Establishing Interrater Agreement for Scoring Criterion-based Assessments: Application for the MEES

Mark Hogrebe, Washington University in St. Louis October 2018

The purpose of this document is to give guidance on how to establish interrater agreement when scoring performance using a criterion-based assessment that incorporates a detailed rubric for each assessment category. It describes the components required for determining how closely two or more raters score performance using this type of assessment rubric. Examples of how to calculate interrater agreement using MEES-type data are provided along with a free calculator. Finally, a procedure for how interrater agreement data could be collected, summarized, and presented for a MEES technical manual is presented.

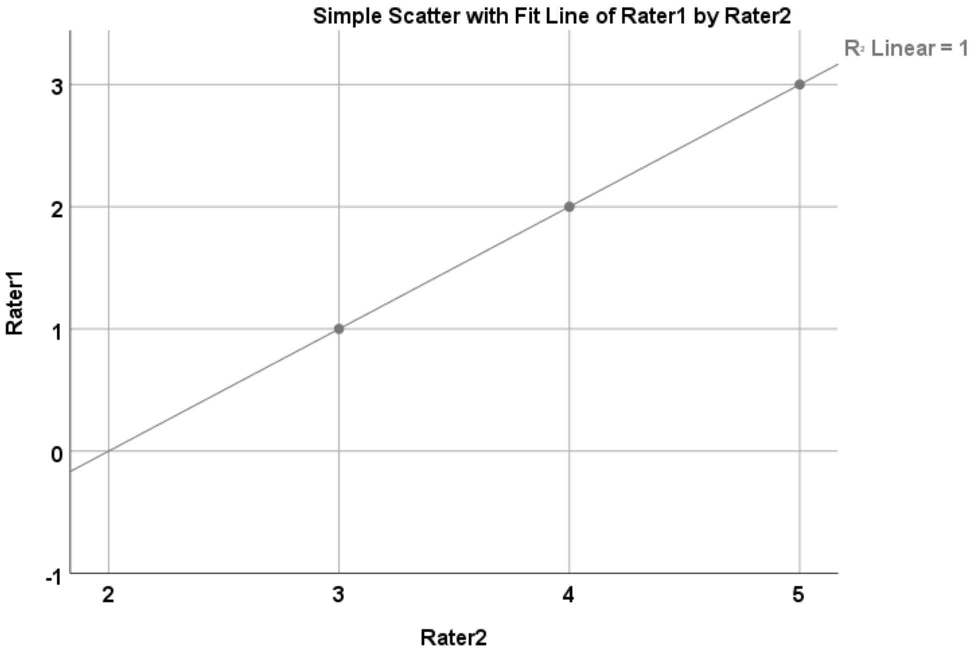
Difference between Interrater Agreement and Interrater Reliability

First, it is important to understand the difference between interrater agreement and interrater reliability (Bandalos, 2018; LeBreton & Senter, 2008; McGraw & Wong, 1996). Interrater ***agreement*** emphasizes the extent to which two or more raters give the exact same rating using the rubric (e.g., Rater1 scores 3 and Rater2 scores 3). In the case where the rubric is intended to be a criterion-based assessment such as the MEES observation assessment (MEES - Missouri Educator Evaluation System), *absolute* interrater agreement is the goal. The raters should give the exact same rubric score when rating the same performance for an observation or response.

In contrast, interrater ***reliability*** refers to the consistency of the rank order of the scores between raters. For example, Rater1 could give scores of 1, 2, 3, while Rater2 scores 3, 4, and 5. In this case the interrater *agreement* is 0.00, but the interrater *reliability* using a Pearson correlation coefficient is 1.00. Clearly, when using a criterion-based rubric, a measure of interrater reliability would not be appropriate.

Perfect Interrater Reliability Equal to 1.00, but Interrater Agreement = 0

Rater1	Rater2
1	3
2	4
3	5



What is an acceptable magnitude for interrater agreement coefficients?

Although there is no ultimate standard for the magnitude for interrater agreement coefficients, the following guidelines for reliability coefficients are widely recognized by the assessment and testing professional community (Reynolds, Livingston, & Willson, 2006, p. 103)

If an instrument is being used to make important decisions that significantly impact individuals and are not easily reversed, it is reasonable to expect reliability coefficients of 0.90 or higher.

This high degree of importance certainly exists with the MEES observations for both teacher education candidates and teacher education programs. Reliable MEES observations are critical for candidates who depend upon accurate ratings in attaining certification. Also, MEES ratings are weighted highly in the Annual Performance Report (APR) for teacher education programs. For both the teacher candidates and teacher education programs, the MEES observation instrument is clearly a high stakes assessment that must demonstrate a high degree of interrater agreement aligned to the target ratings.

Interrater agreement coefficients of 0.85 to 0.89 for a few individual items may be acceptable, but the interrater agreement for all items combined should be 0.90 or above for a high stakes assessment such as the MEES.

Requirements and Procedures for Establishing Absolute Agreement Using a Criterion-based Rubric

Selecting the Target Videos to Rate in Establishing Interrater Agreement

In order to establish interrater agreement, the raters must score more than one target video. It is critical that *multiple* target videos be scored that represent the *range* of possible values in the rubric. For example, if the rubric has five potential scores for a "standard" then the raters have to score target videos representing each of the five points in the rubric that represent each candidate level. The raters should have to score all of the levels that they will potentially observe. The raters' interrater agreement would then be calculated using the 10 scores (5 target ratings and 5 rater scores).

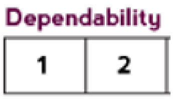
Example interrater agreement task for MEES Standard 1 (content knowledge aligned with appropriate instruction).

Rate five target videos, representing each scoring category in the rubric, i.e., each candidate level:

Example videos for MEES Standard 1	**Targets Ratings**	**Rater**
0 – standard not present	0	0
1 – emerging candidate	1	2
2 – developing candidate	2	2
3 – skilled candidate	3	4
4 – exceeding candidate	4	3

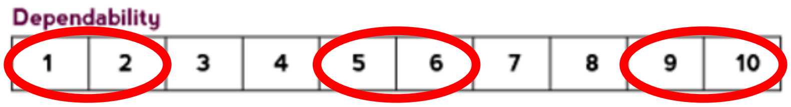
There are several reasons why a range of target videos should be used for establishing interrater agreement in the rating process. First, it is important that the raters view each type of target video that they are expected to encounter and score. This process will help demonstrate the validity of the categories in terms of being distinct and allow the rater to correctly identify the behavior/performance being rated. Second, if only a small subset of categories are used in the interrater agreement calculations, then the strength of agreement will be limited by the restricted range. It is critical that the range of target performances are represented.

Restriction of range example



1 = Not dependable
2 = Dependable

Difference of one unit results in completely opposite ratings.



1 = Not dependable 5 = Sometimes 10 = Always

Difference of one unit results in very similar ratings

Validated Target Ratings

Interrater agreement should be calculated between each rater and the target ratings. Use of the *validated* target ratings as "Rater 1" insures that the agreement calculation is based on accurate representations of the five categories. Need to be sure that raters associate observed behavior to the correct categories. Two raters can have high agreement, but their ratings may be inaccurate in assigning the correct score if the ratings do not correspond to validated target scores. Raters can "miss" the target.

Target scores can be validated as to their accuracy in representing the behavior in the videos by the following procedure:

- For each standard, teacher educator experts rate five videos that presumably represent the five rubric categories.
- Compute interrater agreement among the experts. Review ratings and repeat until there is perfect agreement about the five video rating scores among the experts.
- If perfect agreement cannot be attained among the experts, then evaluate whether one or more of the videos need to be replaced.
- These final ratings become the validated target ratings used in computing all other interrater agreement exercises.

Determining Interrater Agreement at a MEES Training Session

During the MEES training session, several of the key standards would be the focus (e.g., standards 1, 2, 3, and 5). After training and discussing how to rate standard 1, the workgroup would watch a video and then each member assign a rating. Discussion and calibration follows the ratings.

Suggested practice during training sessions

After training and calibration are finished, an estimate of interrater agreement should be attained by having the trainees watch five short validated videos, each representing the five levels on the rubric. The trainees rate each of the five videos for standard 1.

Next compute the interrater agreement on standard 1 for the group members' ratings with target ratings.

> Pair each group member's ratings with the five video target ratings. For 10 trainees there would be 10 pairings of 5 ratings (10 trainee ratings vs target ratings).

The interrater agreement calculation is demonstrated below beginning on page 5. After calculating the interrater agreements, the results can be viewed and reasons for low agreements discussed and explored.

How to Calculate Interrater Agreement

When the rating scale is ordinal, the calculation for interrater agreement should be able to account for intervals between categories that cannot be assumed to be equidistant (e.g., MEES). A versatile interrater agreement measure designed to deal with ordinal data is Krippendorff's alpha (Hayes and Krippendorff, 2007; Krippendorff, 2011; Krippendorff, 2004a; Krippendorff, 2004b). The advantages of Krippendorff's alpha are that it can be used for:

- all levels of measurement (nominal, ordinal, interval, ratio)
- data with missing values
- any number of raters or categories

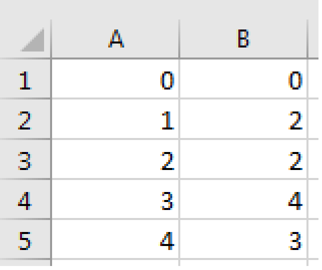
Krippendorff's alpha can be calculated using SPSS in conjunction with a macro, in Matlab, and with Stata module krippalpha. However, there is a free, simple-to-use online Krippendorff alpha calculator called "ReCal" that produces the same results as these other programs. The online ReCal calculator can be found at this link:

http://dfreelon.org/utils/recalfront/recal-oir/#doc

Here is an example using ReCal to calculate interrater agreement for Krippendorff alpha:

Figure 1

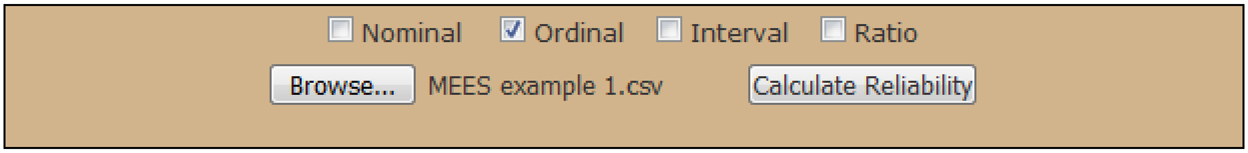
Example videos for MEES Standard 1	Targets Ratings	Rater 1
0 – standard not present	0	0
1 – emerging candidate	1	2
2 – developing candidate	2	2
3 – skilled candidate	3	4
4 – exceeding candidate	4	3



	A	B
1	0	0
2	1	2
3	2	2
4	3	4
5	4	3

Data in Excel spreadsheet saved as a ".csv" file: MEES example 1.cvs

Note there are no column or row headings. Columns represent different raters and the rows represent their ratings/scores for each observation. See Figure 1.
Browse to find the "MEES example 1.cvs" file, then click on the "Calculate Reliability" button.

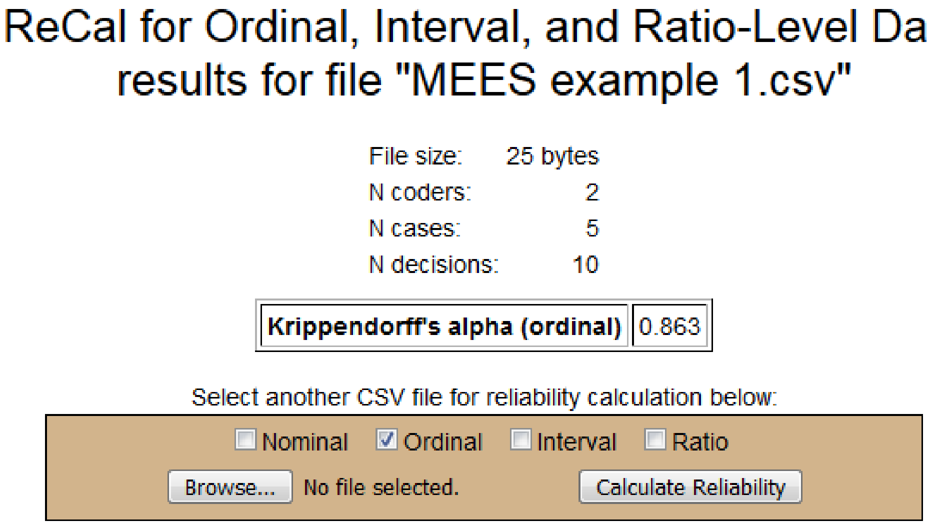


☐ Nominal ☑ Ordinal ☐ Interval ☐ Ratio

MEES example 1.csv

The results are displayed immediately as follows on the next page:

(Notes: For interval level data, the Intraclass Correlation Coefficient (ICC) is sometimes used. See Appendix A.
Also Krippendorff alpha is more versatile than the kappa coefficient or weighted kappa coefficient.)



ReCal for Ordinal, Interval, and Ratio-Level Data
results for file "MEES example 1.csv"

File size:	25 bytes
N coders:	2
N cases:	5
N decisions:	10

Krippendorff's alpha (ordinal)	0.863

Select another CSV file for reliability calculation below:

□ Nominal ☑ Ordinal □ Interval □ Ratio

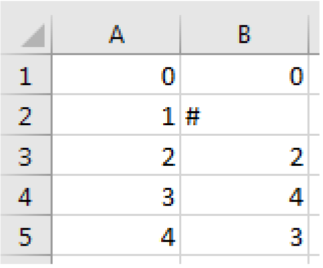
No file selected.

The interrater agreement as calculated by Krippendorff's alpha for Rater 1 against the target ratings is 0.86

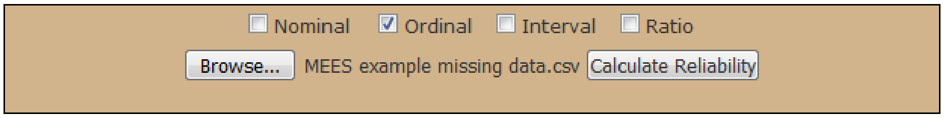
Example with Missing Data

ReCal for Krippendorff alpha can handle a file with missing data.

Data in Excel spreadsheet saved as a ".csv" file: MEES example missing data.cvs
Missing data is coded with a "#"

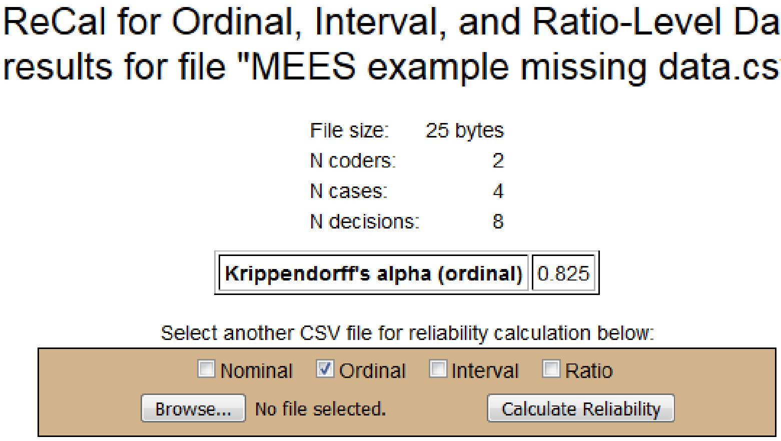


	A	B
1	0	0
2	1	#
3	2	2
4	3	4
5	4	3



□ Nominal ☑ Ordinal □ Interval □ Ratio

MEES example missing data.csv



ReCal for Ordinal, Interval, and Ratio-Level Data
results for file "MEES example missing data.csv"

File size:	25 bytes
N coders:	2
N cases:	4
N decisions:	8

Krippendorff's alpha (ordinal)	0.825

Select another CSV file for reliability calculation below:

□ Nominal ☑ Ordinal □ Interval □ Ratio

No file selected.

The interrater agreement as calculated by Krippendorff's alpha for Rater 1 with one missing rating against the target ratings drops to 0.825

Establishing Interrater Agreement for the MEES as a High Stakes Assessment for the APR

To conduct a large scale interrater agreement study for the MEES, five target videos representing the five rubric levels could be posted online for each teacher candidate standard. Minimally, a combination of 20 cooperating teachers and 20 program supervisors could be recruited to view and rate the five videos for each standard. This would generate 40 pairs on which to calculate the interrater agreement coefficient for each standard (40 raters paired with the target ratings). The mean, standard deviation, standard error, and confidence interval for the 40 interrater agreement coefficients would then be calculated and displayed in a histogram. These statistics and histogram would show the variation in the coefficients. The mean and variation give an estimate of the population or "true" interrater agreement.

Example videos for MEES Standard 1	**Targets Ratings**	**Raters 1 thru 40**
0 – standard not present	0	0
1 – emerging candidate	1	2
2 – developing candidate	2	2
3 – skilled candidate	3	4
4 – exceeding candidate	4	3

Ideally, the following process could be completed for each of the nine standards. The interrater agreement study for each of nine standards would not need to be completed at the same time but could be completed in stages. The viewing and rating of the videos could be conducted online to access a greater pool of cooperating teachers and program supervisors. Some additional research would be needed to verify the appropriateness of having raters score more than one standard for a target video.

Process for establishing interrater agreement for each standard:

Example for teacher education candidate Standard 1

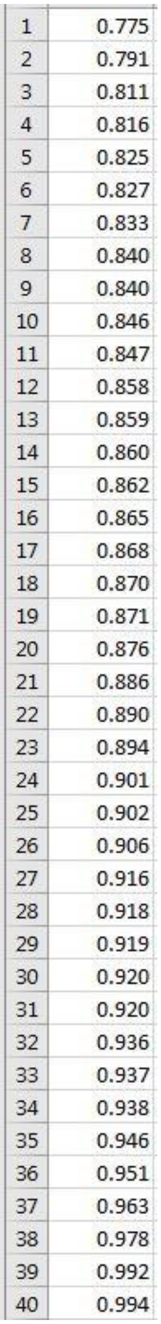
- 20 cooperating teachers and 20 program supervisors rate 5 videos representing 5 levels of the rubric
- Calculate interrater agreement on the 40 pairs of ratings between the target ratings and raters
- Calculate mean, standard deviation, standard error of the mean, and confidence interval for the 40 interrater agreement coefficients
- Display in a histogram

An example of this process and presentation is found on page 8.

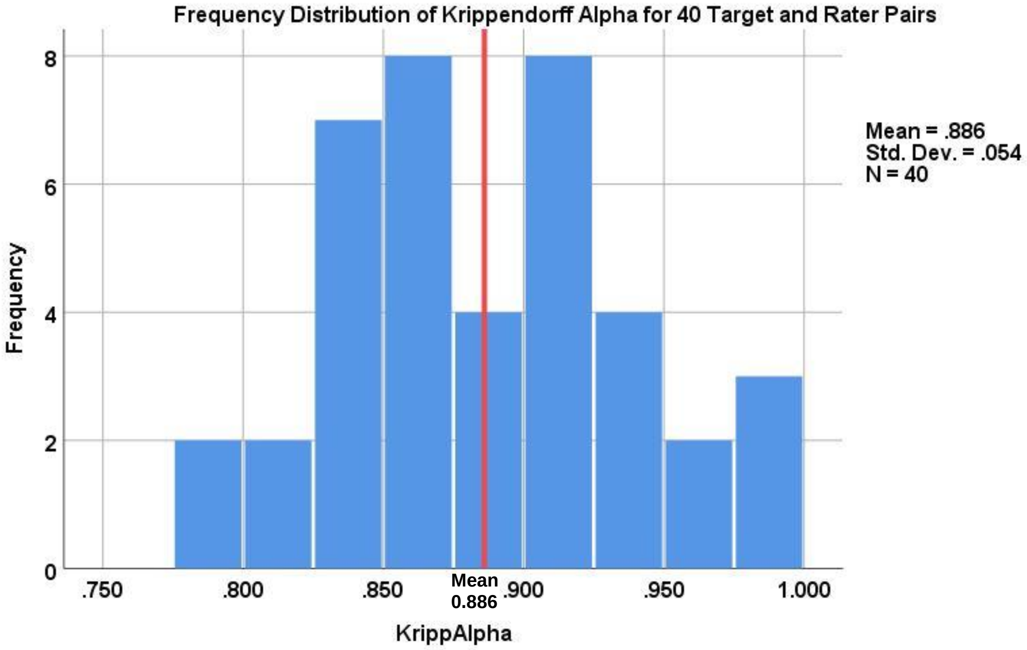
Example Procedure for Determining Interrater Agreement for Standard 1 that Displays the Interrater Agreement Coefficients for 40 Target and Rater Pairs

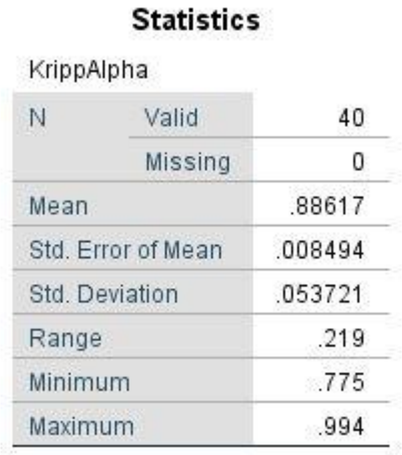
This process could be used for presenting interrater agreement in a technical manual for the MEES.

Randomly generated Krippendorff alpha coefficients for 40 target and rater pairs scoring 5 videos.



1	0.775
2	0.791
3	0.811
4	0.816
5	0.825
6	0.827
7	0.833
8	0.840
9	0.840
10	0.846
11	0.847
12	0.858
13	0.859
14	0.860
15	0.862
16	0.865
17	0.868
18	0.870
19	0.871
20	0.876
21	0.886
22	0.890
23	0.894
24	0.901
25	0.902
26	0.906
27	0.916
28	0.918
29	0.919
30	0.920
31	0.920
32	0.936
33	0.937
34	0.938
35	0.946
36	0.951
37	0.963
38	0.978
39	0.992
40	0.994





Statistics

KrippAlpha

N	Valid	40
	Missing	0
Mean		.88617
Std. Error of Mean		.008494
Std. Deviation		.053721
Range		.219
Minimum		.775
Maximum		.994

95% Confidence Interval for Standard 1 Interrater Agreement as Estimated by Krippendorf's alpha

C.I. = Mean +/– (1.96 * std. error of mean)

C.I. = .886 +/– (1.96 * .008494)

C.I. = .886 +/– .0167

C.I. = (.869, .903)

We can be 95% confident that the interval between .869 to .903 contains the population or "true" interrater agreement coefficient (Krippendorff alpha).

There is a 95% chance that the "true" interrater agreement coefficient is contained in this confidence interval (.869, .903).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Relevant Chapters in *Standards for Educational and Psychological Testing:*

Chapter 2	Reliability/Precision and Errors of Measurement
Chapter 12	Educational Testing and Assessment

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Chapter 9: Interrater agreement and reliability. New York: The Guilford Press

Krippendorff, K. (2011). Computing Krippendorff 's Alpha-Reliability. Retrieved from http://repository.upenn.edu/asc_papers/43

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures 1*(1):77-89.

Krippendorff, K. (2004a). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. Human Communication Research, 30 (3), 411-433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2004b). *Content Analysis: An Introduction to Its Methodology. Second Edition*. Thousand Oaks, CA: Sage.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815-852

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46

Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2006). *Measurement and assessment in education*. Pearson/Allyn & Bacon, Boston, p. 103

Appendix A

ICC Reliability Calculator – Intraclass Correlation Coefficient (ICC)

When data is interval or ratio, interrater agreement can be calculated using the intraclass correlation coefficient (ICC). (Note, the ReCal calculator also handles interval/ratio data.)

The ICC can be calculated using free, simple-to-use online software that is described below.

ICC can also be calculated using the Reliability procedure in SPSS.

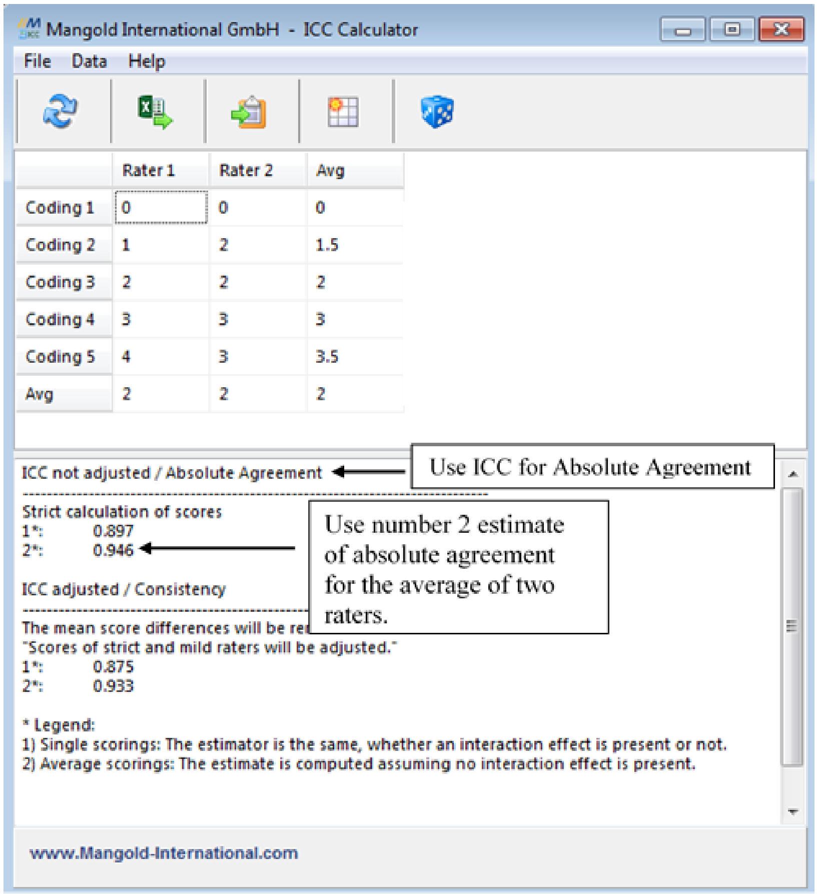
OR

Free software at: http://www.raterreliability.com/
Data entry:

- In Excel, enter scores from Raters 1 and 2
- Copy and paste ratings only into ICC calculator.
  OR
- Click on the "table" icon and enter data directly.

When using reliability ratings for absolute agreement, strict calculation of scores, Number 1 in the results section is the reliability estimate if only one rater will be used to score/record. Number 2 is the reliability estimate when the average of two raters will be used.

When using reliability ratings for consistency, Number 1 in the results section is the reliability estimate if only one rater will be used to score/record. Number 2 is the reliability estimate when the average of two raters will be used. Consistency is used when the rank order of the ratings is important but not the absolute agreement between raters.



	Rater 1	Rater 2	Avg
Coding 1	0	0	0
Coding 2	1	2	1.5
Coding 3	2	2	2
Coding 4	3	3	3
Coding 5	4	3	3.5
Avg	2	2	2

```
ICC not adjusted / Absolute Agreement
------------------------------------------------------------------------------
Strict calculation of scores
1*:        0.897
2*:        0.946

ICC adjusted / Consistency
------------------------------------------------------------------------------
The mean score differences will be re
"Scores of strict and mild raters will be adjusted."
1*:        0.875
2*:        0.933

* Legend:
1) Single scorings: The estimator is the same, whether an interaction effect is present or not.
2) Average scorings: The estimate is computed assuming no interaction effect is present.
```

www.Mangold-International.com

Use ICC for Absolute Agreement

Use number 2 estimate of absolute agreement for the average of two raters.