작문교육론연구

4. 작문 평가에 있어서의 평가자 간 신뢰도 산출을 위한 SAS 프로그램 예시

```
data ratings;
input student @;
do rater = 1 to 4;
input score @;
output;
end;
cards;
9 2 5 8
6 1 3 2
8 4 6 8
7 1 2 6
8 5 6 9
6 2 4 7
5 4 6 7
8 6 7 6
7 5 6 5
5 3 6 2
;
proc print;
proc glm outstat=sums;
class student rater;
model score= student rater student*rater;
test h=student rater e=student*rater;
data sums;
set sums;
retain sss sdf rss rdf iss idf;
if _source_='student' and _type_='SS1' then do;
sss=ss;
sdf=df;
end;
if _source_='rater' and _type_='SS1' then do;
rss=ss;
rdf=df;
end;
if _source_='student*rater' and _type_='SS1' then do;
iss=ss;
idf=df;
output;
end;
proc print;
data sums;
set sums;
n=sdf+1;
```

```
    sms=sss/sdf;
    ems=iss/idf;
    rms=rss/rdf;
    wsms=(rss+iss)/(idf+rdf);
    icc1=(sms-wsms)/(sms+(rdf*wsms));
    icc2=(sms-ems)/(sms+(rdf*ems)+((rdf+1)*(rms-ems)/(sdf+1)));
    icc3=(sms-ems)/(sms+(rdf*ems));
    icc1k=(sms-wsms)/sms;
    icc2k=(sms-ems)/(sms+(rms-ems)/n);
    icc3k=(sms-ems)/sms;
    proc print; var icc1 icc2 icc3 icc1k icc2k icc3k;
    run;
```

5. 평가자 신뢰도 산출 방법 1
 1) 다음 사이트 방문
 https://department.obg.cuhk.edu.hk/researchsupport/IntraClass_correlation.asp
 2) 평가자수(number of raters) 입력
 3) 피험자수(number of cases) 입력
 4) 각 피험자에 대한 각 평가자의 성적 입력
 5) ICC 버턴 누름
 6) 결과에서 model 2에 제시된 결과를 바탕으로 평가자 개인 신뢰도 계수(icc2) 평가자 전체 평균신뢰도 계수(icc2k) 산출*/

Reference: Portney L.G., Watkins M.P. Foundations of Clinical Research. Applications and Practice (1993) Appleton & Lange, Norwalk, Conneticut ISBN 0-8385-1065-5 p. 509-516

Explanation: Intraclass correlation evaluates the level of agreement between raters in measurements, where the measurements are parametric or at least interval. This method is better than ordinary correlation as more than 2 raters can be included, and there is a correction for correlations between raters that becomes apparent when the range of measurement is large.   The coefficient represents concordance, where 1 is perfect agreement and 0 is no agreement at all. In the analysis of variance, F value for between raters test whether the raters significantly differ in their assessment.

Three models are available.

Model 1 assumes that the raters rating different subjects are different, being subsets of a larger set of raters, randomly chosen.

Model 2 assumes the same raters rate all cases, and the raters are a subset of a larger set of raters.

Model 3 makes no assumptions about the raters.

Model 1 is rarely used, and model 3 cannot be generalized.   Model 2 is therefore the usual one used.

Each model has two versions of the intraclass correlation coefficient:

Single measure reliability: individual ratings constitute the unit of analysis. That is, single measure reliability gives The the reliability for a single judge's rating.

Average measure reliability: the mean of all ratings is the unit of analysis. That is, average measure reliability gives the reliability of the average rating assuming that there is a reasonable number of raters.

If each rating score is from an individual rating, then the single or individual form is used.   However if the rating is the mean of multiple ratings (for example, each rater is a team and the rating is the average of the team), then the mean form should be used.

In other words, Model 2 individual form is the usual one to be used.

Example: Let say we wish to measure the length of the fetus in-utero, using either x-ray (col 1), ultrasound (col 2), or magnetic resonance (col 3).   We want to know how much they all agree with each other. We measure 4 babies and obtain the following results.   the 4 rows are 1.1, 1.2, 1.5; 2.2, 2.1, 2.0; 6.3, 6.1, 6.8; 9.4, 9.5, 9.0.   ICC Coefficient=0.9956, analysis of variance F=0.13.   There is no significant difference between raters, and the level of concordance is high.

<참고 문헌> Shrout, P.E. & Fleiss, J.L. (1979) Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 2, 420-428.*