GENERAL RELIABILITY AND INTRACLASS CORRELATION PROGRAM (GRIP)

Ron D. Hays (RAND and University of California, Los Angeles)
Eric Wang (University of Southern California)
Michael S. Sonksen (University of California, Los Angeles)

ABSTRACT

Reliability is the extent to which a measure yields a similar value each time it is administered, all other things being equal (i.e., no true change in the attribute being measured has occurred). The simplest reliability model is derived from a one-way ANOVA with the targets (persons or things) being rated as the between factor and the remaining variance assigned to the within error term. If the number of assessments (raters) is the same across targets, it is possible to estimate the main effect of assessment (i.e., mean shifts in responses). The two-way fixed effects model estimates the reliability of multiple assessments by subtracting the mean square error from the mean square between, dividing by the mean square between. The mean square error is estimated by the interaction between respondents and the multiple assessments (the main effect of multiple assessments is excluded from the error term). The two-way random effects model assumes that the different assessments (e.g., raters) are randomly selected. In this model, the main effect of multiple assessments is incorporated into the estimate of total variability. This paper describes a SAS® macro that computes reliability estimates and intraclass correlations for the one-way and two-way ANOVA models.

INTRODUCTION

Reliability refers to the extent to which the measure yields the same number or score each time it is administered, all other things being equal (i.e., no true change in the attribute being measured has occurred). Observed scores include a true score component, a systematic error component, and a random error component. If no random error is present, the reliability is 1.0. Reliability approaches zero as the relative amount of random error increases. Both the true score component and systematic error contribute to the reliability of the measure because they drive the observed score for an individual towards a consistent value. However, systematic error leads to bias in measurement, because it causes the score to be consistently too high or too low relative to the true score. Reliability assessment involves examining agreement between an individual's score on two or more measures of the same thing. There are four basic categories of reliability estimation, each reflecting somewhat different ways by which random error of measurement is estimated: inter-rater, equivalent-forms, test-retest, and internal consistency reliability.

Inter-rater reliability refers to a comparison of scores assigned to the same target person by two or more raters. Both rater selection and intra-individual response variability influence random error in this case.

Data from an experimental study of the effect of exposure to light on the growth of plants is presented to illustrate the estimation of inter-rater reliability. Ten house plants were randomly assigned to one of two experimental conditions: 1) exposed to indoor light; or 2) not exposed to light (i.e., kept in a dark closet). The intervention lasted 7 days and the dependent variable was growth of the house plants. Height was measured to the nearest 16th of an inch using a wooden 12-inch ruler by two raters. The raw data from this study is provided in Table 1.

Table 1--Raw Data for Ratings of Height of House Plants

Plant	Experimental Condition	Rater	Height Baseline	Height Followup
A1	1			
		1	120	121
		2	118	120
A2	2			
		1	084	085
		2	096	088
B1	2			
		1	107	108
		2	105	104
B2	1			
		1	094	100
		2	097	104
C1	2			
		1	085	088
		2	091	096
C2	1			
		1	079	086
		2	078	092
D1	1			
		1	070	076
		2	072	080
D2	2			
		1	054	056
		2	056	060
E1	1			
		1	085	101
		2	097	108
E2	2			
		1	090	084
		2	092	096

Note: Height was measured to the nearest 16th of an inch.

For data such as these, the Pearson product-moment correlation coefficient is sometimes used to estimate inter-rater reliability. The coefficient indicates the extent to which individuals (plants) who received high scores (ratings of height) from one rater also tend to receive high scores from the other rater(s), and the extent to which those who receive low scores from one rater also tend to receive low scores from the other rater(s). A limitation of product-moment correlations is the fact that systematic differences in mean ratings (e.g., one rater consistently rates people higher than do other raters) are not reflected in the statistic. The intraclass correlation coefficient, in contrast, is sensitive to variation in systematic differences in ratings as well as relative ordering of different respondents. In addition, more than two ratings are easily summarized by the intraclass correlation coefficient.

The simplest variant of intraclass correlation is derived from a one-way ANOVA with the persons or things being rated as the between factor and the remaining variance assigned to the within error term. Table 2 provides the calculating formulas for this and other models discussed below.

The reliability column in Table 2 lists formula for the reliability of the average of the multiple assessments (ratings) and the intraclass correlation column provides formula for the reliability of a single assessment. In inter-rater reliability evaluations such as this house plant study, one would be most interested in the estimated reliability for a single rating or assessment (i.e., the intraclass correlation) if a single rating is all that is available for most subjects (plants) in the study. The reliability estimate for the average of multiple assessments would be of most interest if one has multiple ratings for all or most subjects (plants).

Table 2--Formula to Calculate Reliability and Intraclass Correlation for Various Models

Model	Reliability	Intraclass Correlation
One-way	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS} + (K-1)\, MS_{WMS}}$
Two-way fixed effects	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (K-1)\, MS_{EMS}}$
Two-way random effects	$\frac{N\ (MS_{BMS} - MS_{EMS})}{N\ MS_{BMS} + MS_{JMS} - MS_{EMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (K-1)\ MS_{EMS} + K\,(MS_{JMS} - MS_{EMS})/N}$

Note: Winer (1971) provided an unbiased formula for the one-way model:

$$\text{Theta} = \frac{MS_{BMS} - MS_{WMS}}{K\,M\,MS_{BMS}}$$

$$M = (N\,(K-1)) \;/\; (N * (K-1) - 2)$$

$$\text{Reliability} = (K\ \text{Theta}) \;/\; (1 + K\ \text{Theta})$$

$$\text{Intraclass Correlation} = \text{Theta} \;/\; (1 + \text{Theta})$$

Where N = number of respondents; K = average number of assessments per respondent.

If the number of assessments (raters) is the same across respondents, it is possible to estimate the main effect of assessment (i.e., mean shifts in responses). The two-way fixed effects model estimates the reliability of the average of the multiple assessments by subtracting the mean square error from the mean square between, divided by the mean square between. The mean square error is estimated by the interaction between respondents and the multiple assessments (the main effect of multiple assessments is excluded from the error term). For the house plant example, the estimated reliabilities of the average rating and single rating under this model, respectively, are 0.97 and 0.95.

The two-way random effects model assumes that the different assessments (e.g., raters) are randomly selected, and is appropriate if raters can be said to have been selected at random. In this model, the main effect of multiple assessments is incorporated into the estimate of total variability. For the house plant study, the estimated reliabilities of the average rating and single rating under this model are 0.98 and 0.96, respectively.

Equivalent-forms reliability refers to the agreement between an individual's score on two or more measures designed to measure the same attribute. Both item selection and intra-individual response variability contribute to random error in this method of estimating reliability. If the forms are truly equivalent in terms of item content, then this estimate provides a good estimate of their reliability. However, it is difficult to devise equivalent forms and intervening events or practice effects can distort the results from this method of reliability assessment. The same approach used for inter-rater reliability can be used to estimate equivalent-forms reliability.

Test-retest reliability is the relationship between scores obtained by the same person on two or more separate occasions. Intra-individual response variability is used to estimate random error in test-retest assessments. The approach described above for inter-rater reliability is the same one used for test-retest reliability, with multiple times of assessment substituted for multiple raters. Several factors may influence the reliability of a measure between test dates, such as the conditions of administration, testing effects, specific factors affecting the participants in their daily lives, or the length of time between administrations. The assessment of reliability is further complicated by the fact that changes in the attribute being measured may have occurred between administrations. A low test-retest estimate may therefore not accurately reflect the reliability of the test. Thus, test-retest assessments become less useful to the extent that real changes occur from the first to the second assessment of the attribute being measured.

Internal consistency is a function of the number of items and their covariation within a scale measuring a particular construct. Random error due to item selection is modeled in this type of reliability estimate. Cronbach's (1951) alpha is the coefficient commonly used to estimate the reliability of instruments based on internal consistency. Cronbach's alpha is calculated using the two-way fixed effects model described above with items serving as a main effect (rather than, e.g., raters or retests). Generally, one is most interested in the reliability of the average of the items (instead of the reliability of a single item, intraclass correlation). Formulas for computing the significance of difference between alpha coefficients are provided elsewhere (Feldt, Woodruff, & Salih, 1987).

For each reliability model, the intraclass correlation can be derived from the estimated reliability for multiple assessments using a variant of the Spearman-Brown prophecy formula (Clark, 1935):

$$Rii^* = \frac{Rtt}{K + (1-K)\,Rtt}$$

Likewise, the reliability of the multiple assessments can be obtained from the intraclass correlation using the following formula:

$$Rtt^* = \frac{K\,Rii}{1 + (K-1)\,Rii}$$

USING THE MACRO

Required input to the macro is the name of the input data set, the variable name for the between group factor, the variable name for the replicate factor (e.g., rater), the name of the variable for which reliability is being estimated, the ANOVA model to be estimated (two-way: type=1; one-way: type=0), and title information for the two tables produced by the program. Raw data is arranged with multiple lines of input per case (a separate line of input per replicate).

Output from the program for the house plant data presented in Table 1 is provided in Table 3 (ANOVA summary) and Table 4 (reliability and intraclass correlation estimates).

The GRIP macro is provided in Table 5. The macro invocation is as follows:

```
%GRIP(indata=a, targetv=id, repeatv=rater, dv=height1,
        type=1, t1=source of variance in baseline rating of
        height in house plant study, t2=reliability and
        intraclass correlation estimate for houseplan study)
```

Table 3–Analysis of Variance Output from GRIP Macro

Source of Variance in Baseline Rating of Height in House Plant Study

Source	Degrees of freedom	Mean square	Label for Mean square
Ratees (N-1)	9	628.67	BMS
Within	10	17.70	WMS
Raters (K-1)	1	57.80	JMS
Raters x Ratees	9	13.24	EMS
Total	19		

Table 4–Reliability and Intraclass Correlation Output from GRIP MACRO

Model	Reliability	Intraclass Correlation
One-way		
Biased	0.972	0.945
Unbiased	0.965	0.932
Two-way		
Fixed effects	0.979	0.959
Random effects	0.972	0.946

Table 5–GRIP Macro

```
%MACRO twoway;

proc glm data=&indata outstat=stats noprint;
  class &targetv &repeatv;
  model &dv = &targetv &repeatv ;
  run;

proc sort data=stats;
  by _name_ _SOURCE_;
  run;

data allrel;
  retain bdf bms jdf edf wms jms ems k;
  set stats;
  by _name_;
  if _type_='SS1' then delete;
  if _source_='ERROR' then do;
    ems=ss/df;
    edf=df;
  end;
  if _source_="%upcase(&targetv)" then do;
    bms=ss/df;
    bdf=df;
  end;
  if _source_="%upcase(&repeatv)" then do;
    jms=ss/df;
    jdf=df;
    k=df+1;
  end;
  if last._name_ then do;
   wms=((ems*edf)+(jms*jdf))/(edf+jdf);
   n=bdf+1;
m=(n*(k-1))/(n*(k-1)-2);
theta=(bms-(m*wms))/(k*m*wms);
rii=theta/(1+theta);
rtt=(k*theta)/(1+(k*theta));

fixed=(bms-ems)/(bms+((k-1)*ems));
fixedk=(bms-ems)/bms;
biased=(bms-wms)/(bms+(k-1)*wms);
k=(bms-wms)/bms;
random=(bms-ems)/((bms)+((k-1)*ems)+((k*(jms-ems))/n));
rk=(bms-ems);
randk=rk/(bms+((jms-ems)/n));
   output;
  end;
run;

data _null_;
file print header=hea1 ps=64 notitles;
set  allrel;
kkm=jdf+edf;
kk3=edf+bdf+jdf;
*+++++++++1+++++++++2+++++++++3+++++++++4+++++++++
5+++++++++6+++++++++7+++++++++8+++++++++9++++++++1
0;
put    @20           'Degrees of    mean     Label  for'/
       @5  'Source        freedom' @35 'square' @45 'mean
Square'/
       @5 '--------------------------------------------';

put    @5 'Ratees  (N-1)'  @24 bdf  3.   @34 bms 7.2  @49 'BMS'/
       @5 'Within'       @24 kkm 3.   @34 wms 7.2  @49 'WMS'/
       @7 'Raters  (K-1)'  @25 jdf  3.   @34 jms 7.2  @49 'JMS'/
       @7 'Raters x Ratees ' @25 edf 3. @34 ems 7.2  @49
'EMS'/
       @5 '--------------------------------------------'/
       @5 'Total'         @24 kk3   3. /;
return;

hea1:
  do;
  put @5 "&t1 "/;
  end;
```

```
      return;
run;

data _null_;
file print header=hea1 ps=64 notitles;
set allrel;
*+++++++++1+++++++++2+++++++++3+++++++++4+++++++++
5+++++++++6+++++++++7+++++++++8+++++++++9+++++++++1
0;
put    @5 'Model         Reliability        Intraclass Correlation'/
       @5 '----------------------------------------------------';
put    @5 'One way'/
       @7 'Biased'      @24 k 5.3    @53 biased 5.3 /
put    @7 'Unbiased'    @24  rtt 5.3    @53 rii 5.3 //
       @5 'Two-way'/
       @7 'Fixed effects'      @24  fixedk 5.3    @53 fixed 5.3 /
       @7 'Random effects'     @24 randk 5.3    @53 random 5.3
/
       @5 '----------------------------------------------------'/;
return;

hea1:
  do;
  put @5 " &t2 "//;
  end;
     return;
run;
%MEND twoway;
****************;
%MACRO oneway;
proc anova data=&indata outstat=est1 noprint;
class &targetv;
model &dv= &targetv;
run;

data est;
 set est1;
retain;
if _type_='ERROR' then wms=ss/df;
if _type_='ERROR' then n=df;
if _type_='ERROR' then errdf=df;
if _type_='ANOVA' then bms=ss/df;
if _type_='ANOVA' then betdf=df;
if _type_='ANOVA' then nrated=df+1;
if _type_='ANOVA' then nn=n+df+1;
if _type_='ANOVA' then k=nn/nrated;
OUTPUT;
data est;
 set est;
m=(n*(k-1))/(n*(k-1)-2);
theta=(bms-(m*wms))/(k*m*wms);
rii=theta/(1+theta);
rtt=(k*theta)/(1+(k*theta));
frtt=(bms-wms)/bms;
frii=frtt/(k * (1 + (frtt * (1/k - 1))));
OUTPUT;
run;
DATA EST;
 SET EST;
if k ne .;
RUN;

data _null_;
file print ps=64 notitles;
set  est;
k2=k-1;
n2=n-1;
k3=n2+n;
kk=betdf+errdf;
*+++++++++1+++++++++2+++++++++3+++++++++4+++++++++
5+++++++++6+++++++++7+++++++++8+++++++++9+++++++++1
0;
put    @20          'Degrees of    mean    Label  for'/
       @5  'Source        freedom' @35 'square' @45 'mean
Square'/
       @5 '----------------------------------------------------';

put    @5 'Between    ' @24 betdf  5.   @34 bms 7.2  @49 'BMS'/
       @5 'Within'      @24 errdf  5.   @34 wms 7.2  @49 'WMS'/
       @5 '----------------------------------------------------'/
       @5 'Total'       @24 kk     5. /;
return;

data _null_;
file print ps=64 notitles;
set  est;
*+++++++++1+++++++++2+++++++++3+++++++++4+++++++++
5+++++++++6+++++++++7+++++++++8+++++++++9+++++++++1
0;
put    @5 'Model         Reliability        Intraclass Correlation'/
       @5 '----------------------------------------------------';
put    @5 'One way'/
       @7 'Biased'      @24 frtt 5.3    @53 frii 5.3 /
       @7 'Unbiased'    @24  rtt  5.3    @53  rii 5.3 //
       @5 '----------------------------------------------------'/;
return;

%MEND oneway;
*********************************************;

%MACRO
grip(indata=,targetv=,repeatv=,dv=,nrepeatv=,type=,t1=,t2=);
 %IF %EVAL(&type) %then %DO;
  %twoway;
 %end;
%ELSE
 %DO;
  %oneway;
 %end;
%MEND grip;
*********************************************;
```

ACKNOWLEDGEMENTS

The development of GRIP was supported by RAND from its internal funds. We thank Robert M. Hamer, Ph.D., Virginia Commonwealth University, for code we adopted from his RELIAB.SAS® macro (code obtained from Tor Neilands by electronic mail on April 4, 1995 after submitting the GRIP abstract to WUSS).

REFERENCES

Clark, E.L. (1935). Spearman-brown formula applied to ratings of personality traits. **J Educ Psych, 26**, 552-555.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. **Psychometrika, 16**, 297-334.

Feldt, L.S, Woodruff, D.J., & Salih, F.A. (1987). Statistical inference for coefficient alpha. **Appl Psych Measurement, 11**, 93-103.

Winer, B.J. (1971). **Statistical principles in experimental design.** New York: McGraw-Hill.

FOOTNOTES

* Rii = intraclass correlation; Rtt = reliability of average assessment; K= number of assessments per respondent.

Direct Correspondence to: Ron D. Hays, RAND, 1700 Main Street, Santa Monica, CA 90407-2138, Ronald_Hays@rand.org (internet), office phone: (310) 393-0411 (extension 7581), FAX: (310) 393-4818. GRIP will be sent by electronic mail upon request.