Paper ###-2019

Measuring Test-Retest Reliability: The Intraclass Kappa

Dennis G. Fisher, Grace L. Reynolds, California State University, Long Beach
Eric Neri, Art Noda, Helena Chmura Kraemer, Stanford University

ABSTRACT

Anyone using structured interview, or questionnaire instruments must establish the psychometric properties of their instrument (i.e. reliability and validity). The first property which must be established is reliability, because one cannot have a valid measure unless the measure has sufficient reliability. When the response data are dichotomous (Yes/No, Presence/Absence, Positive/Negative etc.) the most common measure in the literature is Cohen's kappa (Cohen, 1960). This measure is appropriate for interrater reliability in which the responses from two different raters are assessed for agreement. However, many reliability studies have data from the same rater at two different points in time. This is known as intrarater reliability or test-retest reliability. Cohen's kappa "forgives" rater bias which is not desirable for a measure that is used in test-retest reliability assessment. The correct statistic to use is the Intraclass kappa (Kraemer, Periyakoil, & Noda, 2002). We present a SAS macro which uses a bootstrap procedure to obtain both the point value and the confidence limits of the intraclass kappa so that applied researchers reporting test-retest reliability will be able to report the correct statistic.

INTRODUCTION

Anyone using structured interview or questionnaire instruments must establish the psychometric properties of their instrument. The basic psychometric properties that must be established are reliability which is concerned with the extent to which any measuring device yields that same results on repeated trials (Carmines & Zeller, 1979), or the degree to which the scores are free from measurement error (Price, 2017). Validity is the extent to which the device measures what it is supposed to measure (Price, 2017). The first property that must be established is reliability which is a necessary, but not sufficient, condition to establish validity (Price, 2017). In most cases in which psychometric properties are discussed, the data are measured on a continuous, or at least on a Likert-type scale. The usual measure of reliability in those instances has been the Pearson or Spearman correlation, with a more recent trend to using an intraclass correlation. However, there are many situations in which the data are dichotomous such as yes/no, presence/absence, ever/never, positive/negative, disease/not disease etc. In those situations most of the applied literature has reported a Cohen's kappa (Cohen, 1960).

Cohen's kappa has been reported as the amount of agreement between either two raters or between the same rater at two time points, corrected for the amount of agreement from chance alone. However, just as in the case of the intraclass correlation replacing the Pearson correlation, the intraclass kappa has been advocated to be used in some situations instead of the Cohen's kappa (Kraemer et al., 2002). The major advantage of the intraclass kappa over the original has to do with how it treats bias.

The basic formula for kappa is the following:

Kappa = (observed agreement – chance agreement) / (1 – chance agreement).

BIAS

Bias, in the case of interrater reliability as measured by kappa, is the extent to which the two raters determine that the cases are positive, for example, if we have two clinicians who are determining diagnoses and one clinician decides that 75% of the cases have the disease and the other clinician decides that only 50% of the same cases have the disease, then there is bias in that either the first clinician is over-diagnosing or the second clinician is under-diagnosing the cases. This can be extended to the case of intrarater reliability or test-retest reliability. This is reflected in the marginals of a two-by-two table (Sim & Wright, 2005). If the row marginals are the same as the column marginals, then there is no bias. It is unusual in applied research to have absolutely no bias. Given that most applied research has some bias in the measurement, how the bias is treated by the coefficient is important.

The chance agreement for Cohen's kappa also termed the Percentage Agreement Corrected for Chance (PACC) is PQ + P'Q' where P is the proportion positive for time 1 and Q is the proportion positive for time 2, or P' = 1 – P, Q' = 1 -Q. This is considered to be fairly "forgiving" which is not desirable for a measure of test-retest reliability. The PACC for the intraclass kappa is $P^2 + P'^2$. This has been shown to equal σ_p^2 / PP' (Kraemer, 1979). The intraclass kappa counts bias as error which is more appropriate for a reliability measure. Many reliability studies are test-retest reliability studies or intrarater reliability studies. In this type of study (i.e. test-retest or intrarater reliability) the correct statistic to use is the intraclass kappa (Kraemer et al., 2002). The problem for the applied researcher is that the intraclass kappa has not been accessible in the same way that Cohen's kappa has been. If a SAS analyst wanted to obtain Cohen's kappa, then they could do that using:

```
proc freq;
  tables var1 * var2 / agree;
run;
```

However, intraclass kappa has not been available in this fashion. To remedy this, we present a SAS macro that uses a bootstrap procedure to obtain both the point value of kappa, which is the median of the bootstrap replications, and the upper (97.5 percentile) and the lower (2.5 percentile) confidence limits

SOURCE OF DATA USED IN EXAMPLE

The original data come from the National Institute on Drug Abuse (NIDA) Cooperative Agreement for AIDS Community-Base Outreach/Intervention program (NIDA CA). This was a test-retest reliability substudy that attempted to recruit 20 injection drug users (IDU) at each of the following sites: Anchorage, Alaska; Denver, Colorado; Detroit, Michigan; Houston, Texas; Long Beach, California; Miami, Florida; New York, New York; Philadelphia, Pennsylvania; Portland, Oregon; San Francisco, California; and Tucson, Arizona. Each site attempted to recruit 20 IDU at two points in time, 48-hours apart (Dowling-Guyer, Johnson, Fisher, & Needle, 1994). 48 hours was used because it is how long urine tests for illicit drugs are valid for. The questions come from the instrument that was used in the NIDA CA, termed the Risk Behavior Assessment (RBA). In this section of the RBA there were the following:

"Now, I'm going to ask you some questions about health-related conditions and your use of health care services." "Have you been told by a doctor or a nurse that you had _____?' The diseases listed were: Hepatitis B, Gonorrhea (GC, clap, dose), Syphilis (Syph), Genital Warts (HPV-human papilloma virus), *Chlamydia* (nongonococcal urethritis (NGU)), Genital herpes (herpes). The IDUs were also asked "Have you ever been told that you were infected with the AIDS virus (HIV)?" When asked about the last time they were tested for HIV, the RBA also asked: "Did you get your test results the last time?" The example data in the macro are for *Chlamydia*. (Data have been truncated to only show the first seven observations).

```
/* ------------------------------------------------------------------
   | SUMMARY: this macro uses data for M=2 raters or time points and
   K=2 categories and calculates the intraclass kappa and 95% bootstrap
   confidence intervals
   |
   | INDSN    = input dataset
   | RVALUE1  = the value of the categories,
   | RVALUE2    for example RVALUE1=1 and RVALUE2=2
   | NBOOTS   = number of bootstrap replications (e.g. 1000)
   | BOOTSEED = a seed for generating random bootstrap replications
   | R1       = the variable name for the first rater or time point
   (e.g. anyct)
   | R2       = the variable name for the second rater or time point
   (e.g. anyct2)
   |
```

```
| FORMATTING INPUT DATA
|   There must be one row per subject and one column must represent
the first rater or time and another column must represent the second
rater or time
|
|     Example:
|   ID     rater1     rater2
|   15        1          1
|   22        1          2
|   43        2          2
|
|  NOTE:  Remove any data where rater1 or rater2 is missing
|         before running the SAS macro below
*/


options mprint ls=120 ps=40 nocenter pageno=1;

*  Read in Raw data *;
Data Kraemer ;
Input presid  anyct  anyct2 ;
Label presid = 'Identification Number'
      anyct  = 'Time 1 Ever diagnosed with Chlamydia trachomatis'
      anyct2 = 'Time 2 Ever diagnosed with Chlamydia trachomatis' ;
cards ;
101234 1 1
101235 1 1
101236 1 1
101237 2 2
101238 1 1
101240 1 2
101242 1 1
;
run;

* Remove cases where anyct or anyct2 is missing;
data kraemer;
set kraemer;
if anyct=. or anyct2=. then delete;
run;

* SAS Macro to calculate the intraclass kappa and 95% bootstrap
confidence intervals;
%macro
Intraclass_Kappa_K(INDSN,RVALUE1,RVALUE2,NBOOTS,BOOTSEED,R1,R2);

title1 "Input Data";
proc freq data = &INDSN;
      tables &R1 * &R2/norow nocol nopercent;
run;

* create bootstrap replicates *;
proc surveyselect data = &INDSN
      out = _bootsample seed = &BOOTSEED method = urs
```

```
      samprate = 1 outhits rep = &NBOOTS noprint;
run;

* get kappa from proc freq for sample, suppress output *;
      ods exclude all;
      proc freq data = /* _sub */ &indsn;
            tables &R1 * &R2 /agree outpct;
            ods output CrossTabFreqs =_ctq;
      run;
      ods exclude none;

* calculate intraclass kappa *;
      data _sampkap(keep=po pacc pe ikc); set _ctq end=EOF;
            retain p11 p22 p1dot pdot1 .;
            if      &R1=1 and &R2=1 then p11=percent;
            else if &R1=2    and &R2=2    then p22=percent;
            else if &R1=1 and &R2=.     then p1dot=percent;
            else if &R1=.     and &R2=1 then pdot1=percent;
            if EOF then do;
                  po     = (p11 + p22)/100;
                  pacc = ((p1dot + pdot1)/2)/100;
                  pe   = pacc**2 + (1-pacc)**2;
                  ikc  = (po - pe)/(1 - pe);
                  output;
            end;
      run;

* use freq to get kappas for all replicates *;
      ods exclude all;
      proc freq data = _bootsample;
            by replicate;
            tables &R1 * &R2 / agree outpct;
      ods output CrossTabFreqs = _bctq KappaStatistics = _bks;
      run;
      ods exclude none;

* median of bootstrap simple kappa and 2.5th and 97.5th percentile
for quantile CI *;
      proc univariate data=_bks noprint;
            var value;
            output out=_bootkap median=kapmed PCTLPTS=2.5 97.5
            PCTLPRE=kapmed;
      run;

* calculate intraclass kappa for all bootstrap replicates *;
      data _ikc(keep=p11 p22 p1dot pdot1 po pacc pe ikc);
      set _bctq;
      by replicate;
            retain p11 p22 p1dot pdot1 .;
            if first.replicate then do;
                  p11=.; p22=.; p1dot=.; pdot1=.;
            end;
            if      &R1=1 and &R2=1 then p11  =percent;
            else if &R1=2 and &R2=2 then p22  =percent;
```

```
            else if &R1=1 and &R2=. then p1dot=percent;
            else if &R1=. and &R2=1 then pdot1=percent;
            if last.replicate then do;
                  po     = (p11 + p22)/100;
                  pacc = ((p1dot + pdot1)/2)/100;
                  pe   = pacc**2 + (1-pacc)**2;
                  ikc  = (po - pe)/(1 - pe);
            output;
            end;
            run;

* get median of intraclass kappa and 2.5th and 97.5th percentile for
quantile CI *;
      proc univariate data=_ikc noprint;
            var ikc;
            output out = _bootikc median = ikcmed PCTLPTS=2.5 97.5
            PCTLPRE = ikcmed;
      run;

* combine results on one line *;
      proc sql;
      create table _interleave as
      select input("1",1.0) as rater1,input("2",1.0) as
      rater2,"&NBOOTS" as reps, ikc,ikcmed2_5,ikcmed97_5  from
      _sampkap,_bootikc,_bootkap;
      quit;

* there will be one line of output *;
      proc append base=_results data=_interleave force nowarn; run;
quit;
      run;

* print the single line of results *;
      title1 "Kappa (K=2)";
      proc print data = _results noobs label;
            format rater1 rater2 2.0
            ikc ikcmed2_5 ikcmed97_5 f6.3;
            label
            rater1="Rater1/ Time1"
            rater2="Rater2/ Time2"
            reps="Bootstrap replications"
            ikc = "Intraclass Kappa"
            ikcmed2_5 = 'Intraclass Kappa 2.5th %tile'
            ikcmed97_5 = 'Intraclass Kappa 97.5th %tile';
      run;

* delete temporary working datasets *;
      proc datasets library=work nodetails nolist nowarn;
      delete _results _bootsample _sub _ctq _ks _sampkap _bootsub
            _bctq _bks _bootkap _ikc _bootikc _interleave;
      run; quit;
      %mend;

* run macro *;
```

```
%Intraclass_Kappa_K(INDSN=kraemer,RVALUE1=1,RVALUE2=2,NBOOTS=1000,BO
OTSEED=12345,R1=anyct,R2=anyct2);
```

RESULTS FROM RUNNING MACRO ON SOURCE DATA

The following table shows the results of running both PROC FREQ; Tables / Agree; on the original data and the intraclass kappa macro. As you can see from the table, the values of Cohen's kappa are the same as intraclass kappa when the values are fairly high. The confidence intervals are different. However, when the value of Cohen's kappa is low as in the "Get result" entry in the table on the last line, then the value of the intraclass kappa is different. This is because of the different methods that the two statistics use to handle bias.

Infection	Cohen's Kappa	LCL	UCL	Intraclass kappa	LCL	UCL
Hepatitis B	.8524	.7715	.9333	.852	.765	.920
Gonorrhea	.8581	.7863	.9299	.858	.779	.922
Syphilis	.7771	.6370	.9172	.777	.616	.906
Chlamydia	.8288	.5960	1.000	.829	.495	1.000
Genital Warts	.9450	.8376	1.000	.945	.793	1.000
Genital Herpes	1.000	1.000	1.000	1.000	1.000	1.000
HIV	.8041	.6177	.9906	.804	.572	.959
Get result?	.2250	.1185	.3315	.106	-.064	.267

TABLE 1. TEST-RETEST RELIABILITY FOR SEXUALLY TRANSMITTED INFECTIONS "HAVE YOU BEEN TOLD BY A DOCTOR OR A NURSE THAT YOU HAD _____?"

Time 1	No	Yes	
No	214	0	214
Yes	0	5	5
	214	5	219

TABLE 2. TABLE OF HERPES * HERPES2

Table 2 shows the two-way table for genital herpes which table 1 shows has both perfect Cohen's kappa and intraclass kappa. However, kappa should not be computed unless there are at least 20 for each marginal. Here we only have 5 for the 2. And .2 marginals.

WHAT IS A GOOD VALUE OF KAPPA?

Even though a rule of thumb is an arbitrary opinion, there have been several proposed in the literature. The usual rule of thumb for how to judge values of kappa (Landis & Koch, 1977) are as follows:

Kappa Statistic	Strength of Agreement
<.00	Poor
.00 - .20	Slight
.21 - .40	Fair

Kappa Statistic	Strength of Agreement
.41 - .60	Moderate
.61 - .80	Substantial
.81 – 1.00	Almost Perfect

TABLE 3. RULE OF THUMB FOR STRENGTH OF AGREEMENT (LANDIS & KOCH, 1977)

Another set of values (Fleiss, Levin, & Paik, 2003) has also been given as follows:

Kappa Statistic	Strength of Agreement
<.40	Poor
.41 - .75	Fair to Good
>.76	Excellent

TABLE 4. RULE OF THUMB FOR STRENGTH OF AGREEMENT (FLEISS ET AL., 2003)

CONCLUSION

This paper has introduced a macro for calculating intraclass kappa which is the statistic that should be reported for test-retest reliability studies instead of Cohen's kappa. The reason why the intraclass kappa is preferred for this application is because it handles bias in a more appropriate fashion than Cohen's kappa. Now that the intraclass kappa has been made accessible by this macro, there is no reason that applied researchers who are reporting test-retest reliability measures for dichotomous responses should be using the incorrect statistic.

REFERENCES

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. 07-017). Beverly Hills, CA: Sage Publications.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46. doi:10.1177/001316446002000104

Dowling-Guyer, S., Johnson, M. E., Fisher, D. G., & Needle, R. (1994). Reliability of drug users' self-reported HIV risk behaviors and validity of self-reported recent drug use. *Assessment, 1*(4), 383-392.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Kraemer, H. C. (1979). Ramifications of a population model for !k as a coefficient of reliability. *Psychometrika, 44*(4), 461-472. doi:10.1007/BF02296208

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine, 21*, 2109-2129.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Price, L. R. (2017). *Psychometric methods: Theory into practice*. New York: The Guilford Press.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257-268.

ACKNOWLEDGMENTS

The project described was supported in part by Award Numbers R01DA030234 from the National Institute on Drug Abuse (NIDA), P20MD003942 from the National Institute of Minority Health and Health Disparities (NIMHD), ID10-CSULB-008 from the California HIV Research Program (CHRP), funding from the Sierra-Pacific Mental Illness Research, Education, and Clinical Center (MIRECC) and the VA. The content is solely the responsibility of the authors and does not necessarily represent the official view of the NIDA, NIMHD, CHRP, MIRECC, or the VA. The NIDA, NIMHD, CHRP, MIRECC, and VA had no role

in the study design, collection, analysis, programming, or interpretation of the data, writing of the manuscript, or the decision to submit the paper to this conference.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Dennis G. Fisher
California State University, Long Beach
Dennis.Fisher@csulb.edu

1.