Relationship Between Intraclass Correlation and Percent Rater Agreement

When raters are involved in scoring procedures, inter-rater reliability (IRR) measures are used to establish the reliability of measures. Commonly used IRR measures include percent rater agreement, intraclass correlation coefficients (ICC), and Cohen's Kappa. Several researchers recommend using ICC and Cohen's Kappa over Percent Agreement (Hallgren, 2012; Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979). However, there are misconceptions and inconsistencies when it comes to proper application, interpretation, and reporting of these measures (Kottner et al., 2011; & Trevethan, 2017). Moreover, researchers tend to recommend different thresholds for poor, moderate, and good level of reliability (see Table 2). These inconsistencies, and the paucity of detailed reports of test methods and results, perpetuate the misconceptions in the application and interpretation of IRR measures.

Current recommendations regarding the thresholds of reliability estimates suggest considering purposes and consequences of tests, and the magnitude of error allowed in test interpretation and decision making (Trevethan, 2017; AERA, NCME, & APA, 2014; Kottner at al., 2011). Furthermore, Kottner et al. (2011) also recommend reporting multiple reliability estimates. A low ICC might be due to lack of variability between subjects so by reporting different reliability coefficients (e.g. percent agreement) readers can get a more complete understanding of the degree of reliability.

Research Questions

Given the different types of ICC and guidelines for interpretation, this paper is guided by the following research questions:

1. What is the relationship between ICC and PRA?
2. Are the published guidelines for interpreting ICC appropriate for all rating designs?

Method

Simulations were sued to explore the relationship between intraclass correlation (ICC) and percent rater agreement (PRA). The `IRRsim` R package was developed to facilitate the simulation and analyses of interrater reliability statistics. Here, we will focus on one common design in educational research whereby *n* scoring events are evaluated by two raters from *k* available raters. The following matrix represents the first six scoring events from a total of six available raters where a score ranges between 1 and 3; this matrix was generated using the `simulateRatingMatrix` function.

```
  a b c d e f
1     1   3
2   1 1
3   2       1
4 1         1
5   3       1
6   1       1
```

Parameters are available to generate rating matrices for various designs including parameters for the number of scoring levels, number of available raters, number of scoring events, percent rater agreement, and the distribution (or frequency) of scores. The `simulateICC` function generates multiple rating matrices and calculates IRR statistics for each individual rating matrix. The `summary` and `plot` functions provide an overview the simulated scoring matrices.

Results

Figure 1 represents the results of 200 100 x 6 scoring matrices with three scoring levels. Each point represents one scoring matrix with the corresponding calculated PRA and ICC1. A quadratic regression was fit and is superimposed along with Cicchetti's (2001) guidelines for interpreting ICC. The resulting R^2 for the quadratic model is 97%. Figure 2 is of the same design, but includes the results with 2, 4, 8, and 16 raters chosen two at time; results of all models had $R^2 > 92\%$.

A second analysis was conducted where 158,400 100 x *k* rating matrices were simulated where *k* ranged from 2 through 12; the number of scoring levels ranged from 2 to 5; and four different response distributions were used (i.e. uniform, lightly skewed, moderately skewed, and high skewed). Table 3 provides a summary of the R^2 for quadratic models fit for each *k* and number of scoring levels. These results indicate that PRA accounts for at least 82% of the variance in the ICC statistic.

Discussion

Methodologists have consistently argued that ICC is preferred over PRA for reporting inter-rater reliability (Hallgren, 2012; Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979). Although some recommendations for interpreting ICC have been given (Table 2), the form of ICC (Table 1) those recommendations apply to has not specified by the authors. Furthermore, the nature of the design, especially with regard to the number of possible raters, has substantial impact on the magnitude of ICC (Figure 2). For example, all other things kept equal, increasing the design from 2 to 12 raters changes the required PRA from 61% to 91% to achieve Cicchitti's (2001) "fair" threshold. And with eight or more raters, "good" or "excellent" reliability are not even possible under this design.

We concur with Kottner et al (2011) and Koo and Li (2016) recommendation that the design features along with multiple IRR statistics be reported by researchers. Given the ease of interpretability of PRA, this may be a desirable metric during the rating process. To assist researchers on interpreting ICC in relation to PRA, we have developed an R Shiny application (Figure 3). This application allows researchers to specify their rating design and explore the relationship between various IRR metrics and PRA, superimpose multiple recommendations (Table 2), and predict ICC values from PRA.

References

Altman, D. G. (1990). *Practical statistics for medical research*. London: Chapman & Hall/CRC press.

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.

Brage, M. E., Rockett, M., Vraney, R., Anderson, R., & Toledano, A. (1998). Ankle fracture classification: A comparison of reliability of three x-ray views versus two. *Foot & Ankle International*, *19*(8), 555-562.

Cicchetti, D. V. (2001). Methodological Commentary. The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology*, *23*(5), 695-700.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*(2), 127-137.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York: John Wiley & Sons.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments.* New York: Wiley.

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.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., … & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International Journal of Nursing Studies*, *48*(6), 661-671.

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

Martin, J. S., Marsh, J. L., Bonar, S. K., DeCoster, T. A., Found, E. M., & Brandser, E. A. (1997). Assessment of the AO/ASIF fracture classification for the distal tibia. *Journal of Orthopaedic Trauma*, *11*(7), 477-483.

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

Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research: applications to practice.* Upper Saddle River: Pearson/Prentice Hall.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*(3), 301-317.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428.

Svanholm, H., Starklint, H., Gundersen, H. J. G., Fabricius, J., Barlebo, H., & Olsen, S. (1989). Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. *Apmis*, *97*(7-12), 689-698.

Trevethan, R. (2017). Intraclass correlation coefficients: Clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, *17*(2), 127-143.

Zegers, M., de Bruijne, M.C., Wagner, C., Groenewegen, P.P., van der Wal, G., de Vet, H.C. (2010). The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. *Journal of Clinical Epidemiology, 63*(1), 94–102.

Tables and Figures

Table 1. *Descriptions and formulas of the IRR measures*

	Description	Formula
Percent agreement	Absolute agreement	$\frac{number\ of\ observations\ agreed\ upon}{total\ number\ of\ observations}$
ICC (1, 1)	One-way random effects, absolute agreement, single measures	$\frac{MS_R - MS_W}{MS_R + (k-1)\ MS_W}$
ICC (2, 1)	Two-way random effects, absolute agreement, single measures	$\frac{MS_R - MS_E}{MS_R + (k-1)\ MS_E + \frac{k}{n}(MS_C - MS_E)}$
ICC (3, 1)	Two-way mixed effects, consistency, single measures.	$\frac{MS_R - MS_E}{MS_R + (k-1)\ MS_E}$
ICC (1, *k*)	One-way random effects, absolute agreement, average measures.	$\frac{MS_R - MS_W}{MS_R}$
ICC (2, *k*)	Two-way random effects, absolute agreement, average measures.	$\frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}}$
ICC (3, *k*)	Two-way mixed effects, consistency, average measures.	$\frac{MS_R - MS_E}{MS_R}$
Cohen's Kappa (κ)	Absolute agreement	$\frac{P_o - P_e}{1 - P_e}$

Note. MS_R = mean square for rows; MS_W = mean square for residual sources of variance; MS_E = mean square error; MS_C = mean square for columns; P_o = observed agreement rates; P_e = expected agreement rates.

Table 2. G*uidelines for IRR estimates*

Reference	IRR Metric	Guidelines
Cicchetti & Sparrow (1981) Cicchetti (2001)	ICC & Cohen's Kappa	<.40 poor .40 - .59 fair .60 - .74 good >.75 excellent
Zeger et al. (2010)	Cohen's Kappa	00 - .20 slight .21 - .40 fair .41 - .60 moderate .61 - .80 substantial .81 – 1.00 almost perfect
Fleiss (1981, 1986) Brage et al. (1998) Martin et al. (1997) Svanholm et al. (1989)	Cohen's Kappa	< .40 poor .40 - .75 fair (to good) >.75 excellent (or good)
Altman (1990)		00 - .20 poor .20 - .40 fair .40 - .60 moderate .60 - .80 good >.80 very good
Shrout (1998)		00 - .10 virtually none .11 - .40 slight .41 - .60 fair .61 - .80 moderate .81 – 1.00 substantial
Landis & Koch (1977)	Cohen's Kappa	<0.00 poor 00 - .20 slight .21 - .40 fair .41 - .60 moderate .61 - .80 substantial .81 – 1.00 almost perfect
Portney & Watkins (2009)	ICC	<.75 poor to moderate .75 - .90 reasonable for clinical measurement
Koo & Li (2016)	ICC	<.50 poor .50 - .75 moderate .75 - .90 good >.90 excellent

Table 3. *R^2 for quadratic models for varying scoring levels and raters*

Number of Scoring Levels	Number of Raters (*k*)	R^2 ICC1	ICC2	ICC3	ICC1k	ICC2k	ICC3k
Average R^2		0.92	0.92	0.92	0.83	0.82	0.83
2	2	0.99	0.99	0.99	0.82	0.80	0.82
2	3	0.99	0.99	0.99	0.79	0.68	0.79
2	4	0.98	0.98	0.98	0.82	0.80	0.82
2	5	0.98	0.98	0.98	0.77	0.76	0.77
2	6	0.97	0.97	0.97	0.78	0.77	0.78
2	7	0.96	0.96	0.96	0.77	0.75	0.77
2	8	0.95	0.95	0.95	0.82	0.80	0.81
2	9	0.95	0.95	0.95	0.80	0.78	0.79
2	10	0.95	0.94	0.94	0.81	0.80	0.81
2	11	0.93	0.93	0.93	0.82	0.81	0.82
2	12	0.93	0.93	0.93	0.69	0.75	0.70
3	2	0.95	0.95	0.95	0.88	0.88	0.88
3	3	0.95	0.95	0.95	0.88	0.88	0.88
3	4	0.95	0.94	0.94	0.87	0.87	0.87
3	5	0.94	0.94	0.94	0.88	0.88	0.88
3	6	0.93	0.93	0.93	0.87	0.87	0.87
3	7	0.92	0.92	0.92	0.87	0.87	0.87
3	8	0.92	0.92	0.92	0.88	0.88	0.88
3	9	0.91	0.91	0.91	0.88	0.87	0.87
3	10	0.90	0.90	0.90	0.88	0.87	0.87
3	11	0.90	0.90	0.90	0.88	0.87	0.87
3	12	0.89	0.89	0.89	0.88	0.87	0.87
4	2	0.92	0.92	0.92	0.84	0.84	0.84
4	3	0.93	0.92	0.92	0.84	0.83	0.84
4	4	0.92	0.92	0.92	0.83	0.83	0.83
4	5	0.92	0.92	0.92	0.84	0.83	0.83
4	6	0.91	0.91	0.91	0.84	0.84	0.84
4	7	0.90	0.90	0.90	0.84	0.84	0.84
4	8	0.90	0.90	0.90	0.82	0.82	0.82
4	9	0.89	0.89	0.89	0.84	0.83	0.83
4	10	0.89	0.89	0.89	0.84	0.84	0.84
4	11	0.87	0.87	0.87	0.83	0.83	0.83
4	12	0.88	0.88	0.88	0.83	0.83	0.83
5	2	0.90	0.90	0.90	0.81	0.81	0.81
5	3	0.91	0.91	0.91	0.82	0.82	0.82
5	4	0.91	0.91	0.91	0.81	0.81	0.81
5	5	0.90	0.90	0.90	0.81	0.81	0.81
5	6	0.90	0.90	0.90	0.81	0.81	0.81
5	7	0.90	0.90	0.90	0.82	0.82	0.82
5	8	0.89	0.89	0.89	0.82	0.82	0.82
5	9	0.88	0.88	0.88	0.81	0.80	0.81
5	10	0.87	0.87	0.87	0.81	0.80	0.80
5	11	0.87	0.87	0.87	0.80	0.80	0.80
5	12	0.86	0.86	0.86	0.81	0.81	0.81

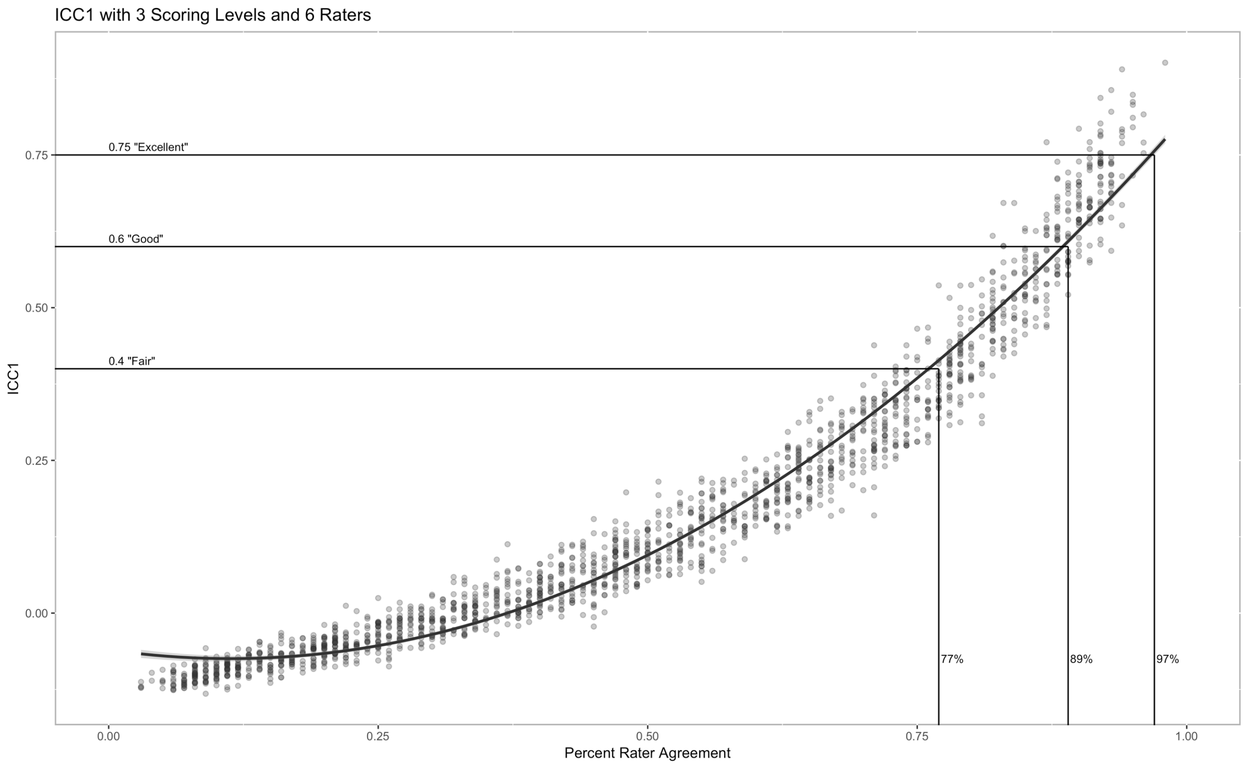

Figure 1. Percent Rater Agreement and ICC1 for 3 scoring levels with 6 raters. Quadratic regression line and Cicchetti's (2001) guidelines superimposed.

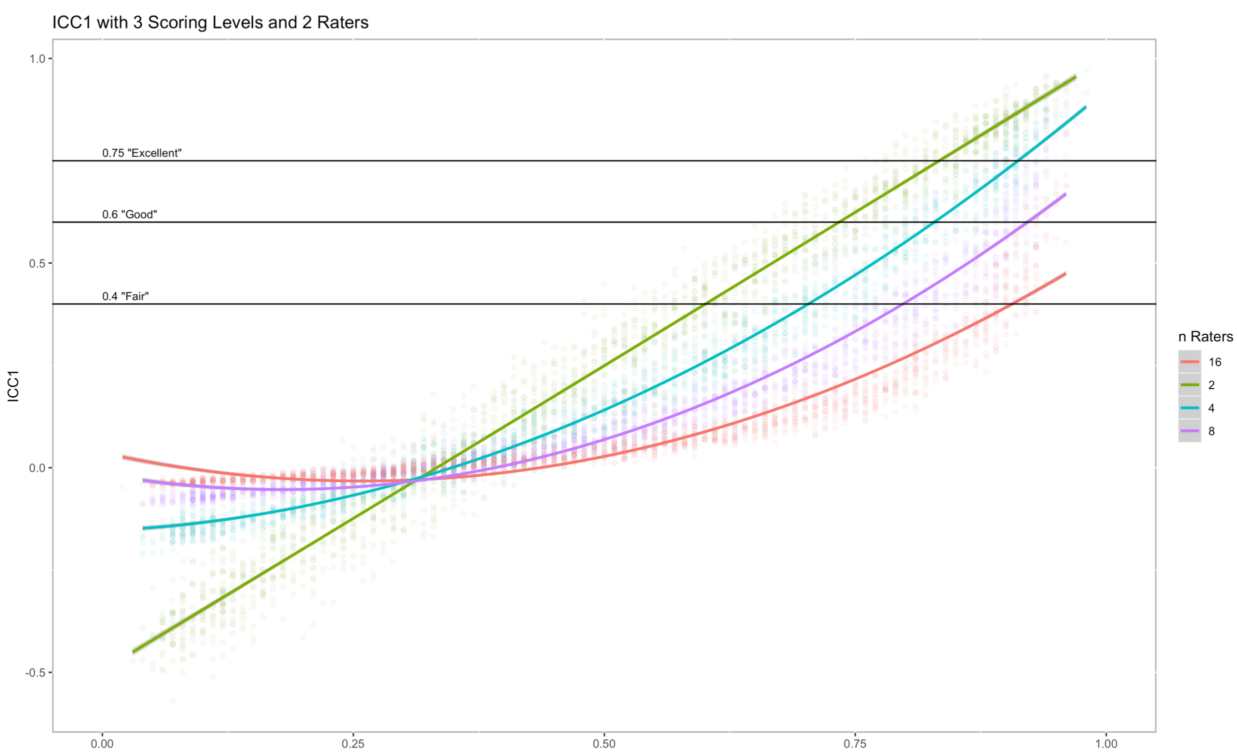

Figure 2. Percent Rater Agreement and ICC1 for 3 scoring levels and 2, 4, 8, and 16 raters. Quadratic regression line and Cicchetti's (2001) guidelines superimposed.

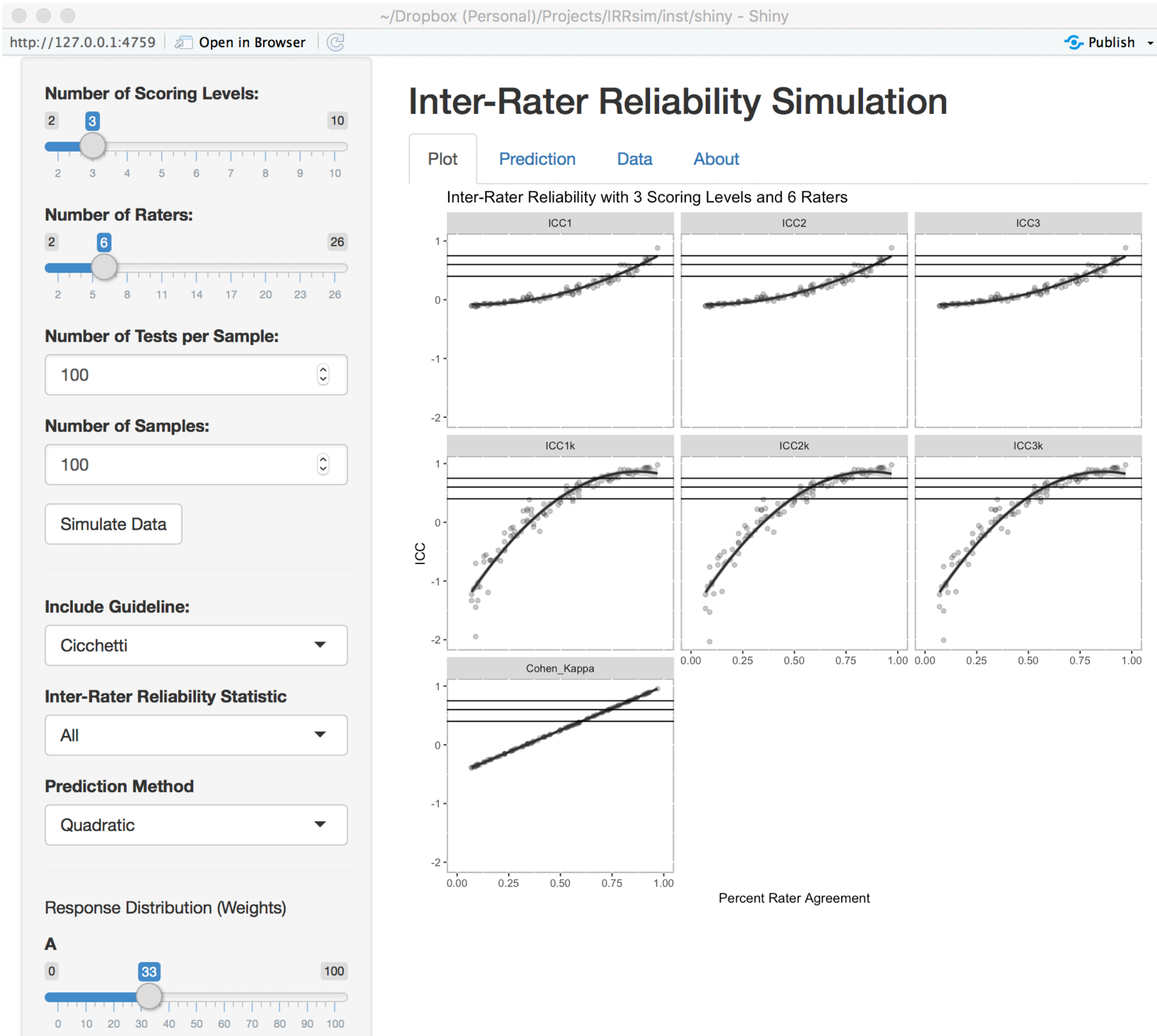

Figure 3. Screenshot of the IRRsim R Shiny App.

Appendix A: R Code

```
library(IRRsim)
data("IRRguidelines")

set.seed(2112) # For reproducibility

##### Example: Single rating matrix
test <- simulateRatingMatrix(nLevels = 3, k = 6, agree = 0.6, nEvents = 100)
print(head(test), na.print = '')

##### Example 1: 3 scoring levels with 6 raters
test1 <- simulateICC(nSamples = 200, nLevels = 3, nRaters = 6)
icc1.summary <- summary(test1, stat = 'ICC1', method = 'quadratic')
summary(icc1.summary$model)

# Calculate the corresponding PRA for Cicchetti's guidelines
newdata = data.frame(agreement = seq(0.01, 1, 0.01))
predictions <- predict(icc1.summary$model, newdata = newdata)
tab <- data.frame(Label = paste0(IRRguidelines[['Cicchetti']], ' "',
                                names(IRRguidelines[['Cicchetti']]), '"'),
                  ICC = IRRguidelines[['Cicchetti']],
                  Agreement = sapply(IRRguidelines[['Cicchetti']],
                                  FUN = function(x) {
                                   min(which(predictions >= x)) / 100 }))

# Figure 1
plot(test1, stat = 'ICC1', method = 'quadratic') +
   geom_segment(data = tab, color = 'black', x = -Inf,
                aes(y = ICC, yend = ICC, xend = Agreement)) +
   geom_segment(data = tab, color = 'black', y = -Inf,
                aes(x = Agreement, xend = Agreement, yend = ICC)) +
   geom_text(data = tab, aes(x = 0, y = ICC, label = Label),
             color = 'black', vjust = -0.5, size = 3, hjust = 'left') +
   geom_text(data = tab, aes(x = Agreement, y = min(predictions),
                          label = paste0(round(Agreement*100), '%')),
             color = 'black', size = 3, hjust = -0.1)
ggsave('NCME1.png', width = 13, height = 8)

##### Example 2: 3 scoring levels with 2, 4, 8, and 16 raters
test2 <- simulateICC(nSamples = 200, nLevels = 3,
                     nRaters = c(2, 4, 8, 16))

icc2.summary <- summary(test2, stat = 'ICC1', method = 'quadratic')

# Add Cicchetti's guidelines
guide <- data.frame(label = paste0(IRRguidelines[['Cicchetti']], ' "',
                                names(IRRguidelines[['Cicchetti']]), '"'),
                    y = IRRguidelines[['Cicchetti']])

# Figure 2
plot(test2, stat = 'ICC1', method = 'quadratic', point.alpha = 0.05) +
   geom_hline(yintercept = IRRguidelines[['Cicchetti']]) +
   geom_text(data = guide, aes(x = 0, y = y, label = label),
             color = 'black', vjust = -0.5, size = 3, hjust = 'left')
ggsave('NCME2.png', width = 13, height = 8)
```