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

<table>
<thead>
<tr>
<th>Plant</th>
<th>Condition</th>
<th>Rater</th>
<th>Baseline</th>
<th>Followup</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1</td>
<td>1</td>
<td>120</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>118</td>
<td>120</td>
</tr>
<tr>
<td>A2</td>
<td>2</td>
<td>1</td>
<td>084</td>
<td>085</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>096</td>
<td>088</td>
</tr>
<tr>
<td>B1</td>
<td>2</td>
<td>1</td>
<td>107</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>105</td>
<td>104</td>
</tr>
<tr>
<td>B2</td>
<td>1</td>
<td>1</td>
<td>094</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>097</td>
<td>104</td>
</tr>
<tr>
<td>C1</td>
<td>2</td>
<td>1</td>
<td>085</td>
<td>088</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>091</td>
<td>086</td>
</tr>
<tr>
<td>C2</td>
<td>1</td>
<td>1</td>
<td>079</td>
<td>086</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>078</td>
<td>092</td>
</tr>
<tr>
<td>D1</td>
<td>1</td>
<td>1</td>
<td>070</td>
<td>075</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>072</td>
<td>080</td>
</tr>
<tr>
<td>D2</td>
<td>2</td>
<td>1</td>
<td>054</td>
<td>056</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>056</td>
<td>060</td>
</tr>
<tr>
<td>E1</td>
<td>1</td>
<td>1</td>
<td>085</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>097</td>
<td>108</td>
</tr>
<tr>
<td>E2</td>
<td>2</td>
<td>1</td>
<td>090</td>
<td>084</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>092</td>
<td>096</td>
</tr>
</tbody>
</table>

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 formulae for this and other models discussed below.

The reliability column in Table 2 lists formulae for the reliability of the average of the multiple assessments (ratings) and the intraclass correlation column provides formulae 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>
<thead>
<tr>
<th>Model</th>
<th>Reliability</th>
<th>Intraclass Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-way</td>
<td>$\frac{MS_{BMS} \cdot MS_{WMS}}{MS_{BMS}}$</td>
<td>$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS} + (K \cdot 1) MS_{WMS}}$</td>
</tr>
<tr>
<td>Two-way fixed effects</td>
<td>$\frac{MS_{BMS} \cdot MS_{EMS}}{MS_{BMS}}$</td>
<td>$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (K \cdot 1) MS_{EMS}}$</td>
</tr>
<tr>
<td>Two-way random effects</td>
<td>$\frac{N (MS_{BMS} \cdot MS_{EMS})}{N MS_{BMS} + MS_{EMS} - MS_{EMS}}$</td>
<td>$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (K \cdot 1) MS_{EMS} + K (MS_{JMS} - MS_{EMS}) / N}$</td>
</tr>
</tbody>
</table>

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

$$\text{Theta} = \frac{MS_{BMS} \cdot MS_{WMS}}{KM MS_{BMS}}$$

$$M = \frac{(N (K \cdot 1))}{N (K \cdot 1) - 2}$$

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

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

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) alphas 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 differences between alpha coefficients are provided elsewhere (Feldt, Woodruff, & Sallin, 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, 1985):

\[
R_{tt} = \frac{K + (K-1) \cdot R_{II}}{1 + (K-1) \cdot R_{II}}
\]

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

\[
R_{II} = \frac{R_{tt} - 1}{K - 1}
\]

**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:

\[
\text{%GRIP(indata=a, targetv=id, repeatv=rater, dv=height1, type=1, t1=source of variance in baseline rating of height in house plant study, t2=reliability and intraclass correlation estimate for houseplant study)}
\]

**Table 3—Analysis of Variance Output from GRIP Macro**

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Mean square</th>
<th>Label for Mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rates (N-1)</td>
<td>9</td>
<td>628.57</td>
<td>BMS</td>
</tr>
<tr>
<td>Rates (K-1)</td>
<td>1</td>
<td>57.50</td>
<td>JMS</td>
</tr>
<tr>
<td>Rates x Rates</td>
<td>9</td>
<td>13.24</td>
<td>EMS</td>
</tr>
<tr>
<td>Within</td>
<td>10</td>
<td>17.70</td>
<td>WMS</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4—Reliability and Intraclass Correlation Output from GRIP MACRO**

<table>
<thead>
<tr>
<th>Model</th>
<th>Reliability</th>
<th>Intraclass Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-way</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biased</td>
<td>0.972</td>
<td>0.945</td>
</tr>
<tr>
<td>Unbiased</td>
<td>0.985</td>
<td>0.932</td>
</tr>
<tr>
<td>Two-way</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed effects</td>
<td>0.979</td>
<td>0.959</td>
</tr>
<tr>
<td>Random effects</td>
<td>0.972</td>
<td>0.946</td>
</tr>
</tbody>
</table>

**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 stat;
   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-1)*(k-1));
      theta=(bms-(m*ems))/((k*m*ems));
      rii=theta/(1-theta);
      Rtt=(k*theta)/(1+(k*theta));
   end;
   if last._name_ then do;
      wms=((ems*edf)+(jms*jdf))/(edf+jdf);
      n=bdf+1;
      m=(n*(k-1))/((n-1)*(k-1));
      theta=(bms-(m*ems))/((k*m*ems));
      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*ems)/n);
    kkm=jdf+bdf+jdf;
    kk3=edf+bdf+jdf;
    *+++++++1+++++++++2+++++++++3+++++++++4+++++++++5+++++++++6+++++++++7+++++++++8+++++++++9+++++++++1 0;
    put @20 'Degrees of mean Label for/Square/';
    put @5 'Source freedom @35 'square' @45 'mean Square/';
    put @5 'Coevers @35 'square' @45 'mean Square/';
    put @5 'Rate @24 bdf 3. @34 bms 7.2 @49 BMS/';
    put @5 'Within' @24 bms 7.2 @49 WMS/';
    put @5 'Rates (K-1)' @25 jdf 3. @34 jms 7.2 @49 JMS/';
    put @5 'Rates x Rates' @25 edf 3. @34 ems 7.2 @49 EMS/';
    put @5 'Total' @24 kkm 3. 1.;
    return;
end;
return;
run;

data _null_;                  
file print header=hea1 ps=64 notitites;       
set allio;                        
*+++++++++1+++++++++2++++++++3++++++++4++++++++5++++++++6++++++++7++++++++8++++++++9++++++++10;   
put @5 'Model Reliability Intraclass Correlation'/                  
*5 '-------------------------------';                       
put @5 'One way'                  
put @7 'Biased' @24 k 5.3 @53 biased 5.3 /                  
put @7 'Unbiased' @24 r 5.3 @53 ri 5.3 //                  
put @5 'Two-way'                  
put @7 'Fixed effects' @24 fixedf 5.3 @53 fixed 5.3 /                  
put @7 'Random effects' @24 randf 5.3 @53 random 5.3 /                  
/                  
@5 '-------------------------------';                       
return;

heat:                  
do;                     
put @5 *)&t2 */;  
end;                     
run;                     
%MEND twoway;                    
******************************;                     
%MEND oneway;                    
******************************;                     
proc anova data=aindata outstat=est1 noprint;                     
class &targetv;                     
model &dv= &targetv;                     
run;                     

data est;                     
set est1;                     
retain;                     
if _type_ = "ERROR" then wms=ss/dd;                     
if _type_ = "ANOVA" then r=dd;                     
if _type_ = "ERROR" then errdf=dd;                     
if _type_ = "ANOVA" then bms=ss/dd;                     
if _type_ = "ANOVA" then betdf=dd;                     
if _type_ = "ANOVA" then nrated=dd/d+1;                     
if _type_ = "ANOVA" then k=n/n/nrated;                     
OUTPUT;                     
data est;                     
set est;                     
m=(n*(k-1))/((n*(k-1)-2);                     
theta=(bms-(m*wms))/(k*m*wms);                     
r=theta(1+theta);                     
rtt=(theta/(1-theta));                     
frtt=(m-wms)/bms;                     
ri=frtt*(k*(k-1)/(k*(k-1)-1)));                     
OUTPUT;                     
run;                     
DATA EST;                     
SET EST;                     
if k ne ;                     
RUN;

data _null_;                  
file print ps=64 notitites;                  
set est;                  
k2=k-1;                  
n2=n-1;                  
k3=n2+n;                  
k=kbetdf+errdf;                  
*+++++++++1+++++++++2++++++++3++++++++4++++++++5++++++++6++++++++7++++++++8++++++++9++++++++10;                  
@20 'Degrees of mean Label for'/                  
@5 'Source freedom' @35 'square' @45 'mean

Statistics & Research

%MACRO grip(indata=,targetv=,repeatv=,dv=,nrepeatv=,type=,n1=,n2=);                  
%IF %EVAL(type) %then %DO;                  
%weway;                  
%end;                  
%ELSE;                  
%weway;                  
%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).

SAS is a registered trademark of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

REFERENCES


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.