Educational and Psychological Measurement
1970, 30, 61-70.

ESTIMATING THE RELIABILITY, SYSTEMATIC ERROR AND RANDOM ERROR OF INTERVAL DATA

KLAUS KRIPPENDORFF
The Annenberg School of Communications
University of Pennsylvania, Philadelphia

In content analysis as in all situations in which unstructured observations play an important role, the reliability with which data are generated is of crucial importance. Low data reliability limits the confidence in the validity of subsequent inferences and the reliability of a population of data must be *estimated from the agreement* among many observers regarding a sample.

The way reliability of data is assessed is not different in principle from the way the reliability of psychological tests is measured. However, in the process of developing recording instructions, defining units of analysis and operationalizing scales, the researcher requires more detailed information about the sources and kind of unreliability. Over-all measures of agreement do not provide such information readily.

More specifically, the analyst of a recording instrument may wish to obtain:

1. An estimate of the reliability of a population of data over all observers in the universe using that recording instrument. This measure is called *data reliability* and can be interpreted as a measure of the confidence in data.
2. An estimate of the extent to which data reliability could be improved if scale values were to be transformed or their definitions were to be modified for the individual observers. This measure assesses the *systematic error* of the recording process, which, together with a measure of the *random error* may be said to account for the lack of data reliability.

3. An estimate of the reliability associated with each individual observer, often called *individual reliability*. Such an estimate permits the identification of observers who are detrimental to achieving high data reliability. Deviant observers need either more instruction or cannot be employed in the process of collecting data.
4. An estimate of the extent to which each observer is corrigible by further instruction. Such an estimate would assess *systematic observer biases* which together with the individual's *random error* account for lack of individual reliability.
5. Finally, there is needed an indication of the extent to which a random sample of observers agree on the scoring of each unit of recording. This measurement may be called *unit reliability* and allows to identify for further inspection sources of unreliability within the set of observations.

The following aims at explicating these analytical devices.

Reliability Data

Situations in which a set of n independent observers record m unstructured units into an interval-scale representation usually generate the following kind of data:

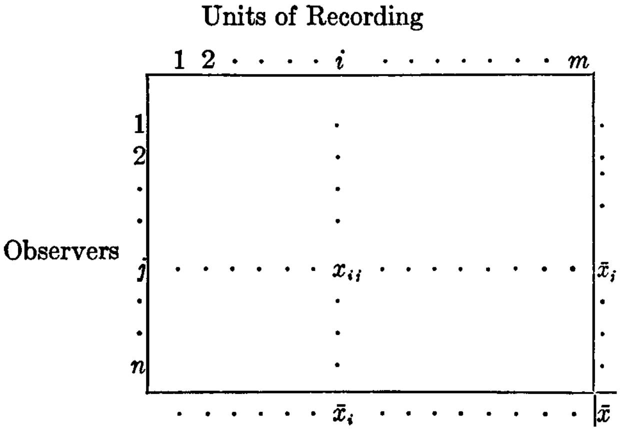

Figure 1. Form of reliability data.

Here, x_{ij} is the scale value that the j-th observer assigned to the i-th unit of recording, $\bar{x}_i$, $\bar{x}_j$ are the means over columns and rows respectively, and $\bar{x}$ is the mean score over all reliability data.

Analysis of Variance

Such interval-scale data almost always lead to some form of the analysis of variance and this calculus has often been used to measure reliability. Kendall's (1948) coefficient of concordance is one familiar example and the measure of agreement proposed by Robinson (1957) is another. Hoyt (1951), Ebel (1951), Alexander (1947), Tryon (1957), Engelhart (1959) and others discuss the implications of this approach to the reliability of psychological tests.

One of the problems encountered in such applications is to define measures of reliability that are unbiased with respect to both the sample size of the recording units and of the number of observers. Rajaratnam (1960) solved this problem by defining the n by m ratings x_{ij} in terms of an estimated true variance over the population of recording units and an estimated error variance for each such unit over the universe of observers. Our reliability formuli essentially agree with this approach except that a distinction is introduced between a systematic error and a random error:

$$x_{ij} = \text{est}\,(\bar{x}) + \text{est}\,(\bar{x}_i - \bar{x}) + \text{est}\begin{pmatrix}\text{Systematic}\\ \text{Error}\end{pmatrix} + \text{est}\begin{pmatrix}\text{Random}\\ \text{Error}\end{pmatrix} \qquad (1)$$

In assigning scale values to given units of recording, observers can deviate from each other in two ways: observers may differ in their tendency to use one side of the scale rather than the other. This becomes manifest in the deviation $(\bar{x}_j - \bar{x})$ above chance. Observers may also prefer or avoid the extreme values of a scale and thus magnify or reduce the true differences among recording units. This tendency appears in the deviation $[(\bar{x}_i - \bar{x}) - (x_{ij} - \bar{x}_j)]$. Both deviations can be corrected once they are known either by specialized instructions to the observers or by subsequent data transformation. Both deviations thus contribute to the systematic error of a recording insrument.

It is, therefore, appealing to use a model in which the rating x_{ij} is defined in terms of an estimated true score for the i-th observation over all observers, a transformation that accounts for the j-th observer's idiosyncracies for all observations and the random error.

$$x_{ij} = \text{est}\,(\bar{x}_i) + \text{est}\,(\alpha_j + \beta_j(x_{ij} - \bar{x}_i)) + \text{est}\begin{pmatrix}\text{Random}\\ \text{Error}\end{pmatrix} \qquad (2)$$

from (1) and (2) follows:

$$\text{est}\begin{pmatrix}\text{Systematic}\\ \text{Error}\end{pmatrix} = \text{est}\ (\alpha_i + \beta_i(x_{ii} - \bar{x}_i)) \tag{3}$$

$$\text{est}\begin{pmatrix}\text{Random}\\ \text{Error}\end{pmatrix} = \text{est}\ ((1 - \beta_i)(x_{ii} - \bar{x}_i) - (\bar{x}_i - \bar{x})) \tag{4}$$

Following Snedecor (1956), Edwards (1958) and others, Rajaratnam defined variance by the sample sum of squared deviations divided by the degrees of freedom. He showed that this definition gives rise to variance estimates that are unbiased with respect to the number of observers and the number of units recorded. We will base our formuli on his results.

It can be shown that the sum of the squared random error becomes minimum if the two constants are:

$$\alpha_i = (\bar{x}_i - \bar{x}) \tag{5}$$

and

$$\beta_i = 1 - \frac{\sum_i (x_{ii} - \bar{x}_i)(\bar{x}_i - \bar{x})}{\sum_i (x_{ii} - \bar{x}_i)^2} \tag{6}$$

In proceeding analogously to a two-way analysis of variance, unbiased variance estimates are obtained according to Table 1, which gives an estimate of the variance among units of recording, an estimate of the error variance that can be eliminated by systematic transformation and an estimate of the random error variance.

Data Relibaility

Following Rajaratnam, data reliability, is defined as the proportion of the estimated variance over the population of recording units in data and the estimated total variance for the universe of observers over the population of data. It is an estimate of the agreement among observers:

$$\hat{a} = \frac{\hat{V}_a}{\hat{V}_t} = \frac{m(n-1)SS_a - mSS_r}{m(n-1)SS_a + n(m-1)SS_s + (mn - m - n)SS_r} \tag{18}$$

In general, Rajaratnam's reliability estimate $\hat{p}$ and the proposed measure are related as follows: $\hat{a} \geq \hat{p}$. If the systematic error is merely caused by $\alpha_i \neq 0$ then $\hat{a} = \hat{p}$. However, if $\beta_i \neq 0$ then in the proposed formula some of Rajaratnam's residual error variance

TABLE 1

Variance Estimates from Reliability Data

	Sum of Squares		Mean Squares		Unbiased Variance Estimates	
Among Units (Columns)	$SS_a = \sum\sum(\bar{x}_i - \bar{x})^2$	(7)	$MS_a = \frac{SS_a}{m - 1}$	(11)	$\hat{V}_a = \frac{1}{n}(MS_a - MS_r)$	(14)
Systematic Error	$SS_s = \sum\sum(\alpha_j + \beta_j(x_{ij} - \bar{x}_j))^2$	(8)	$MS_s = \frac{SS_s}{n - 1}$	(12)	$\hat{V}_s = \frac{1}{m}(MS_s - MS_r)$	(15)
Random Error	$SS_r = \sum\sum((1 - \beta_j)(x_{ij} - \bar{x}_j) - (\bar{x}_i - \bar{x}))^2$	(9)	$MS_r = \frac{SS_r}{(m - 1)(n - 1)}$	(13)	$\hat{V}_r = MS_r$	(16)
Total	$SS_t = \sum\sum(x_{ij} - \bar{x})^2$	(10)			$\hat{V}_t = \hat{V}_a + \hat{V}_s + \hat{V}_r$	(17)

appears in the systematic error variance associated with the set of observers and is not used to estimate the variance over the units of recording. Consequently the proposed measure is larger than $\hat{p}$. (In our example below $\hat{a} = .27$ while $\hat{p} = .22$.) However, if the errors are low, which is required for acceptable data, then the difference is negligible.

Robinson's measure of agreement, to mention only one more example, boils down to the proportion SS_a/SS_t. While of great simplicity, the coefficient does not provide an estimate. Its value tends to be too large.

Systematic Error

The extent to which data reliability might be improved if α_j and β_j become known is consequently defined as the proportion of the estimated variance that is systematically dependent on the observers and the estimated total variance:

$$\hat{s} = \frac{\hat{V}_s}{\hat{V}_t} = \frac{n(m-1)SS_s - nSS_r}{m(n-1)SS_a + n(m-1)SS_s + (mn - m - n)SS_r} \tag{19}$$

The remaining random error is defined as:

$$\hat{r} = \frac{\hat{V}_r}{\hat{V}_t} = \frac{mnSS_r}{m(n-1)SS_a + n(m-1)SS_s + (mn - m - n)SS_r} \tag{20}$$

It follows from the above that:

$$\hat{a} + \hat{s} + \hat{r} = 1 \tag{21}$$

Individual Reliability and Observer Biases

To associate a measure of agreement with each observer, attention is focussed on his covariation with the estimated true scores $\bar{x}_i$ for each unit of recording, and on the two errors for which he can be held responsible. Analoguously to Table 1 and the definitions (18), (19) and (20) estimates for individual reliability, $\hat{a}_j$, the systematic observer bias, $\hat{s}_j$, and the random error, $\hat{r}_j$, can be defined as in Table 2. By analogy to (21):

$$\hat{a}_j + \hat{s}_j + \hat{r}_j = 1 \tag{28}$$

Unit Reliability

The agreement achieved on the i-th recording unit by a random sample of observers from the universe cannot be derived in the

TABLE 2

Individual Reliability and Observer Biases

	Partial Sum of Squares	Individual Reliability and Observer Biases
Among Units (Columns)	$SS_{aj} = \sum_i (x_{ij} - \bar{x})^2 - \sum_i (x_{ij} - \bar{x}_i)^2$ (22)	$\hat{a}_j = \frac{m(n-1)SS_{aj} - mSS_{rj}}{m(n-1)SS_{aj} + n(m-1)SS_{sj} + (mn - m - n)SS_{rj}}$ (25)
Systematic Error	$SS_{sj} = \sum_i (\alpha_j + \beta_j(x_{ij} - \bar{x}_i))^2$ (23)	$\hat{s}_j = \frac{n(m-1)SS_{sj} - nSS_{rj}}{m(n-1)SS_{aj} + n(m-1)SS_{sj} + (mn - m - n)SS_{rj}}$ (26)
Random Error	$SS_{rj} = \sum_i ((1 - \beta_j)(x_{ij} - \bar{x}_i) - (\bar{x}_i - \bar{x}))^2$ (24)	$\hat{r}_j = \frac{mnSS_{rj}}{m(n-1)SS_{aj} + n(m-1)SS_{sj} + (mn - m - n)SS_{rj}}$ (27)

same way as above. This is because the equivalent to the systematic observer bias is quite meaningless as far as the units of recording are concerned. In expressing the agreement on the i-th observation, use is made of the fact that the errors are contained in the sum of squares within columns, $SS_{wc} = SS_s + SS_r$. Consequently the total error variance for which the i-th observation accounts may be estimated by:

$$\hat{V}_{ei} = (\hat{V}_s + \hat{V}_r) \frac{\sum_i (x_{ij} - \bar{x}_i)^2}{SS_s + SS_r} \tag{29}$$

and the agreement can then be defined as:

$$\hat{a}_i = \frac{\hat{V}_t - m\hat{V}_{ei}}{\hat{V}_t} \tag{30}$$

The relation between $\hat{a}_i$ and $\hat{a}$ then becomes:

$$\hat{a} = \frac{1}{m} \sum_i \hat{a}_i \tag{31}$$

It should be noted that such a relation does not exist between $\hat{a}_j$ and $\hat{a}$.

Example

Figure 2 shows the hypothetical results of the rating of 10 observations by six observers using the same instrument. With a data reliability of $\hat{a} = .27$, the data generated by this means are not acceptable by any standard. However, transforming the data will help and making the observers aware of their biases might improve the recording procedure. In fact the proposed measures indicate that reliability could increase by the amount of the systematic error, $\hat{s} = .58$, from $\hat{a} = .27$ to .85.

Upon inspection of the measures associated with each individual, deviant observers will be detected immediately. For example, the 2nd observer fails to differentiate among observations and is, therefore, in disagreement with the other observers far above chance. The proportionately small systematic error suggests no significant improvements. The proposed measures also indicate how observers differ in the interpretation of the recording instrument. The 3rd observer seems to have not only used a different means but also inverted the scale. However, the large systematic error suggests that

	Units of Recording $i =$													
	1	2	3	4	5	6	7	8	9	10	$\bar{x}_j$	$\hat{\alpha}_j$	$\hat{s}_j$	$\hat{r}_j$
Observer $j =$ 1	5	3	3	2	1	2	3	4	5	2	3	1.00	0.00	0.00
2	2	2	2	2	2	2	2	2	2	2	2	−2.16	1.00	2.16
3	7	4	5	6	7	2	5	4	3	7	5	−.40	1.15	.25
4	7	5	3	1	−1	2	3	4	7	−1	3	.63	.34	.03
5	4	3	2	1	0	2	2	3	3	0	2	.43	.40	.17
6	5	1	3	0	−3	2	3	7	10	2	3	.50	.47	.03
$\bar{x}_i$	5	3	3	2	1	2	3	4	5	2	3	$\hat{\alpha}$	$\hat{s}$	$\hat{r}$
$\hat{\alpha}_i$	.40	.67	.80	.27	.94	1.00	.80	.53	−.54	−.27		.27	.58	.15

Figure. 2. Numerical example.

if he were to be informed about his biases, individual reliability might be improved from −.40 to +.75. Similarly does the systematic error of the 4-th observer account for the fact that he exaggerates the differences among recording units. If the population of data from which the reliability sample was taken is already recorded then the systematic observer bias, $\hat{s}_4 = .34$, implies that a transformation of the data generated by the 4-th observer will yield an individual reliability of .97, etc.

Little needs to be said about unit reliability. The 6-th observation contributes most and the 9-th observation contributes least to the observed disagreement. A subsequent inspection of those observations on which agreement is low may yield further information about the sources of unreliability contained in the recording instrument.

It might be noted that the nature of the proposed measures does not lead to formuli that can be easily computed. However, since most of these assessments are done by computers there is no reason to give preference to a simpler solution which yields less information.

REFERENCES

Alexander, H. W. The estimation of reliability when several trails are available. *Psychometrika,* 1947, 12, 79–99.

Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika,* 1951, 16, 407–424.

Edwards, A. L. *Statistical analysis* (Rev. Ed.), New York: Rinehart, 1958.

Engelhart, M. D. A method of estimating the reliability of ratings compared with certain methods of estimating the reliability of tests. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1959, 19, 479–588.

Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika,* 1951, 6, 153–160.

Kendall, M. G. *Rank correlation methods,* London: Griffin, 1948.

Rajaratnam, N. Reliability formulas for independent decision data when reliability data are matched. *Psychometrika,* 1960, 25, 261–271.

Robinson, W. S. The statistical measure of agreement. *American Sociological Review,* 1957, 22, 17–25.

Snedecor, G. W. *Statistical methods* (5th ed.), Ames, Iowa: Iowa State College Press, 1956.

Tryon, R. C. Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin,* 1957, 54, 229–249.