the sum of the squared factor loadings for any given factor, and can be thought of as the amount of variance in the judges' ratings accounted for by that factor. Factor analytic computer programs generally give latent roots or eigenvalues for each factor extracted, so that theta is very easy to obtain in practice. Armor (1974) has pointed out the close relationship between theta and Cronbach's coefficient alpha.

For an illustration of the use of theta we refer to the standardization of a test of sensitivity to nonverbal cues—the Profile of Nonverbal Sensitivity (PONS) (Rosenthal *et al.* 1979). When the 220 items of that test were subjected to a principal components analysis, the eigenvalue or latent root (L) of the first (unrotated) component was 13.217. Therefore, from equation 7 we find: