International Journal of Internet Science
2013, 8 (1), 10-16
ISSN 1662-5544

International Journal of Internet Science

IJIS.NET

ReCal OIR: Ordinal, Interval, and Ratio Intercoder Reliability as a Web Service

Deen Freelon

American University School of Communication, Washington, DC, USA

Abstract: This article introduces ReCal OIR, an extension to the existing ReCal suite of online intercoder reliability modules that accommodates ordinal, interval, and ratio levels of measurement. It includes a discussion of the currently available options for calculating reliability for the ordinal, interval, and ratio levels; explains what ReCal OIR does; validates its calculations; and discusses its usage data. The process of validating its output reveals ReCal OIR to be slightly more accurate than one of its major competitors.

Keywords: Intercoder reliability, interrater reliability, ReCal, web service, content analysis

Introduction

Intercoder or interrater reliability is an essential means of verifying the quality of data produced through subjective observation (Hayes & Krippendorff, 2007; Krippendorff, 2004; Neuendorf, 2002). Formally, the term refers to any technique that evaluates the level of agreement of multiple raters' subjective judgments of particular data points (Freelon, 2010). Such techniques are widely used throughout the social and medical sciences, notably as part of the method of content analysis. Intercoder reliability can be expressed numerically as any of several coefficients, each of which incorporates slightly different assumptions into its formula. These coefficients include simple percent agreement, Scott's pi, Cohen's kappa, and Krippendorff's alpha. Until fairly recently, software to calculate most of these coefficients was in short supply (Freelon, 2010; Hayes & Krippendorff, 2007; Neuendorf, 2002).

One recent addition to the set of intercoder reliability calculators is ReCal (Freelon, 2010), a web-based service which offers a number of advantages over other calculators. Its only system requirement is web access, which means it can be run from nearly any desktop or mobile operating system, unlike statistical packages which support only the most popular OSes. ReCal is also available free of charge, which is especially important for students and scholars in developing countries. Further, most users who have left feedback have found it relatively easy to use in comparison to alternative free and proprietary calculators. Given the number of program executions since its launch (66,234 at the time of this writing), it seems clear that ReCal is continuing to serve an important purpose for its user base.

However, ReCal is not without its limitations. For one, it cannot accept missing data at this time, which requires users to eliminate all empty data cells prior to using the application. More importantly, the original ReCal modules only accept data at the nominal level of measurement. Researchers studying variables comprised of discrete, non-numerical values were well-served by these modules, but those working with ordinal, interval, and/or ratio-level data needed to find alternative solutions. In 2010 ReCal's developer (who is also the author of this

Address correspondence to Deen Freelon, American University School of Communication, 4400 Massachusetts Ave NW, Washington, DC 20016, USA. Freelon@american.edu

paper) addressed this limitation by developing a new version of the tool that accepts ordinal, interval, and ratio data. This new version is called ReCal OIR (ordinal, interval, ratio) and implements the ordinal, interval, and ratio variants of Krippendorff's alpha in the programming language PHP (Krippendorff, 2007). The purpose of this paper is to demonstrate the research purposes ReCal OIR fulfills, list the formulae it implements, verify its output, and report on its usage.

Ordinal, interval, and ratio-level intercoder reliability

The simplest intercoder reliability metric, percent agreement, is a nominal-level coefficient. It is calculated by dividing the number of agreements between two independent coders by the total number of agreements plus disagreements. Though percent agreement does not account for agreement by chance, its mathematical logic is nevertheless valid for variables comprised of unordered discrete categories (i.e., nominal variables). For example, a content analysis task in which two coders are asked to judge whether the target audience of a series of advertisements is men, women, or both would have three possible categorical answers. Percent agreement in this case would express the extent to which the coders agreed in their evaluations of the ads. Mathematically, it treats all differences in evaluations equally: the coefficient is not given any additional "credit" for disagreements between any particular pair(s) of answers among the three candidates. The same is true for reliability coefficients such as Cohen's kappa and Scott's pi that, unlike percent agreement, account for intercoder agreement by chance.

However, coefficients designed for nominal variables do not suffice for variables at other levels of measurement. A widely-used typology of data measurement (Stevens, 1946) divides metric units into four types: nominal, ordinal, interval, and ratio. All of these types except for nominal assume that the proximity between assigned values is meaningful. Ordinal measures assume that adjacent values are more similar than distant values, but that the degree of similarity between values is not necessarily constant. A common example can be found in rank measurements, such as when survey participants are asked to rank their preferences from most- to least-preferred. Such ranked preferences can be arranged on a sequential scale, but they do not represent fixed degrees of distinction – one participant's perceived difference between her first and second choice may differ greatly from another's and the perceived differences between two subsequent options versus two other subsequent options may differ greatly as well. Interval measures, by contrast, do represent fixed degrees of distinction between all possible values. Two examples here would be the Fahrenheit and Celsius scales of temperature – in each case, the difference between any two adjacent integer degree measurements is the same. Ratio measures are formally identical to interval measures with the exception that zero in the former represents a complete absence of the quantity in question. On the Fahrenheit and Celsius scales, 0° does not indicate an absence of temperature, only one degree colder than +1° and one degree warmer than –1°. But the Kelvin scale of temperature is at the ratio level, because zero K (i.e., absolute zero) is defined as the complete absence of temperature. All count data are ratio-level by definition, while other quantities (such as time) may differ in terms of their level of measurement depending on the metric used.

Most of the coefficients developed specifically to measure intercoder reliability are only valid for nominal-level data (Neuendorf, 2002). While many subjective data analysis tasks operate at the nominal level, many do not. For example, medical doctors may use ordinal-level scales to quantify the progression of a particular medical condition. Further, the proliferation of tools for administering online surveys and experiments are making interval and ratio measurement scales easier to apply than ever (Reips & Funke, 2008). Any coding task that requires counting – for example identifying the number of references to the American president in newspaper articles about foreign policy – would need to account for numerically similar judgments. Equations that treat the difference between counts of seven and eight references as equal to that between zero and eight will yield coefficients that are inappropriately conservative. Thus, scholars whose research involves repeated, subjective observations of non-nominal data need the proper reliability coefficients to help validate their data.

To this end, Krippendorff (2007; Hayes & Krippendorff, 2007) performs the immensely helpful service of demonstrating how his eponymous reliability coefficient can be adapted to all four levels of measurement. The result is a suite of four mathematically distinct Krippendorff's alpha formulae, each calibrated to fit the contours of one of the measurement levels. These variants represent some of the only available coefficients appropriate for calculating reliability for non-nominal data (their major alternative is Lin's concordance, which is appropriate for interval and ordinal data and computationally similar to Krippendorff's alpha)[1]. The Krippendorff's alpha variants

[1] Some have suggested using Pearson's product-moment correlation for interval-level intercoder reliability (Hayes & Hatch, 1999), but Krippendorff (2004) argues convincingly against this practice.

also have the advantage of sharing the same underlying assumptions, which facilitates comparisons of coefficients between levels of measurement (Krippendorff, 2004).

At present, options for calculating Krippendorff's alpha are limited in terms of both system requirements and learning curve. Hayes and Krippendorff (2007) offer a pair of scripts that can calculate any of the four alpha variants, but they only work with the proprietary statistical packages SAS and SPSS, respectively. Artstein (2010) offers a free script in the Perl programming language that can calculate the nominal and interval levels of alpha, but it omits the ordinal and ratio variants and requires a Perl interpreter to function (not to mention knowledge of Perl). Gamer, Lemon, Fellows, and Singh (2012) have created a package called *irr* for the open-source statistical platform R that can calculate a wide range of intercoder reliability coefficients, including all four variants of Krippendorff's alpha. However, R's learning curve is substantially higher than those of its commercial competitors and it can be difficult for non-statisticians to use (Williams, 2009). For example, irr's implementation of Krippendorff's alpha cannot analyze coder judgments directly: users must take the non-intuitive step of transforming their datasets into matrices first. Finally, unlike all of these packages, ReCal OIR requires only a graphical web browser to use, meaning that users need not worry about system requirements for downloading and installing software (for more on this point, see Freelon, 2010).

ReCal OIR: what it is and what it does

The ReCal OIR module extends the functionality of the original two nominal-only ReCal modules, ReCal2 and ReCal3 (Freelon, 2010), to ordinal, interval, and ratio-level data types. It can be accessed at the following URL: http://dfreelon.org/utils/recalfront/recal-oir/. Like its sibling modules, it accepts as input CSV (comma-separated values) and TSV (tab-separated values) files, which are non-proprietary formats that can be exported by most spreadsheet, statistical and database applications. Each file must be arranged such that each row corresponds to an individual unit of analysis and each column corresponds to an individual coder's judgments. As with ReCal3, each file represents a single variable; thus ReCal OIR can only compute intercoder reliability for one variable per execution. The program accepts only files containing the judgments of a minimum of two coders; that is, two or more columns of data. All data must consist solely of integers representing variable values, and all cells within each row and column must be filled in. Violating any of these data formatting requirements will result in a program error. Tables 1 and 2 provide examples of data formatted properly for ReCal OIR.

Table 1
ReCal OIR Dataset A

Coder 1	Coder 2
6	6
6	6
7	6
7	6
7	7
7	7
7	7
6	6
6	6
7	7
7	7
6	6
6	6
6	6
6	6
7	7
7	7
6	6
6	6
6	6

Note. Rows indicate units of analysis; columns indicate individual coders.

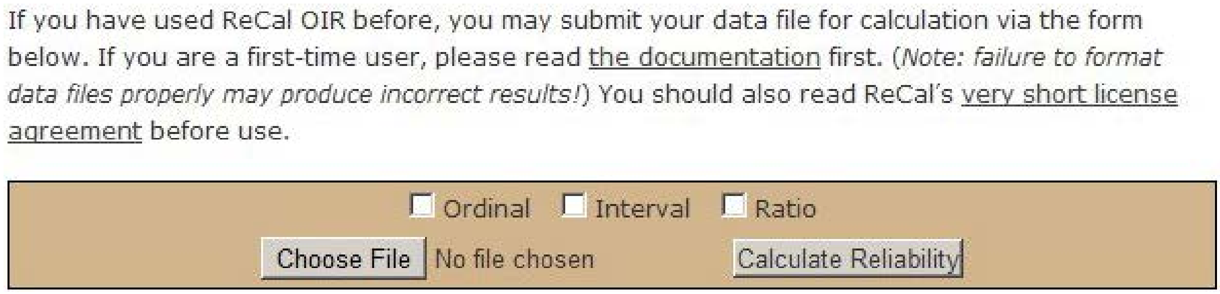
Table 2
ReCal OIR Dataset B

Coder 1	Coder 2	Coder 3
0	0	0
1	2	1
2	2	1
0	1	1
0	0	1
2	2	2
0	0	1
1	1	1
2	2	2
0	1	2
1	0	1
2	1	2
0	0	0
2	1	3
2	2	2
2	1	1
1	1	1
1	1	1
2	1	2
0	0	0

Note. Rows indicate units of analysis; columns indicate individual coders.

Unlike its sibling modules, ReCal OIR does not automatically compute all of the coefficients it offers when it executes. Therefore, before submitting files for analysis, users must decide which coefficient(s) they would like to calculate. This is done by ticking the appropriate checkboxes for ordinal, interval, and/or ratio data in the ReCal

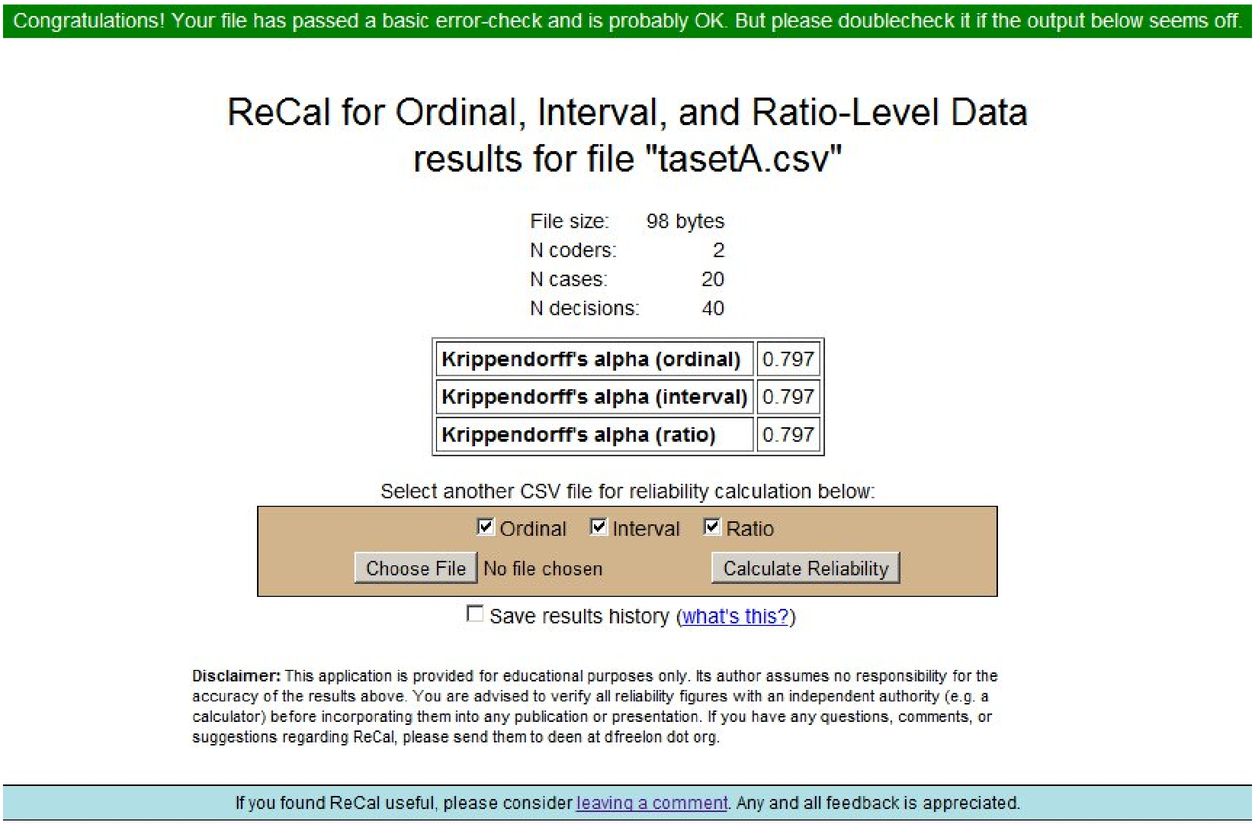
OIR front-end interface (see Figure 1). This design choice reflects the fact that since most variables unambiguously belong to a single level of measurement, only one of the available coefficients will be appropriate for each file. Assuming the file is formatted correctly and does not exceed the file size limit of 100kb, the resulting page will return the coefficient(s) requested (Figure 2). This page can either be saved as HTML for future reference or the user can manually copy the coefficient into another document.



If you have used ReCal OIR before, you may submit your data file for calculation via the form below. If you are a first-time user, please read the documentation first. (*Note: failure to format data files properly may produce incorrect results!*) You should also read ReCal's very short license agreement before use.

Figure 1. Submission interface for ReCal OIR.

Of course, any statistics program is only as valuable as its output is correct. Therefore the following section will present results from test executions of both ReCal OIR and Andrew Hayes' Krippendorff's alpha script for SPSS (Hayes & Krippendorff, 2007). The sample data used in these tests are provided for those readers who want to check the results for themselves.



Congratulations! Your file has passed a basic error-check and is probably OK. But please doublecheck it if the output below seems off.

ReCal for Ordinal, Interval, and Ratio-Level Data
results for file "tasetA.csv"

File size:	98 bytes
N coders:	2
N cases:	20
N decisions:	40

Krippendorff's alpha (ordinal)	0.797
Krippendorff's alpha (interval)	0.797
Krippendorff's alpha (ratio)	0.797

Select another CSV file for reliability calculation below:

☑ Ordinal ☑ Interval ☑ Ratio
Choose File
No file chosen
Calculate Reliability
☐ Save results history (what's this?)

Disclaimer: This application is provided for educational purposes only. Its author assumes no responsibility for the accuracy of the results above. You are advised to verify all reliability figures with an independent authority (e.g. a calculator) before incorporating them into any publication or presentation. If you have any questions, comments, or suggestions regarding ReCal, please send them to deen at dfreelon dot org.

If you found ReCal useful, please consider leaving a comment. Any and all feedback is appreciated.

Figure 2. ReCal OIR output page for Dataset 1.

Validation of output

The basic form of the Krippendorff's alpha equation is:

$$\alpha = 1 - \frac{D_o}{D_e}$$

with D_o being the observed disagreement (hence the subscripted "o") and D_e being the expected disagreement accounting for chance (Krippendorff, 2007). The computational definitions for D_o and D_e differ between the variants of alpha designed for each level of measurement. As Krippendorff (2007) has already provided complete formulae and worked examples for each alpha variant in an open-access article, his efforts will not be duplicated here.

Two files submitted by users to ReCal OIR were randomly chosen to serve as test data in the validation process. This was possible because ReCal saves all data submitted to it (and states as much in its terms of service) and labels each file according to the module used. The two selected files were truncated to their first 20 units of analysis each so that they could be easily reproduced as tables in this article. Dataset 1, in Table 1, consists of two columns (and therefore coders) and contains a total of two expressed values (the numbers 6 and 7). Dataset 2, in Table 2, contains three columns and four expressed values (the numbers 0 through 3). Each dataset was tested using the Hayes macro version 3.1, irr version 0.83, and ReCal OIR.

Working from the same CSV files, both the Hayes macro and ReCal OIR produced equivalent results for both datasets, albeit to differing numbers of significant digits. However, the irr package for R produced slightly different coefficients for Dataset 2. Ordinal, interval, and ratio tests for Dataset 1 all yield the same alpha value of 0.7970 in ReCal OIR, 0.797 in irr, and 0.7969 in the Hayes macro. The results from all three programs for Dataset 2 can be seen in Table 3. All three pairs of ReCal OIR and Hayes coefficients fall within rounding error of one another to three significant digits, while irr's values are a few thousandths different from each pair. Given that the Hayes macro's output has been endorsed by Krippendorff himself (Hayes & Krippendorff, 2007), these results strongly suggest that irr's calculations are lacking in accuracy by a small degree. They further indicate that while the Hayes macro produces slightly more accurate results than ReCal OIR, the underlying math in the two programs is essentially the same.

Table 3
Ordinal, interval, and ratio Krippendorff's alpha coefficients for Dataset B from three calculators

	Alpha Coefficient		
Program	Ordinal	Interval	Ratio
ReCal OIR	0.6310	0.6180	0.5500
Hayes macro	0.6309	0.6177	0.5497
irr	0.6280	0.6140	0.546

A brief note about missing data is warranted here. Like the other ReCal modules, ReCal OIR does not accept data files in which any cells are blank, though the math of Krippendorff's alpha can accommodate such data. This is a clear limitation of ReCal OIR that is not shared by the Hayes macro. Researchers who wish to calculate intercoder reliability statistics for incomplete data using ReCal OIR should conduct manual listwise deletion on their dataset prior to submitting it. That is, they should delete all rows in which at least one cell is missing. While not a perfect solution, this should suffice for data in which only a small percentage of cells are blank. Data sets with substantial amounts of empty cells have broader validity problems which should be addressed at the data collection level if possible.

ReCal OIR usage data

According to its web analytics reports, ReCal OIR has seen consistent usage, though not quite as much as the other ReCal modules. Between its launch on June 23, 2010 and December 4, 2012 (the time of this writing), ReCal OIR has been executed 13,827 times by 2,524 unique visitors.[2] The former number yields an average of 15.4 executions per day. By comparison, ReCal3 was executed 19,953 times and ReCal2 20,905 times during the same time period. ReCal OIR's usage also grew at a faster rate than either of the other modules – in its first three months of availability it attracted 981 executions, whereas ReCal2 garnered only 289 executions and ReCal3 only 190 in their first three months. Overall, however, ReCal's patrons seem to overwhelmingly prefer nominal data, but a fair number of them are clearly interested in the other measurement levels.

Like ReCal2 and 3, ReCal OIR has been executed by users working from a wide range of academic institutions, Internet service providers, government agencies, hospitals, and nonprofit organizations. Site statistics report that 462 unique domains accessed ReCal OIR at least once since its launch. These domains represent 61 countries across all six inhabited continents. The top ten countries by ReCal OIR executions are, respectively: the US, the UK, Cambodia, Canada, Belgium, the Netherlands, Germany, Thailand, Singapore, and South Korea. Some of the less likely suspects on this list are probably there due to heavy usage by one or two particularly active research

[2] These data come from Google Analytics.

groups. If nothing else, the geographic diversity represented in these top countries demonstrates ReCal OIR's international utility.[3]

More specific data on ReCal's users – for example, their disciplines and academic ranks – are limited by my decision not to collect any user information through the web site. This decision was made in the interest of lowering ReCal's barriers to access as much as possible, as it already has a slight learning curve. Thus, although it is quite clear that ReCal is being used by many individuals from around the world, much less can be said about which disciplines are making greatest and least use of it. Some limited and non-representative data on this point come from citations to the paper that originally introduced ReCal (Freelon, 2010), though these are likely skewed by the fact that some disciplines are more likely to cite software packages than others. Table 4 displays the disciplines of the 36 documents that have cited the original paper as of December 4, 2012, in descending order of number of citations.[4] These documents include a combination of peer-reviewed journal articles, conference papers, doctoral dissertations, and master's theses. Communication is the best-represented discipline with eight citations, which is unsurprising given that contemporary content analysis methods were originally developed by communication scholars. Information science, which also makes extensive use of content analysis, is also well-represented with five citations. The remaining disciplines only contain a handful of citations each, but collectively represent a wide range of disciplines that includes the natural sciences, the social sciences, applied fields, and critical theory. Of course, because not every journal article or dissertation that uses ReCal cites it formally, its true scholarly impact is almost certainly much greater than the numbers listed here suggest.

Table 4
Disciplines citing ReCal: Intercoder Reliability Calculation as a Web Service

Discipline	N of citations
Communication	8
Information science	5
Medicine	4
Criminology	3
Education	3
Critical studies	3
Psychology	2
Business	2
Political science	2
National security studies	1
Public health	1
Geography	1
[unknown]	1
Total	36

Note. These data were collected through Google Scholar. The document listed as "[unknown]" was written in Slovak and its discipline could not be ascertained.

User comments on the web site and emails provide further qualitative evidence consistent with the notion that ReCal is serving an important research function. The ReCal OIR front page features a handful of comments from users, some of which are complimentary and others of which are questions about how to use the program. Emails sent to the author reflect a similar mix of sentiments: they typically include both praise and questions relating to specific datasets. Indirect evidence of ReCal OIR's value comes from comments to the main ReCal page that address users' difficulties in getting other intercoder reliability calculators to function as advertised. For example on June 24, 2012, "Cynthia" wrote: "I spent about 6 hours mucking my way through other calculators/SPSS/ Excel trying to get an IRR I could use. 20 minutes here, and I've got the scores I need! Wow, I can't thank you enough!!!"[5] Comments such as these indicate that despite its alternatives, the simple presence of a free, intuitive, web-based reliability calculator made a genuine difference for some users. Extending this general principle to additional levels of measurement opens new research possibilities for an existing user base.

[3] International use of ReCal is almost certainly limited by the fact that it is currently only available in English, which is also true of the other calculators discussed in this article. The author welcomes the volunteer assistance of anyone interested in translating ReCal's instructions and output annotations into other languages.
[4] Intercoder reliability was not assessed for these categories because each document bore its own preexisting category (for example, a study published in the Howard Journal of Communications was counted in the communication category).
[5] This comment can be viewed in its original context at http://dfreelon.org/utils/recalfront/.

Conclusion

This article has introduced, justified, validated, and discussed the usage of a recent update to the ReCal web service for intercoder reliability. While options for nominal intercoder reliability calculation remain limited, those for ordinal, interval, and ratio coefficients are even more so. ReCal OIR will not be an ideal solution for everyone – it cannot process files from which data are missing, and it does not offer confidence intervals as the Hayes macro does. But the advantages demonstrated in this article, including its superiority over irr in accuracy, make it an attractive and convenient option for many, as feedback on its web site attests. Like its sibling modules, ReCal OIR's purpose is primarily to expand the universe of intercoder reliability calculation options, and secondarily to inspire future social science software developers to invest time and energy creating open-access applications for the benefit of their fields.

References

Artstein, R. (2010). *Calculate Krippendorff's alpha, including confidence intervals using bootstrapping.* Retrieved from http://ron.artstein.org/resources/calculate-alpha.perl

Freelon, D. G. (2010). ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, *5*(1), 20–33.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). irr: Various coefficients of interrater reliability and agreement. Retrieved September 11, 2012, from http://cran.r-project.org/web/packages/irr/index.html

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.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability correlation versus percentage of agreement. *Written Communication*, *16*(3), 354–367. doi:10.1177/0741088399016003004

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, *30*(3), 411–433.

Krippendorff, K. (2007). Computing Krippendorff's alpha reliability. *Departmental Papers (ASC)*, 43.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, *40*(3), 699–704. doi:10.3758/BRM.40.3.699

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. doi:10.1126/science.103.2684.677

Williams, G. J. (2009). Rattle: A data mining GUI for R. *The R Journal*, *1*(2), 45–55.