Computing interrater reliability on the Apple Macintosh computer

JOHN O. BROOKS III

Stanford University School of Medicine, Stanford, California

and

LAURA L. BROOKS

Rice University, Houston, Texas

and SKOPOS Corporation, Los Altos, California

A program is described for computing interrater reliability by averaging, for each rater, the correlations between one rater's ratings and every other rater's ratings. For situations in which raters rate more than one ratee, raters' reliabilities can be computed for either each item or each ratee. The program reads data from a text file and puts the reliability coefficients in a text file. The standard Macintosh interface is implemented. The QuickBASIC program is distributed both as a listing and in compiled form; it can be run with advantage with math coprocessors.

The agreement among raters, or interrater reliability, is a central issue in the assessment of judgments in many areas of psychology. There are several indices of interrater reliability, such as the intraclass correlation (Guilford & Fruchter, 1978), but one that is frequently used is the average of the correlations between one rater's ratings and every other rater's ratings (Rosenthal & Rosnow, 1984). Suppose, for example, that several raters rate a person on 20 personality characteristics. If there are only two raters, then the best estimate of interrater reliability is the correlation between their ratings. For three raters, each rater's reliability coefficient is the average of two correlation coefficients: The correlations between one rater's ratings and the ratings of the other two raters. Thus, each rater's reliability coefficient is based on the average of $(n - 1)$ correlations, where n is the number of raters.

Interrater reliability coefficients can be used in a variety of ways. As Rosenthal and Rosnow (1984) have pointed out, before conducting an experiment one can use effective reliability, which may be thought of as the collective reliability of a set of raters, to determine the number of raters that should be used. Effective reliability is computed by applying the Spearman-Brown prophecy formula (Rosenthal & Rosnow, 1984; Walker & Lev, 1953) to the average of the interrater reliabilities. In the prophecy formula, shown in Equation 1, R = effective reliability, n = number of raters, and $\bar{r}$ = average reliability coefficient across raters.

$$R = \frac{n \bar{r}}{1 + (n - 1) \bar{r}} \quad (1)$$

Correspondence should be addressed to John O. Brooks III, Department of Psychiatry and Behavioral Sciences (TD-114), Stanford University School of Medicine, Stanford, CA 94305-5490.

Thus, if one knows both the desired effective reliability and the number of raters to be used, the corresponding average interrater reliability may be determined with the aid of the table provided by Rosenthal (1982). Similarly, if the desired effective reliability and an estimate of interrater reliability are known, one may determine how many raters are required to achieve this level of reliability.

The average interrater reliability coefficient is sometimes used as a dependent measure in behavioral research and so may be taken as an index of the effects of experimental manipulations on judgments. Interrater reliability is applicable in areas other than those that involve rating other persons. Experimental manipulations may, for example, affect the reliability of subjects' ratings of job characteristics (Hahn & Dipboye, 1988; Harvey & Hayes, 1986).

Unfortunately, to the best of our knowledge, no statistics packages provide a convenient way of computing interrater reliability. In this paper, we report a QuickBASIC program for the Macintosh that provides a rapid and convenient way of computing average interrater reliability coefficients.

The Program

The data must be in a text, or ASCII, file, which can be created by most Macintosh applications. Each rater's ratings must be on a separate line, and the ratings on each line must be separated by commas. Because some applications create tab-delimited files, a program is included to convert tab-delimited files to comma-delimited files. To increase the speed of the calculations, the program uses machine language routines (see Brooks, 1987) that are in a machine language library called ExtraLib. The routines in ExtraLib are from the machine language library CLR MathStatLib and are distributed with the permission of Clear Lake Research (2476 Bolsover, Suite 343, Houston, TX 77005). Because of the use of machine language routines, the program runs faster on Macintosh computers with a math coprocessor (e.g., Macintosh SE/30, II, IIfx, IIfx, and IIfx) than on those without coprocessors.

Figure 1. A sample set of ratings.

Figure 2. The display with which parameters of the data file are entered.

When raters rate more than one ratee, it is possible to compute interrater reliabilities either across items (i.e., for each ratee) or across ratees (i.e., for each item). If reliabilities are computed across items, each rater will have as many reliability coefficients as there are ratees. Similarly, if reliabilities are computed across ratees, each rater will have as many reliability coefficients as there are items. To compute reliability coefficients across ratees, there must be more than one ratee. When there is more than one ratee, the data should be grouped as in Figure 1, where each ratee is on a separate line and the ratees are in the same order for each rater. Thus, if 10 raters rate six ratees on 14 dimensions, the data file should have 60 lines of 14 numbers. When each rater rates only one ratee, there is only one line of ratings per rater, and interrater reliability must be computed across items. Because there is no established technique for dealing with missing data in computations of interrater reliability, all raters must have complete data.

After the program has been started, the screen shown in Figure 2 appears. After the user enters the number of raters, ratees, and items, the file containing the ratings can be selected by clicking on the Ratings button. Next, the output file, which will contain the reliabilities, can be created by clicking on the Output file button. The Begin button will remain dimmed and inactive until all parameters have been set and all files have been selected.

The program options are set by clicking on the Options button, after which the window shown in Figure 3 will appear. Fisher's r to Z transformation can be used to average the correlations (Fisher, 1928; Hays, 1988; Kraemer & Thiemann, 1987; Silver & Dunlap, 1987). With this technique, each r is converted to a Z and the Z s are averaged. The average Z is then converted back to an r . Either the average r , obtained from the Z to r transformation, or the average Z may be saved in the output file. Because some applications require tab-delimited text files, either commas or tabs may be used in the output file to separate the variables. After all options have been selected, the Done button can be clicked and the original screen will

Figure 3. The display that allows the setting of optional features.

- .033, - .432, .103, - .191	← rater 1
.410, .091, .011, .054	← rater 2
.216, .157, .367, .289	← rater 3
↑ ↑ ↑ ↑	
1 2 3 4	
	Ratee

Figure 4. The reliability coefficients for the sample data that are computed across items. The results have been truncated to three decimal places for illustrative purposes.

- .134, .385, -.006, -.943, .193	← rater 1
-.571, -.127, -.097, -.408, .114	← rater 2
-.734, .571, .268, .280, .574	← rater 3
↑ ↑ ↑ ↑ ↑	
1 2 3 4 5	
	Item

Figure 5. The reliability coefficients for the sample data that are computed across ratees. The results have been truncated to three decimal places for illustrative purposes.

Table 1
Pairwise Correlations for the Three Raters for Each Item

Rater Pair	Item Number				
	1	2	3	4	5
r_{12}	.152	-.355	-.361	-.957	-.330
r_{13}	-.399	.828	.351	-.924	.626
r_{21}	.152	-.355	-.361	-.957	-.330
r_{23}	-.896	.114	.181	.778	.517
r_{31}	-.399	.828	.351	-.924	.626
r_{32}	-.896	.114	.181	.778	.517

Note—These correlations are derived from the sample data in Figure 1. The subscripts next to each r denote the rater pair for the correlation.

Table 2
Interrater Reliability Coefficients:
Average r s After an r to Z Transform of the Data in Table 1

Rater	Item Number				
	1	2	3	4	5
1	-.134	.385	-.006	-.943	.193
2	-.571	-.127	-.097	-.408	.114
3	-.734	.571	.268	-.280	.574

Table 3
Pairwise Correlations for the Three Raters for Each Ratee

Rater Pair	Ratee Number			
	1	2	3	4
r_{12}	.182	-.485	-.264	-.411
r_{13}	-.244	-.376	.445	.050
r_{21}	.182	-.485	-.264	-.411
r_{23}	.596	.612	.283	.497
r_{31}	-.244	-.376	.445	.050
r_{32}	.596	.612	.283	.497

Note—These correlations are derived from the sample data in Figure 1. The subscripts next to each r denote the rater pair for the correlation.

Table 4
Interrater Reliability Coefficients:
Average r 's After an r to Z Transform of the Data in Table 3

Rater	Ratee Number			
	1	2	3	4
1	-.033	-.432	.103	-.191
2	.410	.091	.011	.054
3	.216	.157	.367	.289

appear. After the Begin button is clicked, a status window displays the number of the rater for which reliability is being computed.

The reliability coefficients are placed in an output file and are delimited by either commas or tabs. The reliability coefficient(s) for each rater are on a single line. The coefficients for the sample data from Figure 1 are shown in Figures 4 and 5. To document the accuracy of the program, the values of the intermediate calculations for the sample data shown in Figure 1 are provided in Tables 1–4. Tables 1 and 2 provide the intermediate values used to compute interrater reliability across ratees, and Tables 3 and 4 provide the intermediate values used to compute reliability across items.

Data files can contain a maximum of 32,767 lines and 32,767 items, as long as there is enough available

memory; each number in the data file requires 4 bytes of RAM. The reliability coefficient and all intermediate variables are kept in single precision, which provides seven significant digits. All calculations are performed in at least extended precision.

Hardware and Software Requirements

An Apple Macintosh computer with at least 1 MB of RAM is required to run the program. A hard disk, though not required, will greatly increase the program's speed, as will a math coprocessor. Microsoft QuickBASIC is *not* required for use of the compiled version of the program. The necessary machine language routines from CLR MathStatLib have been placed in ExtraLib, which is distributed with the program. Included with the program disk is a sample file of data on which interrater reliabilities have been computed.

Availability

An 800K disk containing the program listing, the compiled version of the program, the machine language routines, and the sample data may be obtained from J. O. Brooks, by sending \$6 for shipping and handling. J. O. Brooks may be contacted through Bitnet by sending a message to MAILER@Stanford and placing the following on the first line of the message: To: johnb%alois.uucp@med. A copy of the program that is sent through Bitnet may be sent either as Macintosh files encoded in BinHex format (which can be decoded with programs such as Stuffit) or as text files. Program requests through Bitnet should include a mention of the desired file format.

REFERENCES

- BROOKS, J. O., III. (1987). Enhancing and degrading visual stimuli. *Behavior Research Methods, Instruments, & Computers*, *19*, 260-269.
- FISHER, R. A. (1928). *Statistical methods for research workers*. London: Oliver & Boyd.
- GUILFORD, J. P., & FRUCHTER, B. (1978). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.
- HAHN, D. C., & DIPBOYE, R. L. (1988). Effects of training and information on accuracy and reliability of job evaluations. *Journal of Applied Psychology*, *73*, 146-153.
- HARVEY, R. J., & HAYES, T. L. (1986). Monte Carlo baselines for interrater reliability correlations using the position analysis questionnaire. *Personnel Psychology*, *39*, 345-357.
- HAYS, W. L. (1988). *Statistics* (4th ed.) New York: Holt, Rinehart and Winston.
- KRAEMER, H. C., & THIEMANN, S. (1987). *How many subjects?* Newbury Park, CA: SAGE.
- ROSENTHAL, R. (1982). Conducting judgment studies. In K. R. Scherer and P. Ekman (Eds.), *Handbook of methods in nonverbal behavior research* (pp. 287-361). New York: Cambridge University Press.
- ROSENTHAL, R., & ROSNOW, R. L. (1984). *Essentials of behavioral research*. New York: McGraw-Hill.
- SILVER, N. C., & DUNLAP, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, *72*, 146-148.
- WALKER, H. M., & LEV, J. (1953). *Statistical inference*. New York: Holt, Rinehart and Winston.