Interrater Agreement and Combining Ratings

Richard S. Bogartz

University of Massachusetts, Amherst

RUNNING HEAD: Interrater agreement

Address correspondence to:

Richard S. Bogartz

Department of Psychology

University of Massachusetts

Amherst, MA 01003

bogartz@psych.umass.edu

Interrater Agreement and Combining Ratings

Abstract

Some behaviors such as smiles require human raters for their measurement. A model of the rating process is explored that assumes that the probability distribution of overt rating responses depends on which of several underlying or latent responses occurred. The ideal of theoretically identical raters is considered and also departures from such identity. Methods for parameter estimation and assessing goodness of fit of the model are presented. A test of the hypothesis of identical raters is provided. Simulated data are used to explore different measures of agreement, optimal numbers of raters, how the ratings from multiple raters should be used to arrive at a final score for subsequent analysis, and the consequences of departures from the basic assumptions of identical raters and constant underlying response probabilities. The results indicate that often using two or three raters to rate all of the data, assessing the quality of their ratings by assessing their pairwise correlations, and using their average rating rather than the ratings of a single primary rater will provide improved measurement of the underlying response. Simulations show that reduced rater sensitivity to response differences can be compensated for by using more raters. Methods are considered for using the model to select raters and to gauge which aspects of the rating process need additional training. Suggestions are offered for using the model in pilot rating of pilot data for deciding on the number of raters and the number of values on the rating scale.

Interrater Agreement and` Combining Ratings

Introduction

Is the Mona Lisa smiling? Some believe she is. Others suggest she wears an expression common to those who have lost their front teeth. Still others argue her expression is typical of facial muscle contracture that develops after Bell's palsy. What might we do to decide? We could ask someone to rate smiles on various pictures and see how Mona Lisa's smile is rated relative to ratings of other smiles. But then how could we know whether to rely on the rater? Well, we could use two or three raters and see if they agree. But how should we assess their agreement? And what should we do with their ratings? Use the first rater if he or she agrees with the others? Combine the ratings somehow to obtain a composite rating scale value? Simply average the ratings? And how many raters should we use? At first it seems that if we could just ask Leonardo what he intended, we could settle the matter. But not really because he may or may not have fulfilled his intentions. But surely he is the expert we could rely on. But what qualifies Leonardo as more of an expert on what is a smile than one of our other raters?

After much perplexity we may decide that we cannot know with certainty whether Mona Lisa is smiling or not and that there is no final authority. We can obtain ratings but the truth of the situation is latent rather than directly observable. This is the same situation that researchers are in when trying to decide whether an infant smiled or looked with focused attention on a given trial. There is behavior that can be rated, but the truth of whether a smile or focused attention occurred on a given trial remains latent. Although we can reach our own personal decision in the case of the Mona Lisa, in the handling of research data we require a more systematic, repeatable procedure with respect to rating observable behavior to reach conclusions about latent responses.

This paper concerns the assessment of interrater agreement and the rating of responses when a behavior is assigned a discreet rating by more than one rater: the infant looked or did not look and was assigned a 1 or 2; what might be a continuous variable of smiling is rated at three levels, no smile, partial smile, or full smile, and the ratings used are 1, 2, or 3. It is not necessary that the number of rating values equal the number of trait or behavior levels. A simple model is considered here together with its consequences for measuring interrater agreement, assigning values to the rated responses responses, and choosing the number of raters.

In much research there is a primary observer who rates or scores all of the responses by all of the participants and a secondary observer who rates all of the responses for some of them. A measure of interrater agreement is obtained between the two raters for the participants they both rated. The use of multiple observers serves a gatekeeping function to assure that the data are based on reliable judgment. Provided that the agreement measure is high enough, the ratings by the primary observer are then used as the basic data for statistical analysis. This approach is traditional but it foregoes the advantages of using the multiple judgments provided by more than one rater as the basic data, such as the greater reliability of the mean of multiple ratings than of a rating by a single rater. This is considered below in the light of the model.

In early work the proportion of agreement between the two observers' ratings was used as the measure of agreement. Then, it was pointed out that a high proportion of agreement could be obtained even though the raters were independent of each other if they shared the same bias. An alternative measure, Cohen's kappa (Cohen, 1960), was introduced to assess the amount of agreement above and beyond what would be expected from independent raters. The use of kappa has hung on to the present despite the limitations of kappa that have been indicated (Maxwell, 1977; Carey, 1978, Grove, Andreason, N. C., McDonald-Scott, P., Keller, B., &

Shapiro, R. W., 1981; Spitznagel, 1985; Uebersax, 1987). As Uebersax (1988) points out, other statistical indices such as the random error coefficient (Janes, 1979; Maxwell, 1977; Zwick, 1988), Yule's Y (Spitznagel, 1985), and the log-odds ratio (Sprott, 1987) have been proposed but many of the criticisms of kappa apply to these also.

Around the mid to late 1980s, emphasis shifted from attempts to arrive at an omnibus index of interrater agreement to attempts at modeling the data obtained from more than one rater. Many researchers appear to be largely unaware of these innovations. Uebersax & Grove (1993) cited emphasis on modeling (Agresti, 1992); Uebersax, 1992), discussions of log-linear models (Tanner, 1985a, 1985b), extensions of log-linear models based on association models (Agresti, 1988; Becker,1989, 1990); and work based on category distinguishability and quasisymmetry (Darroch, 1986).

The modeling approach taken here, although developed independently, is similar to the treatment of the discrete case presented by Uebersax (1988) in his study of validity inferences from interrater agreement. An important difference is that in the present approach it is not assumed that raters are randomly sampled from a population of raters. In many areas of research random sampling of raters from a population of raters is virtually never done. Rather, it is considered a part of good experimental procedure that the raters in hand are trained, with achieving identical raters being the ideal. In the modeling explored here it is assumed as a working hypothesis that the raters are identical, with some attention then given to testing this hypothesis and studying the consequences of it being wrong. To begin with, the class of situations that we are interested in is where a small number of raters rate each subject's response. However some attention is given to using a single primary rater.

Some Closed Form Results

The purpose of this first inquiry was to find evidence that would suggest how large an average pairwise correlation between raters would be needed to be confident that the mean of their ratings would be correlated with the underlying latent response value at a value of .95 or more. The intent was to examine the situation for from two to 5 raters, for scales using from 2 through 10 points, with underlying distributions of latent response values ranging from 2 through 10 values, using four different shapes of the underlying latent response distributions, and examining a reasonably broad range of conditional probability distributions of the assignment of a scale value given a latent response.

Method

Statistics. If, for a group of R identical raters using a rating scale with S scale points to rate a manifest response resulting from the occurrence of one of L latent responses, we know the probability distribution for the latent responses and the conditional probability distribution of the rating scale values given the latent response, we can compute both the mean pairwise correlation between raters, r_{RR} and the correlation of the mean of the R ratings with the latent response, r_{rL}.

Let p(l) be the probability of latent response l occurring on a given trial. For each of the R raters (the identity assumption), let q(s,l) be the conditional probability that the scale value s is the rating given that latent response l occurred. Since all of the raters are assumed to be identical, to obtain the mean pairwise correlation between the raters it is sufficient to obtain the expected value of the correlation between an arbitrary pair of raters. This is

$$Cor(rr) = \{E(rr) - [E(r)]^2\} / Var(r) \qquad [1]$$

where

$$E(rr) = \sum_i \sum_j \sum_k p(l)\, jkq(j,l)$$

$$E(R) = \sum_l \sum_j p(l)\, jq(j,l)$$

and

$$Var(r) = \sum_l \sum_j p(l)\, j^2 q(j,l) - [E(r)]^2$$

.

To calculate the correlation of the mean of the <u>*R*</u> ratings with the latent response, let $\bar{r}$ be the mean of the ratings for the raters on a single trial and *l* be the latent response on that trial. Then the correlation of $\bar{r}$ with *l* is

$$Cor(\bar{r}L) = (E(\bar{r}L) - E(\bar{r})E(L)]/[SD(\bar{r})SD(L)] \qquad [2]$$

$$E(\bar{r}L) = \sum_l \sum_{r_1} \sum_{r_2} \ldots \sum_{r_S} lp(l)[(r_1 + r_2 + \ldots + r_S)/S]\Pi_s q(s,l)$$

where

$$E(\bar{r}) = \sum_l \sum_{r_1} \sum_{r_2} \ldots \sum_{r_S} p(l)[(r_1 + r_2 + \ldots + r_S)/S]\Pi_s q(s,l)$$

$$E(L) = \sum_l lp(l),$$

and

$$Var(L) = \sum_l l^2 p(l) - [E(L)]^2$$

Thus for every combination of number of raters, number of scale points, probability distributions of the latent responses, and conditional probability distributions for the raters assigning a scale value given a latent response, both the mean pairwise correlation of two raters and the correlation of the mean rating with the underlying latent response value can be obtained.

Suppose it is desired that the correlation of the mean rating with the latent response be at least .95. Then, if we can find a value of the mean pairwise interrater correlation at or above which r_{RL} is always above .95, we can use the pairwise interrater correlation as a gauge for the predictability of the latent response using the mean rater response as the predictor.

Latent Response Distributions. Four latent response probability distributions were used. The first was the uniform distribution with probabilities of all of the latent responses being equal. The second was the binomial distribution with n = the number of latent responses minus 1 and p = .5. Thus the probability of latent response x was binomial(L - 1, x - 1, .5). The third

distribution decayed exponentially as the number of the latent response increased, giving the j-shaped probability distribution

$$p(l) = \alpha^{l} / \sum_{l=1}^{L} \alpha^{l}$$

with α set equal to .6 for the present applications. The fourth distribution assigned to each latent response l the area under a normal curve over the interval from l - .5 to l + .5, where the mean of the normal distribution was L(.5) and the standard deviations was $\sqrt{[L(.5)^2]}$. These distributions are referred to as the uniform, binomial, skewed, and normal, respectively.

Rater bias and response dispersion. To provide additional generality to the results, various degrees of rater bias and rater accuracy were introduced. The conditional probability distribution of rating s given latent event l, $q(s,l)$, was assumed to be a normal distribution with mean

$$\mu_{(s,l)} = \beta\bar{L} + (1-\beta)l$$

and standard deviation σ, where $\bar{L}$ is the mean of the latent event values and β is a bias parameter. The value of $\mu_{(s,l)}$ is thus a weighted average of the actual latent event and the mean of all of the latent events. The parameter β gives the amount of weight or bias toward the overall mean and away from the actual latent event value. The parameter σ determines the precision with which the scale value is assigned to $\mu_{(s,l)}$. The greater is σ, the more likely is a large departure of the scale value from the weighted average. For the present purposes β ranged from 0 to .5 in steps of .1 and σ ranged from .2 to 2.0 in steps of .2. Thus the 6 levels of β times the

10 levels of σ resulted in 60 different normal probability distributions of $q(s,l)$ for each combination of a latent event distribution, number of raters, and number of scale values.

Procedure. Again, the purpose of the inquiry was to determine if, under a broad range of conditions, values of the mean pairwise rater correlation could be taken to indicate a correlation of the mean rating with the latent response that was at least .95. If this was found to be the case, then what would these mean pairwise rater correlation values be and how would they covary with number of raters and number of scale values. To achieve this, for each cell in the cross of four latent event distributions X four numbers of raters (2 - 5) X 9 numbers of scale values (2 - 10), the 60 different (β,σ)-pairs were used to give 60 different $q(s,l)$ distributions. These were used with equations [1] and [2] to obtain in each cell 60 pairs of values, the first member of each pair being the mean pairwise correlation between the raters and the second being the correlation of the mean rating with the latent response.

In each cell, the 60 pairs were sorted to find the lowest value of the mean pairwise correlation such that correlation of the mean rating with the latent response paired with that mean pairwise correlation was at least .95 provided that no value of the mean pairwise correlation greater than it was paired with a correlation of the mean rating with the latent response less than .95. This lowest value of the mean pairwise rater correlation was taken as the value for that cell that could be used by researchers as an indicator that a strong correlation of at least .95 between the mean of the ratings and the actual latent response existed.

Provided that these lowest values were similar over the various assumed distributions of the latent responses, such lowest values could then be used as a methodological standard.

Results

To provide a simple overview of the results, for a given combination of number of raters, number of latent events, and number of rating scale values, the largest of the four values for the four distributions of latent responses at that combination was obtained. These values are shown in Table 1. A dash in a cell indicates that for that combination of variables there was no mean pairwise correlation of raters that ever paired with a correlation of mean rating with latent value that was as large as .95.

Insert Table 1 here

Examination of this table yields a number of broad generalizations. First, it is clear that there is always a slight advantage to having at least 3 points on the rating scale and that when information is available as to how many latent responses might occur, one should use at least that same number of points on the rating scale. Furthermore, there is virtually never an advantage to using more points on the rating scale than there are latent responses that may occur. Of course that number will usually be unknown. By and large, unless there are as many as 10 underlying responses, it will rarely be useful to have more than seven points on the rating scale.

Averaging over the four distribution of latent values and focusing on the use of a 7-point scale, if we average over number of latent values (2 - 10), the average required mean rater correlation required for 2 raters is .86, for 3 raters is .78, for 4 raters is .76 and for 5 raters is .72. Instead of averaging over the four distributions, we can use the maximum values in Table 1. Averaging over latent values 2-10 and again focusing on a 7-point rating scale the corresponding values for 2, 3, 4, and 5 raters are .89, .82, .80, and .75.

From these results it appears that a very solid approach for the range of cases considered here is to use a rating scale containing at least 7 points and require a mean interrater correlation of .90 for 2 raters, .85 for 3 raters, .80 for 4 raters, and .75 for 5 raters. This should pretty much assure that the mean of the ratings for a subject on a given trial will correlate with the underlying latent event at a value of .95 or higher.

Some Simulation Results

In this part of the paper we consider results obtained by simulating the behavior assumed to characterize the raters. Methods of estimating parameters of the model and testing the assumptions of the model are considered. Alternative methods of assigning an individual score to the subject on a trial are compared. Also, procedures that will help when the assumptions of the model such as identical raters are violated.

Method

The model.

Assume that on each of a series of trials or occasions, the subject makes the latent response l_1 with probability p and the response l_2 with probability $1 - p$. On each trial each of R identical raters assigns a rating from 1 to S to the observed behavior independently of the other raters. Assume further that for each rater the identical probability distribution for the S ratings exists and depends only on which one of the L latent responses occurs. Let p_{sl} be the probability that a rater will assign rating s to latent response l. Then for each latent response l the probability distribution of the ratings for the K raters will be the multinomial distribution

$$[(K!/(n_1!\ n_2!\ \ldots\ n_S!)]\ p_{1l}^{\ n1}\ p_{2l}^{\ n2}\ \ldots\ p_{Sl}^{\ nS}$$

where n_s is the number of times rating s is given to latent response l by the K raters.

For example, suppose that an infant either smiles or does not smile on each of a series of trials and each of three raters independently assigns the values 1 or 2 depending on whether in the rater's individual judgment the infant did or did not smile. Then the conditional probability distribution of ratings for the three raters given that the infant smiles will be $f_2 = [(3!/(n_2!\ n_1!)]\ p_{22}^{n2}\ (1 - p_{22})^{n1}$ and the conditional probability distribution given the infant does not smile will be $f_1 = [(3!/(n_2!\ n_1!)]\ p_{21}^{\ n2}\ (1 - p_{21})^{n1}$. The marginal probability distribution of ratings will then be

$$p\{[(3!/(n_2!\ n_1!)]\ p_{22}^{\ n2}\ (1 - p_{21})^{\ n1}\} + (1 - p)\{\ [(3!/(n_2!\ n_1!)]\ p_{21}^{\ n2}\ (1 - p_{21})^{\ n1}\}$$

$$= pf_2 + (1 - p)f_1\ . \qquad [1]$$

which is a mixture of two binomial distributions.

Smiling and not smiling are two latent categories in that whether a smile occurred is not a part of the data but is assumed to probabilistically underlie the ratings which do make up the data. The parameter estimation task for this example is to estimate p, the probability of a smile, and p_{22} and p_{21} . If the model assumed that there were three latent response categories and each rater assigned one of three ratings on each trial, then there would be two latent response probabilities to be estimated and 3 sets of 2 conditional rating probabilities for a total of eight parameters. In what follows we limit our attention to the case of a binary latent response with raters using either a binary rating or a three-point rating scale. The bulk of the results concerns the case of three raters but some additional results for more raters are also reported.

Parameter estimation.

Let S_i be the sum of the ratings for three raters on trial i. For the case of binary rating with ratings of 2 or 1, S_i can take the values 3, 4, 5, or 6 as n_1 takes the values 3, 2, 1, or 0 and n_2 takes the values 3 - n_1 . The probability distribution for these values is given by Equation [1]. Let z_j be the observed proportion of trials on which the sum S_i occurs.

To find the parameter estimates a search of the parameter space for values of p and the conditional probabilities can be done using a procedure like that described by Chandler (1969) This procedure simply loops through all possible combinations of parameter values, using a grid of a certain grain on each parameter dimension. The combination of parameter estimates which are provisionally held are those which, up to that point in the search through the grid, minimize the $G2$ statistic, $\Sigma_i \{2^* z_j \ln(z_i / pred_i)\}$, where $pred_i$ is $pf_{2i} + (1 - p)f_{1i}$ with the parameter estimates inserted and is therefore the predicted proportion of occurrences of S_i. Minimizing the corresponding chi square goodness of fit statistic gives virtually the identical estimates. (See Uebersax (1988) for discussion of alternative methods of parameter estimation.) A grid grain of .01 was used for each parameter dimension in the present work.

An example with binary ratings. An example is presented here with data simulated according to the proposed model. On each of 1000 trials a simulated infant latently smiled with probability p = .80 and did not smile with probability .20. When a smile occurred, each of three identical simulated raters independently rated it as a 2 with probability p_{22} = .85 and a 1 with probability .15. When a smile did not occur, the probability of a rater using the 2 rating was p_{21} = .15 and using the 1 rating was .85. Thus, the raters are pretty good but not perfect at identifying the latent response.

Given the simulated rating data we can proceed as we would with real data to estimate the parameters of the model. The parameter search was constrained to the space of parameter values in which $p_{21} < p_{22}$. This is the minimal constraint that a rater will be more likely to say that a smile occurred when it actually occurred than when it did not. The three parameters p, p_{22}, and p_{21} then were .80, .85, and .15. The obtained estimates using the search routine on the simulated data gave corresponding estimates of .78, .86, and .16, with a value of $G2$ = .00006178.

By inserting the parameter estimates into Equation [1] we obtain the predicted proportions of trials on which the number of 2 ratings equals 3, 2, 1, and 0 and therefore the sum of the ratings equals 6, 5, 4, or 3. A comparison of the predicted to the observed proportions is given in Table 2. This agreement indicates that the simulation went as expected and the estimates of the parameters were close enough to provide good agreement with the observed proportions.

Insert Table 2 here

An example with a three-point rating scale. In this second example, the same procedure was followed except that the simulated raters made the rating responses 1, 2, or 3, with a greater number being associated with greater confidence that a smile occurred. The sum of the ratings on each trial, S_i, now can take one of the seven values 3, 4, 5, 6, 7, 8, or 9 and z_i, the observed proportion of occurrences of S_i, has the predicted probability

$$p\{ 3!/(n_1!\, n_2!\, n_3!)]\, p_{12}^{\,n1}\, p_{22}^{\,n2}\, p_{32}^{\,n3} \} + (1 - p)\{ 3!/(n_1!\, n_2!\, n_3!)]\, p_{11}^{\,n1}\, p_{21}^{\,n2}\, p_{31}^{\,n3} \}$$

$$= pf_2 + (1 - p)f_1 .$$

Simulated data were obtained with $p = .60$, $p_{12} = .15$, $p_{22} = .15$, $p_{32} = .70$, $p_{11} = .70$, $p_{21} = .15$, $p_{31} = .15$. Search of the parameter space yielded corresponding parameter estimates of .62, .15, .16, .69, .69, .16, and .15. The observed and predicted proportions of the sums of the ratings are given in Table 3.

Insert Table 3 here

Goodness of fit.

In the case of simulation we know the underlying probabilities of the latent events and the parameters for the three raters. In the research context these are of course unknown. One question we would like to answer is whether the assumption of identical raters is reasonable. Another is whether the model is fitting well otherwise. We exemplify the goodness of fit test of the assumption of identical raters using the simulated data from the example of the three-point rating scale. If the raters are identical, then we expect that over trials they should make the three ratings with approximately the same relative frequency. Let T be the number of trials, f_{ij} be the observed proportion of the T ratings that rater i made rating j, and f_j be the observed proportion of occurrences of rating j over the three raters. Then under the null hypothesis of identical raters the statistic $T\ \Sigma_i\Sigma_j\ [(f_{ij} - f_j)^2/f_j]$ is distributed as chi square on $(R - 1)(S - 1)$ df, where R is the number of raters and S is the number of different ratings, so in the example the df are $(3 - 1)(3 - 1) = 4$. For the simulated data the obtained value of chi square was 2.05 whereas the .05 level critical value is 9.49. As we would expect, the hypothesis of identical raters is not rejected. It is interesting to note that this chi square test constitutes a significance test of agreement between

the three raters in that a nonsignificant chi square inferentially supports the assumption that the three raters have the same parameter values.

Ordinarily, in performing the goodness of fit test for identical raters we would be concerned with the power of the test since the objective is to support the null hypothesis. Consequently we would probably choose a higher value of alpha than .05. But, as Uebersax (1988) indicates, with many observations even a small departure from identity will produce a significant result, so the researcher will have to exercise good judgment. Yet another consideration is that, as will be shown below, identical parameter values for the raters may not be necessary. It will turn out that sufficiently high pairwise correlations between the raters will do the job.

We can also calculate a chi square statistic to test the agreement of the observed proportions of the various sums of the ratings with the predicted proportions of those sums (see Table 3). This statistic is $\underline{T}\ \Sigma_i\ [(\underline{z}_i - \underline{pred}_j)^2 / \underline{pred}_j]$. Since there are seven different proportions, the df for the chi square statistic is equal to 7 - 1 minus the number of estimated parameters which is 5, so there is 1 df for the test. For the observed and predicted proportions in Table 3, the obtained chi square is 3.54 and the .05 level critical value is 3.84.

Results

Interrater agreement

Consider the data gathered from the three raters. We know that they were generated by a process that involved theoretically identical raters. One could assert that the raters were thus fundamentally in agreement so far as their criteria for assigning a rating is concerned. And yet, when we examine the observed and predicted proportion of trials on which 0, 2, or 3 raters

agreed in their rating we find the proportions shown in Table 4. These values indicate that only on one third of the trials did all

---

Insert Table 4 here

---

three raters assign the same rating. From this perspective we would conclude that the interrater agreement was not very impressive. If we turn to the parameter estimates for the individual raters we see that 30 per cent of the time that the smile occurs each rater does not assign the highest value, 3, and when the smile does not occur, 30 per cent of the time each rater does not assign the lowest value, 1. From a signal detection perspective we would say that the raters all have the same criteria for the use of the ratings 1, 2, and 3, but their sensitivity to the difference between a smile and a non-smile leaves something to be desired.

If we find that the model fits but the observable agreement needs improvement, we can use the information from the parameter estimates to work further on the training of the raters. For example, the parameter estimates might indicate that the raters are doing quite well in rating smiles when they do happen but not so well when they don't. We might then want to try to sharpen the definition of what is not a smile by using additional examples in training. Using the model to gauge the underlying rating process allows us to adjust the process to improve the observable rater agreement. When we cannot improve the underlying rating process by increasing the sensitivity of the individual raters, the results below suggest that increasing the number of raters may be an acceptable alternative solution.

Using the ratings to assign a score on each trial

Suppose that data from three raters have been gathered and that the raters appear to be sufficiently in agreement so that we can use the working assumption that they are identical. Later we consider the case where they are not. The next step is to use their data to assign a score to the response on each trial. How should this be done? Often the standard approach designates one of the raters as the primary rater and to use that rater's ratings as the basic measure for further analysis after confirming some level of agreement exists with a secondary rater. This is certainly one way to proceed. In what follows some alternative procedures are considered and then they are compared with each other and with the standard way in order to improve decision making on how to arrive at the basic measure to be used for subsequent analyses.

Using a rating scale based on Bayes' theorem. Once we have estimates of the model parameters in hand, it is possible to estimate the conditional probability that a smile occurred given a pattern of ratings by the raters. Let $f_j(n_1, n_2, ..., n_r)$ be the estimated conditional probability of n_1 ratings of 1, n_2 ratings of 2, etc. given that the latent event l_j occurred . These would be estimated by inserting the parameter estimates into the model values. For example, with three raters and a three-point scale, the parameter estimates would be inserted into $3!/(n_1!\ n_2!\ n_3!)]\ p_{12}^{\ n1}\ p_{22}^{\ n2}\ p_{32}^{\ n3}$ to obtain $f_2(n_1, n_2, n_3)$ and into $3!/(n_1!\ n_2!\ n_3!)]\ p_{11}^{\ n1}\ p_{21}^{\ n2} p_{31}^{\ n3}$ to obtain $f_1(n_1, n_2, n_3)$. Using the estimate of p in the following expression, it follows from Bayes' theorem that $g_1 = Pr(w_1 \mid n_1, n_2, ..., n_r) = pf_2(n_1, n_2, ..., n_r) \ / \ [pf_2(n_1, n_2, ..., n_r) + (1 - p)f_1(n_1, n_2, ..., n_r)]$. These conditional probabilities estimates can be viewed as a scaling of the likelihood that a given latent response was of type l_1 given the pattern of ratings displayed by the raters. The greater is the sum of the ratings, the greater will the conditional probability be. These scale values can even be transformed to values on the rating scale used by the raters. For example, if a three-point scale is being used, then $(3)g_1 + (1)(1 - g_1)$ will provide a value on the scale somewhere

between 3 and 1, being closer to 3 the greater is the estimated conditional probability that the latent response l_1 occurred. For the data used in the example above of three raters using a three-point scale, these scale values are shown in Table 5. This method has the additional virtue of using all the data from all the raters to arrive at the subject's score.

Insert Table 5 here

Choosing the best rater. A second method would be to choose the best rater. Since the latent events are unavailable, we might evaluate the raters by the extent to which they agree with each other. Since they make their ratings independently, any correlation between the raters must be based on their using common information resulting from their observation of the same behavior. One way to choose the best rater would be to choose that rater whose pairwise correlations with the other two raters average the highest. Call this the mean correlation method of choosing a single rater. A second way would be to find the pairwise kappa values for each pair of raters and choose the rater who has the highest average pairwise kappa value. Call this the kappa method of choosing a single rater.

Table 6 shows the best rater for the three-rater example data. Using either the mean pairwise correlation or the mean kappa method, rater 3 has a small edge for the title of best rater.

Insert Table 6 here

It should be noted that under the hypothesis that the three raters are identical, there is no best rater in terms of underlying process. Still, the hypothesis might be a little false, in which

case going for the best rater might do no harm. On the other hand, as Table 6 shows, even when the raters are identical, just by chance there can be a "best" rater. This might lead us away from using the data from all three raters when we should. This is discussed further in what follows.

Choose the best pair of raters. Another method would be to use either the pairwise correlations or the pairwise kappas in Table 6 to decide on the best pair of raters and then average their rating on each trial to obtain the basic measure. Both measures give a slight edge to the pair consisting of raters 2 and 3.

Averaging the ratings of the three raters. The final method to be considered is simply ignoring the differences in Table 6 which theoretically are due to chance variation if the raters are identical. Combining the ratings by averaging them then seems like the natural way to proceed since the variability of the mean of all of the raters will be less than the variability of the mean of any subset of them.

Assessing the different methods using the example data. One nice part of this modeling exploration and the simulation possibilities that accompany it is that with the simulation data, as opposed to the case with real data, we know on each trial what latent event actually occurred. Consequently, we can see how well the proposed measures predict the latent event. Table 7 shows the correlation of each of the proposed measures with the latent event.

Insert Table 7 here

It appears that for the example data, even though the scaled values are based on information from all three raters, the scaling is not a very good predictor of the latent response. Further, the more

raters are averaged to obtain a basic measure, the better the measure is at predicting the latent response.

How many raters to use? The "Gold Standard" and a rule of thumb

If the more raters we average, the better the prediction of the latent event, then how many should we use? To explore this question the number of raters was varied from 1 to 7 for sets of raters that ranged from poor to excellent. The definitions of poor to excellent are given in Table 8 which shows the parameter values for the raters for each qualitatively different set of raters.

Insert Table 8 here

For each qualitative level of raters and for number of raters varying from 1 to 7, an estimate of the correlation of the mean of the raters ratings with the value of the latent response was obtained by running the simulation 150 times and averaging the 150 obtained correlations. This was done for values of p, the probability of an l_1 event, equal to .5 and .8. Also, this was done for the ideal situation of 1000 trials being available and also for the more realistic situation where only 25 trials worth of ratings were available.

A correlation based on N observations has an estimated standard error of $\sqrt{[(1 - r^2)/(N - 2)]}$. The average of 150 of these would have a standard error equal to less than one twelfth of that. For 1000 trials and a correlation in the neighborhood of .5 - .8 the standard error of the mean of 150 r's is zero to three decimal places. When the r is based on only 25 observations, when r = .5 the standard error of the average of 150 r's is .015 and when r = .8, it is .010. Thus the values presented here are sufficiently reliable.

The first thing we notice in Table 9 is that when we have excellent raters, one rater can do quite well but three do a little better and there is not much gain after increasing to three except when the raters are less than excellent. The "gold standard" of rating seems to be to have three raters where the average pairwise correlation of the raters is .80 or better. In this case the correlation of the average rating with the latent response is in the range of .96 to .98. Perhaps the next thing we notice is that we can have that correlation greater than .90 with three raters provided that their pairwise correlations average greater than .60. However, with pairwise correlations averaging about .50 it takes about 5 raters to get up to .90. It is interesting to note how although individually these latter raters are not very sensitive to the latent event, the average of the ratings from five of them becomes a much more sensitive measure.

Insert Table 9 here

In most research reports it is never entirely clear what sort of standard is being applied to determine whether interrater agreement is good enough. Correlations are reported anywhere from the low .80s to the high .90s but no justification is given for why a value of say .85 is acceptable. The present results suggest a rough rule of thumb. <u>If you are using the data from a single rater after first checking the correlation of that rater with a second rater, then the correlation of the ratings of the single rater with the latent event will be equal to the square root of the pairwise correlation between the two raters</u>. A theorem to this effect can be proven provided that the two raters are identical and the only dependence between the raters is as a result of their rating the same events so that they are otherwise completely independent of each other. The data in Table 9 conform to the rule of thumb in that the correlation of the single rater

rating-latent response correlation with the square root of the mean pairwise interrater correlation is 1 to three decimal places, the slope of the regression equation is 1.003 and the intercept is -0.002. This serves to validate the simulation process. My recommendation is that if the data from a single rater are to be used as the basic measure for further analysis, then the correlation between that rater and a second one should be at least .90 so that one can be reasonably confident that the correlation of the rater's rating with the latent event is at least .95.

The present results also suggest that if the pairwise interrater correlation of .90 is too difficult to achieve, even after additional training, one should increase the number of raters so that the basic measure for future analysis is the average of their ratings for each trial. If two raters are used, the mean pairwise correlation need only be above .80 to have excellent reliability of the measure. Unless training or conducting the rating is very expensive in time or money, I would recommend that three raters will usually be optimal.

What if the assumption of identical raters is violated?

Thus far the results have been based on the assumption of identical raters. We now consider some results for the case where a mix of nonidentical raters is used. For these purposes the definitions of excellent, good, moderate, and poor raters continues to be as in Table 8. Table 10 shows the average correlations of three raters with each other and the correlation of their mean rating with the latent event for different sets of 3 raters with $\underline{p}$ = .50.

Insert Table 10 here

The results in Table 10 can be compared with the results in Table 9 for raters rating 1000 trials when $p = .50$. As an example, compare the value in Table 10 of .86 for 2 good raters and one moderate one to the value in Table 9 of .87 for 2 good raters without the one moderate one. Nothing is gained by including the moderate rater. On the other hand, compare the value in Table 10 of .94 for 1 excellent rater and two good ones to the value in Table 9 of .93 for 1 excellent rater without the two good ones. A very slight gain is possible by keeping the two good raters.

A provisional summary of the comparison between Table 10 and Table 9 would be that one should always drop the poor rater(s), always drop a lower rater when he or she is outnumbered by higher raters, and keep the lower raters if they are “one step” (meaning fairly close to) below the higher rater and outnumber the higher rater. For example, 1 good and 2 moderate raters are better than 1 good rater, but 2 good and 1 moderate are not quite as good as 2 good raters.

Stability of the process

The results considered thus far have all been based on the assumption that the probabilities of the two latent events remain constant over trials and that the raters remain constant also. To the extent that this is not true an assigned measure such as the mean of the ratings will be in error. The parameter estimates will be some sort of "mean value" while the underlying parameter values on any given trial will in general differ from them in a way related to how the parameters change over trials. A complete answer to how to deal with such a situation cannot be given because it depends on the specific empirical aspects of the research situation. A generalization of the present approach that includes attempting to model the manner in which the parameters change might be in order. With respect to the raters, training

refreshment might be interpolated after a given number of trials and rest periods could be incorporated into the rating task if ratings are of electronically stored imagery rather than of live participants.

The present approach does offer one attempt to gauge the stability of the process provided that enough trials are available. The trials can be blocked and parameters estimated for each block. The stationarity of the process could then be assessed using a likelihood ratio or chi square test.

To get some idea about how things might go if the three raters remained constant at the "good" level indicated in Table 8 but the value of p changed greatly over trials, p was made to depend on the trial number according to the equation $p_n = (.001)^n(.30) + (1 - .001)^n(.80)$ which, over the 1000 trials began at .795 and finished at .294, falling according to an exponential decay curve. The model fit as well as with constant p and the correlation of the mean rating of the three raters with the latent response was .89 which falls between the value of .87 for three good raters when $p = .80$ and the value .91 for three good raters when $p = .50$. With a second run where $p_n = (.001)^n(.20) + (1 - .001)^n(.70)$ taking p down from around .7 to around .2, the fit was again just as good and the correlation of the mean rating of the three raters with the latent response was again .89. These preliminary inquiries suggest that changes in the probability of the latent response over trials will not present much of a problem for the proposed approach.

Some example applications to work in the literature

In this section we provide some examples of how these results apply to work in the literature.

Rosenblum & Lewis (1999). As the first example of potential applications of the kind of modeling discussed above to research we consider the rating of facial attractiveness used by Rosenblum and Lewis (1999) in their study of the relations among body image, physical attractiveness and body mass in adolescence. In order to study the extent to which adolescents' body image was related to how others perceive them, it was necessary to obtain ratings of the facial attractiveness of 115 male and female adolescents. The investigators report that "attractiveness ratings were made via a 9-point rating scale anchored by the descriptors "very attractive" and very unattractive [reference omitted]. Two male and two female recent college graduates rated attractiveness. All four raters rated every participant. Pearson correlation coefficients were used to estimate inter-rater agreement. Mean interjudge agreement (for pairs of judges) was .53. All correlations between rater pairs were significant at $p < .001$. Despite some variability, examination of the data revealed a normal distribution of attractiveness ratings, with no outliers. The attractiveness score used in analyses was the average score across all raters."

To model the rating task, it was assumed that there were 9 latent attractiveness values which were normally distributed over the range from 1 to 9. The judges were assumed to be identical and to spread their ratings out around the true latent event value with a unimodal symmetric probability distribution centered at the true value. For ease of estimation it was assumed that the probabilities were $\underline{p}$ at the true value, $(1 - \underline{p})/3.5$ at the points adjacent to the true value, and $3/4(1 - \underline{p})/3.5$ at the points two steps from the true value. At the end points the symmetric distribution was truncated so that for example when the true value was 1, the predicted rating probabilities were $\underline{p}/(\underline{p} + 1 - \underline{p})/3.5 + 3/4(1 - \underline{p})/3.5)$ at 1, $[(1 - \underline{p})/3.5] /(\underline{p} + (1 - \underline{p})/3.5 + 3/4(1 - \underline{p})/3.5)$ at 2, and $[3/4(1 - \underline{p})/3.5] /(\underline{p} + (1 - \underline{p})/3.5 + 3/4(1 - \underline{p})/3.5)$ at 3.

The simulation was performed using 115 subjects and 150 replications of the experiment. The parameter p was adjusted to obtain a value that yielded a mean pairwise interrater correlation of .53. For this, $p = .247$. With $p = .247$, the mean correlation of the raters with the true facial attractiveness values averaged to .716 over the 150 replications. The correlation of the mean of the four raters with the latent event value was .890. Although the investigators do not provide a rationale for using four raters, it turns out that they were wise to do so. On the basis of the present attempt to model the rating process, they can claim adequately high interrater agreement even though the pairwise correlation was a mere .53.

Without the data it is hard to know whether they might have done better by dropping one or more of the raters. Rosenblum and Lewis state that all interrater pair correlations were significant at $p < .001$. This only requires that the correlation exceed .305. If some of the raters consistently gave pairwise correlations near .305 it is quite possible that more reliable rating was achievable using the average of the other highest raters than by using the average of all four. To check on this a simulation was run where two of the raters were improved by decreasing the variance of their ratings and two of the raters were degraded by increasing the variance of their ratings. The mean pairwise correlation of the raters was kept at .53, but now the two good raters had a pairwise correlation of .95 and the two poor ones had a pairwise correlation of .32, still significant at the .001 level as in the Rosenblum and Lewis study. The correlation of the mean of the four raters with the latent event value remained at .89 but the correlation of the mean of the two good raters with the latent event was now at .98. As we would expect from the material in the results section, substantial improvement in the reliability of the basic measure used for analysis would have been achieved had the two poor raters been removed.

In the light of these results, and since Rosenblum and Lewis found no relation between rated facial attractiveness and the other variables they manipulated, it might be useful for them to reconsider how many raters to use. Of course the differences between the simulated raters here are contrived to result in a worst case scenario. They may well have done the best that could be done if their raters were relatively homogeneous.

The NICHD Early Child Care Research Network (1998). In a report on early child care and self-control, compliance, and problem behavior (The NICHD Early Child Care Research Network,1998), a composite variable Positive Caregiving Ratings based on the sum of five qualitative ratings, was obtained for children aged 6, 15, and 24 months. A sixth and seventh category were added to the composite for children at 36 months. The researchers report interrater reliability estimates in the form of Pearson correlations as .94, .86, .81, and .80 (videotapes) and 90, .89, .89, and .90 (live) at each age. It appears that interrater reliability remained adequately high with the ratings of live child care behavior but that as the children got older the interrater reliability dropped from .94 at 6 months down into the low eighties at 24 and 36 months. The results of the present investigation indicate that the low interrater reliability estimates would have been improved if the average ratings for a pair of raters had been used as the basic data for analysis.

This study involved hundreds of participants. Consequently, having two raters rate all of the videotaped data might have been unmanageable. I could not tell from the report whether the ratings used for interrater reliability computations were over some of the data or all of them. If ratings from two observers had been gathered for all of the data, then of course it would have been an easy matter to use the mean of the two ratings. When very large numbers of participants are involved, a tradeoff of economy of labor versus reliability of the data and the consequent

increase in power will rear its ugly head. Judgments by investigators will be specific to individual cases.

Floyd, Gilliom, & Costigan (1998). A study of marriage and parenting by Floyd, Gilliom, and Costigan (1998) exemplifies the most typical type of interrater agreement assessment. As part of their study they used two measures which they called communication positiveness and negative reciprocity which were based on coding of videotaped discussions by married couples. They report that "reliability was assessed on 20% of the videotapes in this sample, which were independently evaluated by two coders. The overall agreement between coder pairs was $r = .81$ for communication positiveness scores, and $r = .73$ for negative reciprocity scores." Because of the square root law mentioned earlier these correlations can be regarded as the squares of the correlations between the respective variables and the latent attribute being measured, and therefore as proportions of variance accounted for. Values of .81 and .73 seem to me to be too low, especially when in this study there were only 79 couples so that the second coder coding the additional 80% left uncoded and perhaps even introducing a third coder does not seem unduly difficult. This would have had the effect of producing a much more reliable measure, reducing error variance, and therefore increasing power.

Discussion

In perusing a few issues of Child Development to find some examples of the applicability of these results to developmental research I was surprised at how many articles I came across where a measure of interrater agreement was called for but no such assessment was made. Also, quite frequently, where an assessment was made it was reported as, say, "interrater agreement was 93%" with no indication of what the percent referred to. Perhaps the most extreme form of obscurity encountered so far as how agreement was measured is exemplified by the report that

"All tasks were coded by undergraduate research assistants who were trained by the senior author ... . Training continued until coders achieved at least 90% reliability with the senior author." (Matthews, Ellis, & Nelson, 1966). Also of concern was the form of reporting of intraclass correlations as the measure of interrater agreement. It has been known for a quarter of a century that there are at least six different versions of the intraclass correlation, the correct choice of which depends on the "experimental design and the conceptual intent of the study" (Shrout & Fleiss, 1979). In none of the instances where this measure was used was there any indication of which intraclass correlation coefficient was being used or why that one had been selected.

Some standards for assessment and reporting of interrater agreement are needed. The present results can help with this in that they provide a means whereby at least those judgments made with a numerical rating scale can be made increasingly reliable by increasing interrater agreement through using enough raters and removing poor ones coupled with using the mean of the ratings as the basic measure.

Using the approach presented here it should be possible for an editorial requirement that a minimum correlation of .90 with the latent measure be obtained. One could extrapolate in Table 9 to find the needed number of raters for a given level of interrater agreement. For example, suppose that the pairwise interrater correlation were .71. Using the square root rule this would imply that for a single rater the correlation of rating with latent value is 84. Interpolating this value in the first column of Table 9 between the values .71 and .90 we see that it is (.84 - .71)/(.90 - .71) = .619 of the way above .71. Going to column 2 and finding the value that is the same fraction of the way between .82 and .95 we get .82 + .619(.95 - .82) = .90.

Thus, if the pairwise correlation between two raters is .84, the average of their two ratings would correlate .90 with the latent measure. Since it takes a pairwise correlation of .95 for the rating of a single rater to correlate .90 with the latent measure, we have the basis for a journal requirement: interrater pairwise correlations of .95 or above to justify the use of the data from a single observer as the basic data for analysis; pairwise correlations between .84 and .95 require the average of the two raters ratings be used as the basic measure. Below .84 requires more than two raters, with the number being determined by extrapolation.

The results in Table 9 are based on work with a 3-point scale. Comparable analysis using a 9-point scale indicated the square root rule holds very closely for the 9-point scale but that a pairwise correlation of .82 was big enough to get the correlation of the mean of two raters with the latent measure up to .9. This is very close to the .84 required with the 3-point scale. The difference probably was the result in differences in the assumptions about the distribution of judgment responses. Provisionally, the results suggest that .84 be used as the cutoff for small numbers of points on the rating scale and .82 be used for more points, but additional research is required to resolve this.

Fundamentally, the purpose of measuring interrater agreement is to demonstrate that when behavior is measured using human raters, a rater training process, in conjunction with the class of behaviors being rated, has been created such that the trained raters are reliable measuring devices. Such reliability is displayed by raters by their disposition to give the same rating to the same behavior. This in turn is due to their using comparable criteria for assigning one rating rather than another and to their having comparable sensitivity to the differences in the behavior being rated. Since the behavior of interest is not directly observable, confirmation of comparable criteria and sensitivity rests on using the agreement of the raters with each other and evaluating

that agreement within the context of a model of the rating process. The results given here reveal what can be expected to the extent that the rating process is well described by the model that was used.

One nice feature of the approach is that the model makes only minimal assumptions about what is going on: the behavior occurs with a fixed probability, the raters are identical, and they make their ratings in accord with a probabilistic process that remains the same over trials. The results show that when these assumptions are satisfied for the cases considered, firm useful results are available concerning what the probability of the latent behavior is, how reliable the raters are in their rating of the latent behavior, how many raters should be used as a function of how reliable they are individually, and how their individual ratings should be combined to arrive at the basic measure for further analysis: namely, average them.

The simulations also revealed that the interrater pairwise correlations are good measures of the correlations of the ratings with the latent behavior. The square root law provides that with two identical independent raters, the correlation of the ratings with the latent behavior will be the square root of the interrater pairwise correlation.

The simulations also revealed some information about how things fare when some of the assumptions are violated. The preliminary look suggests that trial to trial changes in the probability of the latent event will not interfere with using the model to assess the quality of the raters, their agreement, and arriving at a measure by averaging their ratings. Of course attempts to model the form of the change in p over trials might be informative and improve the whole process.

Much research uses rating scales with more points. Five and ten point scales are common. I would expect the present results to generalize to such scales. The general maxim

would be to assess the quality of the raters by using their pairwise interrater correlations and arrive at a basic measure by averaging the ratings of the good raters. Additional simulation is needed to discover the likely range of pairwise correlations for good raters when scales with more rating values are used but a preliminary look suggests that the results here are likely to generalize to scales with more points quite well. Additional work also would be useful to see how a good choice of the number of raters depends on the underlying response categories and their probabilities, and the number of points on the rating scale. Also additional simulation would be useful in exploring the question of how many rating points should be used.

It seems to me that an improved plan for rating latent behavior would involve some pilot work with raters on pilot subject data so that some preliminary assessment of the number of points on the rating scale and the number of raters needed would be based on the empirics of the particular situation rather than on traditional choices.

REFERENCES

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. Biometrics, 44(539-548).

Agresti, A. (1992). Modelling patterns of agreement and disagreement. Statistical Methods in Medical Research, 1, 201-218.

Carey, G., & Gottesman, I. I. (1978). Reliability and validity in binary ratings: Areas of common misunderstanding in diagnosis and symptom ratings. Archives of General Psychiatry, 35, 1454-1459.

Darroch, J. N., & McCloud, P. I. (1986). Category distinguishability and observer agreement. Australian Journal of Statistics, 28, 371-388.

Floyd, F. J., Gilliom, L. A., & Costigan, C. L. (1998). Marriage and parenting alliance: Longitudinal prediction of change in parenting perceptions and behaviors. Child Development, 69, 1461-1479.

Grove, W. M., Andreason, N. C., McDonald-Scott, P., Keller, B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis. Archives of General Psychiatry, 38, 408-411.

Janes, C. L. (1979). An extension of the random error coefficient of agreement to N X N tables. British Journal of Psychiatry, 134, 617-619.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. British Journal of Psychiatry, 130, 79-83.

Matthews, A., Ellis, A. E., & Nelson, C. A. (1996). Development of preterm and full-term infant ability on AB, recall memory, atransparent barrier detour, and means-end tasks. Child Development, 67, 2658-2676.

Rosenblum, G. D., & Lewis, M. (1999). The relations among body image, physical attractiveness, and body mass in adolescence. Child Development, 70, 50-64.

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

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. Archives of General Psychiatry, 42, 725-728.

Sprott, D. A., & Vogel-Sprott, M. D. (1987). The use of the log-odds ratio to assess the reliability of dichotomous questionnaire data. Applied Psychological Measurement, 11, 307-316.

Tanner, M. A., & Young, M. A. (1985a). Modelling agreement among raters. Journal of the American Statistical Association, 80, 175-180.

Tanner, M. A., & Young, M. A. (1985b). Modeling ordinal scale disagreement. Psychological Bulletin, 98(408-415).

The NICHD Early Child Care Research Network (1998). Early child care and self-control, compliance, and problem behavior at twenty-four and thirty-six months. Child Development, 69, 1145-1170.

Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. Psychological Bulletin, 101, 140-146.

Uebersax, J. S. (1992). A review of modeling approaches for the analysis of observer agreement. Investigative Radiology, 17, 738-743.

Uebersax, J. S., Grove, W. M. (1993). A latent trait finite mixture model for the analysis of rating agreement. Biometrics, 49, 823-835.

Zwick, R. (1988). Another look at interrater agreement. Psychological Bulletin, 103, 374-378.

Table 1. Maximum, over the four latent response distributions, of the minimum mean pairwise correlation required for the correlation of the mean rating response with the latent response to be at least .95.

Raters	Latent Events	Number of Scale Pts. 2	3	4	5	6	7	8	9	10
2	**2**	85	84	84	84	84	84	84	84	84
	3	-	92	92	92	92	92	92	92	92
	4	-	96	92	92	92	92	92	92	92
	5	-	-	93	89	85	85	85	85	85
	6	-	-	96	87	87	87	87	87	87
	7	-	-	-	98	86	85	85	85	85
	8	-	-	-	92	96	87	87	87	87
	9	-	-	-	-	87	90	86	87	87
	10	-	-	-	-	89	98	97	99	96
3		**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	2	85	84	84	84	84	84	84	84	84
	3	-	83	83	83	83	83	83	83	83
	4	-	96	79	79	79	79	79	79	79

5	-	-	86	81	81	81	81	81	81
6	-	-	89	82	82	82	82	82	82
7	-	-	-	94	82	78	78	78	78
8	-	-	-	88	86	87	89	89	89
9	-	-	-	-	79	82	80	79	79
10	-	-	-	-	89	86	80	77	78

4	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	75	75	75	75	75	75	75	75	75
3	-	83	83	83	83	83	83	83	83
4	-	89	78	78	78	78	78	78	78
5	-	-	86	76	76	76	76	76	76
6	-	-	87	82	82	82	82	82	82
7	-	-	-	94	73	73	74	74	74
8	-	-	-	86	82	83	72	73	73
9	-	-	-	-	79	79	73	74	75
10	-	-	-	-	79	91	77	71	72

5	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	75	75	75	75	75	75	75	75	75
3	0	71	71	71	71	71	71	71	71

4	0	82	78	78	78	78	78	78	78
5	0	0	77	68	68	68	68	68	68
6	0	0	84	82	87	87	87	87	87
7	0	0	0	74	68	67	67	67	67
8	0	0	0	84	74	67	68	68	68
9	0	0	0	0	77	73	68	70	70
10	0	0	0	0	79	91	71	69	67

Table 2. Observed and predicted proportions of trials on which the sum of the ratings is S_i.

	Observed	Predicted
6	.500	.497
5	.255	.256
4	.112	.114
3	.133	.133

Table 3. Observed and predicted proportions of trials on which the sum of the ratings is S_i.

S_i	Observed	Predicted
3	.124	.127
4	.088	.094
5	.153	.138
6	.096	.103
7	.197	.188
8	.138	.146
9	.204	.205

Table 4. Observed and predicted proportions of trials on which the raters agreed n times.

Number of rater agreements	Observed	Predicted
0	.087	.099
2	.576	.565
3	.337	.336

Table 5. Scale values for the seven possible sums of the ratings.

Sum of ratings	Scale Value
3	1.033
4	1.143
5	1.524
6	2.240
7	2.765
8	2.944
9	2.987

Table 6. Pairwise correlations and kappas and their averages for each rater.

Raters	Correlation	kappa	Rater	Correlation method	kappa method
1 & 2	0.258	0.497	1	0.300	0.519
1 & 3	0.341	0.541	2	0.305	0.523
2 & 3	0.352	0.549	3	0.347	0.545

Table 7 . Correlation of the various measures with the latent response.

Measure	Correlation with latent event
Scaled value	.363
Best rater	.587
Best pair	.703
Average rating	.766

Table 8. Rater conditional probabilities for four qualitatively different sets of identical raters.

	Rater Type			
Rater Parameters	Poor	Moderate	Good	Excellent
p_{11}	0.30	0.15	0.05	0.01
p_{12}	0.30	0.25	0.20	0.09
p_{13}	0.40	0.60	0.75	0.90
p_{21}	0.40	0.60	0.75	0.90
p_{22}	0.30	0.25	0.20	0.09
p_{23}	0.30	0.15	0.05	0.01

Table 9. Estimated correlation of mean rating of different numbers of raters with latent response value for various qualitative levels of raters for data based on 1000 rating trials and also on 25 trials, together with the mean pairwise correlation of raters.

Raters rating1000 Trials

p = .80

Type of Rater	Number of raters							
	1	2	3	4	5	6	7	r
Excellent	0.90	0.95	0.96	0.97	0.98	0.98	0.98	0.81
Good	0.71	0.82	0.87	0.90	0.91	0.93	0.94	0.50
Moderate	0.44	0.57	0.64	0.70	0.74	0.77	0.79	0.19
Poor	0.10	0.14	0.16	0.19	0.21	0.23	0.25	0.01

p = .50

Type of Rater								
	1	2	3	4	5	6	7	
Excellent	0.93	0.97	0.98	0.98	0.99	0.99	0.99	0.87
Good	0.78	0.87	0.91	0.93	0.94	0.95	0.96	0.61
Moderate	0.52	0.65	0.73	0.77	0.81	0.83	0.85	0.27
Poor	0.12	0.17	0.20	0.24	0.26	0.29	0.31	0.02

Raters rating 25 Trials

p = .80

Type of Rater Number of raters

	1	2	3	4	5	6	7	r
Excellent	0.89	0.94	0.96	0.97	0.97	0.98	0.98	0.80
Good	0.69	0.79	0.85	0.88	0.90	0.92	0.93	0.48
Moderate	0.43	0.55	0.63	0.68	0.72	0.75	0.78	0.18
Poor	0.10	0.13	0.15	0.18	0.20	0.22	0.24	0.01

p = .50

Type of Rater

	1	2	3	4	5	6	7	
Excellent	0.93	0.96	0.98	0.98	0.99	0.99	0.99	0.87
Good	0.78	0.87	0.91	0.93	0.94	0.95	0.96	0.61
Moderate	0.52	0.65	0.73	0.77	0.81	0.84	0.86	0.27
Poor	0.12	0.17	0.20	0.23	0.25	0.29	0.31	0.01

Note. Each data point is the mean of 150 replications.

Table 10. Average correlations of three raters with each other and with latent events for different sets of 3 raters: p = .50.

RATER TYPES				Mean pairwise rater correlation	Correlation of mean of raters with latent event
Excellent	Good	Moderate	Poor		
3				0.88	0.98
2	1			0.78	0.96
2		1		0.60	0.93
2			1	0.34	0.89
1	2			0.70	0.94
1		2		0.41	0.85
1			2	0.07	0.67
1	1	1		0.53	0.90
1	1		1	0.33	0.85
1		1	1	0.22	0.79
	3			0.62	0.91
	2	1		0.48	0.86
	2		1	0.28	0.81
	1	2		0.36	0.81
	1		2	0.05	0.57
	1	1	1	0.20	0.72
		3		0.26	0.73

2	1	0.12	0.61
1	2	0.02	0.43
	3	-0.01	0.22

Note. Each data point is the mean of 150 replications.