Visual and Statistical Methods to Calculate Interrater Reliability for Time-Resolved Qualitative Data: Examples from a Screen Capture Study of Engineering Writing Patterns

Mr. Manoj Malviya, Pennsylvania State University

Master's student

Dr. Catherine G.P. Berdanier, Pennsylvania State University

Catherine G.P. Berdanier is an Assistant Professor in the Department of Mechanical Engineering at Pennsylvania State University. She earned her B.S. in Chemistry from The University of South Dakota, her M.S. in Aeronautical and Astronautical Engineering and Ph.D. in Engineering Education from Purdue University. Her research interests include graduate-level engineering education, including inter- and multidisciplinary graduate education, online engineering cognition and learning, and engineering communication.

Prof. Natascha Trellinger Buswell, University of California, Irvine

Natascha Trellinger Buswell is an assistant teaching professor in the department of mechanical and aerospace engineering at the University of California, Irvine. She earned her B.S. in aerospace engineering from Syracuse University and her Ph.D. in engineering education from the School of Engineering Education at Purdue University. She is particularly interested in teaching conceptions and methods and graduate level engineering education.

Visual and Statistical Methods to Calculate Interrater Reliability for Time-Resolved Qualitative Data: Examples from a Screen Capture Study of Engineering Writing Patterns

Abstract: Traditionally, interrater reliability (IRR) is determined for easily defined events, such as deciding within which category a piece of qualitative data falls. However, for time-resolved or time-dependent observational data and other nontraditional data, complications arise due to the complexity of the data being interpreted and analyzed. In this paper, we present two promising new methods for calculating IRR based on visual representations of analyzed time-resolved data. We compare the IRR calculated using these two visual methods with five of the most common statistical measures for calculating IRR, finding excellent agreement between our new methods and existing statistical formulae. This methods development is exemplified using data for our ongoing research, in which we are working to analyze time-resolved engineering writing data recorded through screen capture technology. The process of developing methods of interrater reliability for our context can also be applied to other researchers who seek to analyze non-traditional data, such as those collected during eye-tracking, screen capture, or observational studies.

Introduction

This research paper presents two novel image-based methods for calculating interrater reliability (IRR) and compares them with statistical methods for calculating IRR. Across fields, establishing quality in the qualitative data analysis process involves calculating a measure of agreement between the human researchers interpreting the data: If researchers cannot agree to an acceptable level, then a coding schema cannot be considered sound and results cannot be considered meaningful, transferrable, or conclusive. The extent to which the classification patterns of two or more coders coincide represents the interrater reliability, sometimes known as interrater agreement. Methods for calculating IRR have been established across the social sciences, such as those documented by Eckes [1], Zhao [2], Krippendorff [3], and Carletta [4], typically calculated for nominal data (i.e., data that can be sorted into categories that are not in any meaningful order.)

As part of our group's ongoing work, we are interested in capturing and studying the time-resolved processes of engineering writers using screen-capture data collected over hours of authentic writing practice. The overarching motivation for the project is to capture similarities and differences in the enacted writing patterns of engineering writers to elicit heuristics and useful writing strategies that can augment engineering students' writing strategies in overcoming procrastination, writer's block, and writing anxiety, which are known to plague engineering students [5]. Data for this project were collected in one semester from three graduate engineering students at a Research-Intensive University as they were applying for the National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP), a competitive fellowship that requires, among other metrics of academic success, a two-page research proposal and a three-page personal statement. During the three participants' writing process, we instructed the participants to enable screen capture software on their computers before starting the writing process, which operated "in the background" to collect movie files of the screen capture. As noted in our past literature, this also captured all the realistic factors that surround writing in the real world, such as checking email, answering instant messages from friends, and changing music types—tasks not

affiliated directly with composition, but that are part of an authentic writing process [5]. To date, we have outlined the choices and justifications involved in characterizing and coding "messy" data through theory-driven coding schemas; defining units of analysis; and introduced strategies for visually representing of hours of time-resolved data in easy-to-interpret figures [5]. In our past work, we simply used two raters working together to analyze the data and code to agreement, we feel that to advance this project, it is essential to develop methods by which to calculate IRR for non-traditional, messy, and time-resolved data such that other researchers using interesting methods to capture data might find our work transferrable and applicable. To code our time-resolved video data, we have established methods for raters to code data in real-time using a web interface [5]; however, with that step forward, the next stem and the focus of this paper is to determine a method to calculate inter-rater reliability on complex data.

Our data has some intricacies which traditional calculations for IRR become burdensome to perform. For example, our data are time-resolved and our codes are not mutually exclusive, such that we have overlapping data. Further, traditional measures of IRR have no thresholds to justify how closely two raters' codes must temporally match: Is a lag of a half a second on the same code still "agreement?" For the purposes of this paper, we are interested in establishing interrater reliability only for two raters for nominal data, and will not present methods for IRR that account for more than two raters or for ordinal, ratio, or interval methods of data categorization, because these methods are not related to our context.

The remainder of the paper is outlined in the following way. First, we introduce interrater reliability as a methodological requirement for qualitative research, reviewing the development of IRR and attributes and limitations of various approaches as a cohesive review for readers new to IRR. Then, we present the methods for two novel calculations of IRR employing image processing techniques that avoid needlessly-complex statistics, and compare the calculations of IRR from these new techniques with five existing methods for statistically calculating IRR.

Review of Statistical Methods determining Interrater Reliability for Nominal Data

Interrater reliability can be conceptualized as a percentage agreement between two raters. Here we present a simple example to calculate agreement between two raters (A and B), who are tasked to classify the same n pieces of data into either of two categories (1 and 2). In this example, Raters A and B categorized n_{11} subjects in category 1 and n_{22} subjects in category 2. However, Rater A categorized n_{12} subjects in category 1, but the same n_{12} subjects are been categorized in category 2 by Rater B. Similarly, Rater B categorized n_{21} subjects in category 1, but the same n_{21} subjects are been categorized in category 2 by Rater B. The next step is to make an agreement matrix, such as the one it Table 1.

Table 1: Agreement matrix defined for interrater estimation adapted from [6]

		Rater B		
Rater A	Category	1	2	Total
	1	n_{11}	n_{12}	a_1
	2	n_{21}	n_{22}	a_2
		b_1	b_2	n

The percent agreement p_a, is defined as the probability of both raters coming to the same conclusion, represented by the mathematical expression,

$$p_a = (n_{11} + n_{22})/n \tag{1}$$

where n_{11} and n_{22} are the categories of agreement and n is the total number of data pieces categorized. The problem with this equation for percent agreement is it doesn't consider the "chance agreement," or agreement that would happen if Rater A randomly assigned the data into categories. To overcome this limitation, several researchers have proposed modifications to the percent agreement calculation, especially Scott, Cohen, Gwet, Krippendorff, and Brenan and Prediger, the contributions of whom we introduce here.

In 1955, William Scott [7] proposed a solution for the percent chance agreement compensation (p_e), developed for text categorization, a measure of the ratio between the difference between the observed and expected (chance) agreement of raters over the maximum possible agreement accounting for expected (chance) agreement. Scott's Pi coefficient (π), is

$$\pi = \frac{p_a - p_e}{1 - p_e} \tag{2}$$

where p_a is the observed agreement between two raters on two categories using the percent observed agreement calculation in Equation (1) and p_e is the expected agreement due to chance. The percent chance agreement (p_e) calculated using joint proportions and assuming raters have the same response distribution, given by the following expression

$$p_e = \sum_{k=1}^{q} \frac{(\pi_{ak} + \pi_{bk})^2}{4} = (\frac{a_1}{2n} + \frac{b_1}{2n})^2 + (\frac{a_2}{2n} + \frac{b_2}{2n})^2 \tag{3}$$

Where q is the number of categories, a corresponds to Rater A and b to Rater B, the subscripts 1 and 2 correspond to categories and π_{xk} is the probability of Rater x categorizing a subject to the k^{th} category defined as the ratio of number of subjects in category k and total number of subjects. However, this method assumes that the chances of raters randomly assigning an item to same category is based on rater's average distribution for each category which is not accurate representation of the experiment.

In 1960, Cohen [8] estimated the expected percent chance agreement and used it to adjust the percent agreement (p_a). Cohen defined the Kappa coefficient (κ_c) as:

$$\kappa_c = \frac{p_a - p_e}{1 - p_e} \tag{4}$$

Where the chance agreement (p_e) is defined as

$$p_e = \sum_{k=1}^{q} \pi_{ak}\pi_{bk} = \frac{(a_1 b_1 + a_2 b_2)}{n^2} \tag{5}$$

As noted by Feinstein & Cicchetti [9], [10], there are two paradoxes of the kappa coefficient where the kappa coefficient is highly inaccurate. First, if the chance agreement is very large, it can convert a relatively large percent agreement into a very small k_c. For example, if rater

A doesn't categorize any subject into category 2 and rater B categorizes the subject such that n_{11} = 100 and n_{12} = 25, the kappa coefficient can be determined using equations 1, 4, and 5 to achieve the following values:

$$p_a = \frac{100 + 0}{125} = 0.8, \quad p_e = \frac{125 \times 100 + 0 \times 20}{125^2} = 0.8$$

$$\kappa_c = \frac{0.8 - 0.8}{1 - 0.8} = 0$$

In this example, the percent chance agreement is very high but the kappa coefficient is zero, suggesting (inaccurately) a total disagreement between raters. The second paradox is that a very high disagreement in distribution of subjects between raters can artificially produce high values of κ_c. For example, if rater A and B categorized the subject such that n_{11} = 20, n_{12} = 15, n_{21} = 5, n_{22} = 20, then the percentage agreement (p_a) based on the data in example is 0.67 and the chance agreement (p_e) is 0.49, resulting in a kappa coefficient of 0.35 which is very low compared to the percentage agreement.

The methods discussed until now cannot be described as a *general method* for measuring IRR since they are limited by severe paradoxes. First proposed by Holley and Guilford in 1964 [12] and generalized later by Brenan and Prediger in 1981 [13] for more than two categories, the G-Index is perhaps the simplest expression for percent chance agreement compensation. It is defined as

$$k_b = \frac{p_a - 1/q}{1 - 1/q} \tag{6}$$

Where q is the number of categories. Since k_b depends only on the number of categories (q) and hence, it is independent of subject rating distribution by raters making it a paradox resistant method. However, the chance agreement coefficient is not constant and depends upon the rating distribution.

In 1970, Krippendorff [11] proposed an agreement coefficient named α (alpha), in which percent chance agreement is identical to Scott's Pi coefficient, but the percentage agreement is the weighted average of the observed percent agreement (p'_a).

$$p'_a = (1 - \varepsilon_n)p_a + \varepsilon_n \tag{7}$$

Where the weight parameter (ε_n) is defined as $\varepsilon_n = 1/2n$, where n is the number of subjects rated by the both raters. The weighted parameter acts as a small sample correction for percentage agreement coefficient, eliminating all the subjects rated by single rater. However, the correction in percent agreement is insignificant for larger number of subjects, yielding in the same result as Scott Pi. This method is preferred over Scott Pi coefficient when all subjects are rated by two or more raters and the number of subject is less than 10.

Most recently, in 2008, Gwet [14] proposed an agreement coefficient called AC_1 to overcome the limitations associated with Cohen's Kappa and Scott Pi coefficient. The expression of coefficient is identical to Cohen's Kappa and the only difference is the expression for (p_e)

$$p_e = \frac{1}{(q-1)} \sum_{k=1}^{q} r_k(1 - r_k) \tag{8}$$

Where q is the number of categories, and the term r_k term is the mean probability of categorizing a subject into k^{th} category and can be calculated using following expression,

$$r_k = \frac{(\pi_{ak} + \pi_{bk})}{2} \tag{9}$$

This method estimates the chance agreement by using a hybrid of average (Pi) and categorical (Kappa) distribution. This approach has been proven to be more stable in a varying marginal probability compared to Cohen's kappa [15] making it paradox- resistant.

One of the limitations of all these methods for calculating IRR is the assumption of an ideal data set with no missing ratings. In practice, though, it is reasonable to assume that a data set will not be ideal and will have *missing* ratings. This happens when one or both raters are not able to classify the data, a common limitation as researchers analyze data in real-time or collect observational data. In case of both raters missing classifications, the percent agreement does not change; however, if one rater codes a behavior and the other misses the occurrence, there will be an effect on the agreement score. To incorporate the missing ratings in agreement matrix, a new row (n_{Bx}) and column (n_{Ax}) is introduced to the agreement matrix. The new row and column contains q-1 elements. Column n_{Ax} contains the number of subjects that only rater A has classified and similarly, Row n_{Bx} contains the number of subjects classified by only rater B (and not rater A). The final generalized agreement matrix is given by

$$A = \begin{bmatrix} n_{11} & n_{1q} & n_{ax1} \\ & & \\ n_{q1} & n_{qq} & n_{axq} \\ n_{bx1} & n_{bxq} & 0 \end{bmatrix}$$

If n_{ax} and n_{bx} are the sums of column X and row X respectively, then to compensate for the missing rating in the percent agreement calculation, the total number of subjects (n) is adjusted by the following expression,

$$n` = n - (n_{ax} + n_{bx}) \tag{10}$$

Where n is the total number of subjects in the experiment and n` is the adjusted total number of subjects.

These five methods summarize some existing statistical methods for calculating interrater reliability for two raters and nominal data. However, these measures are useful for data that is simple and easily categorized, and most of the measures were developed to account for data that was text-based in nature. More recently, with advances in data collection opportunities, technology, and ability to collect and store large quantities of visual and otherwise qualitative data, researchers have needed to begin to apply IRR techniques in modified ways to conduct research across disciplines. As examples of some of this research, Masufumi and Gribble [16] demonstrated a way of estimating intra-rater and inter-rater reliability in image analysis of five image based criteria to investigate foot posture using visually observable criteria. Mishima et al.

[17] implemented inter-class correlation methods to determine IRR to investigate images during a posed smile using a video-based motion analysis system. These examples show that determining how to calculate IRR in image-based and observational data is an issue that is applicable across disciplines and further motivates our methodological research.

Development of Methods to Estimate IRR for Time-Resolved Writing Data

The codebook presented in Table 2 is used to code the screen-recorded data captured from our participants. This codebook is slightly modified from our previous work [5] to reduce complexity. These codes were developed out of cognitive writing theory and have been presented and justified previously [5,18]. For the purposes of calculating IRR, we assigned each code a numerical value so we can manipulate data in MATLAB. The numerical values have no bearing on importance or order; they are simply for computational bookkeeping. Consistent with the methods we have reported in previous literature [5, 18], raters code on a web-interface that transfers the codebook into an interactive GUI through the free online tool called the Generalized Observation and Reflection Platform (GORP), hosted by UC Davis (https://cee.ucdavis.edu/GORP). While there are limitations to the GORP tool, the advantage of being free, intuitive, and able to be run on a touch screen laptop far outweigh limitations. The data are captured in real time and outputs as a spreadsheet file, which reads the categories as a function of time points. The resulting data file can be manipulated in MATLAB or other programs.

Table 2: Codebook and Numerical Values Assigned for Data Processing

Level	Definition of Level	Code	Numerical Value
	Monitoring of process overall	Planning (Sense making)	1
Process Level	Internal and external processes involved in the process of actually writing text	Composing	2
		Editing	3
		Revision	4
		Collaborators & Critics	5
		Technology	6
		Task Requirements and Materials	7
Resource Level	Includes internal memories and general-purpose processes that processes at the other levels can call on	Attention (Away from Writing Task)	8
		No Apparent Activity	9
		Knowledge gathering	10
		Reading Text to Date	11

Elimination of Time Overlap and Lag for Statistical Calculation of IRR

The most pressing issue in handling time-dependent observational data, such as observations, screen-capture, and eye-tracking data is that events are continuous—occurring consistently over a period of time, not at a single time point—and often overlap with each other. The overlapping condition is dependent on theoretical methodological choices: For our time-resolved writing project, cognitive writing theory governs that a writer is often doing multiple things at a time; thus, our data overlaps frequently. A generalized set of time-resolved data here is defined as shown in Table 3.

Table 3: General time-dependent observational data table

Time Start	Time End	Code
S_{1k}	E_{1k}	C_{1k}
⋮	⋮	⋮
S_{ik}	E_{ik}	C_{ik}
⋮	⋮	⋮
S_{nk}	E_{nk}	C_{nk}

Here S_i and E_i are the start time and end time of the event when the rater k coded C_i. Here, an assumption of no data missing is taken which can be mathematically expressed as

$$E_{i-1} \geq S_i \tag{11}$$

If $E_{i-1} > S_i$, there is an overlap between events, meaning that in order to calculate IRR, the overlapping time event will have to be split into distinct non-overlapped events. A four-step process is implemented to manipulate data (Figure 1).

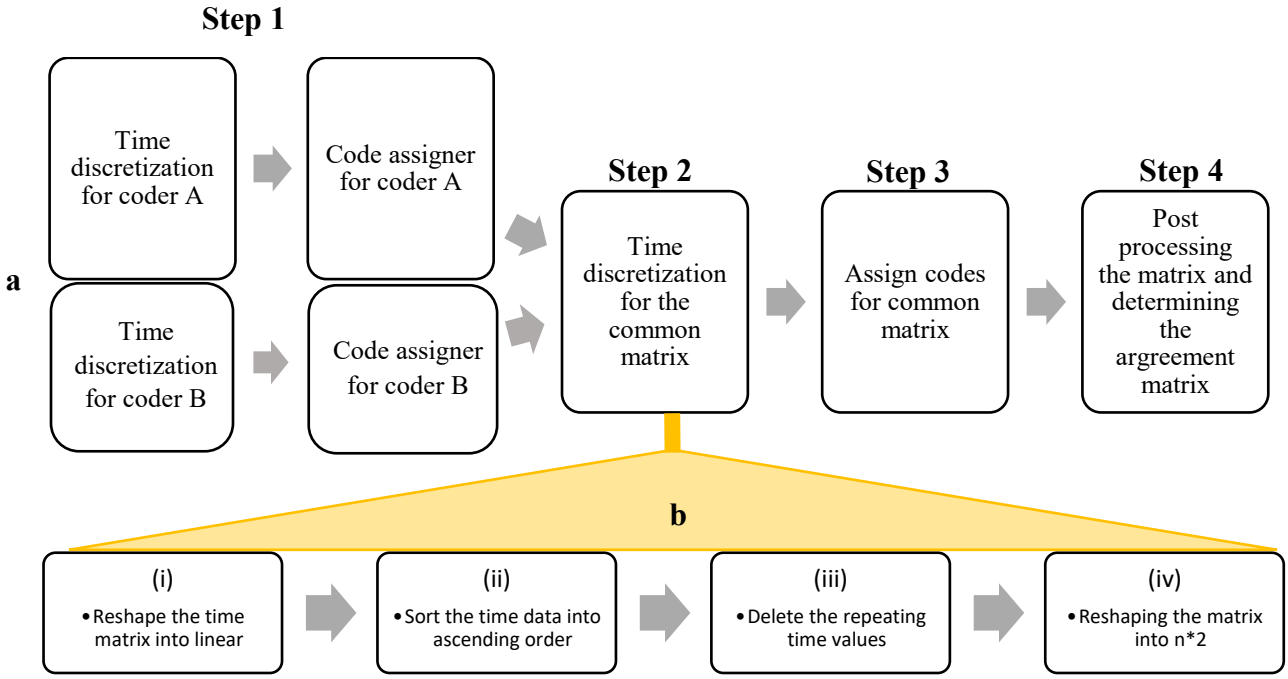


Figure 1. (a) Four-step process to clean data to eliminate overlapping data categories (b) Time discretization process

To illustrate the process at each step, an example is presented in Figure 2. The first step is to split the time intervals into non-overlapping time intervals for each coder. A generalized process to achieve non-overlapping time intervals is shown in Figure 1(b). The time data (start time and end time) is stored in a different matrix and reshaped into a linear matrix. The time data values are sorted in ascending order and the repeating time values are eliminated from the matrix. The sorted matrix is reshaped again with start time and end time. Step 2 assigns codes for the non-overlapping time intervals. A search method is implemented to find codes in a given the time interval. However, the process becomes more complex as it might be possible that the writer is doing multiple things

at once resulting in multiple code for a single time interval. If categories occur simultaneously, the codes are sorted in ascending order to make the data organized. We implemented an overlapping procedure approach presented in [18] to incorporate multiple codes. We selected a data threshold such that short time intervals (< ~3 s) were ignored. After manipulating the data, an agreement matrix is formed similar to shown in Table 1, on which any statistical IRR calculation can be executed (Cohen's Kappa, Krippendorff's Alpha, Scott's Pi, etc.).

Step 1

Rater A	
0 - 1	1
0 - 3	3
3- 4	3
4-5	2

Rater A	
0 - 1	1,3
1 - 3	3
3- 4	3
4- 5	2

Rater B	
0 - 3	3
2 - 3	1
3- 5	2
4-5	3

Rater B	
0 - 2	3
2 - 3	1,3
3- 4	2
4 - 5	2,3

Step 2

Common Time	Rater A	Rater B
0 - 1	1 **3**	0 **3**
1 - 2	**3**	**3**
2 - 3	0 **3**	1 **3**
3- 4	3	2
4 -5	**2** 0	**2** 3

Step 3

Step4

		Rater B			
		0	**1**	**2**	**3**
Rater A	**0**	0	1	0	1
	1	1	0	0	0
	2	0	0	1	0
	3	0	0	1	3

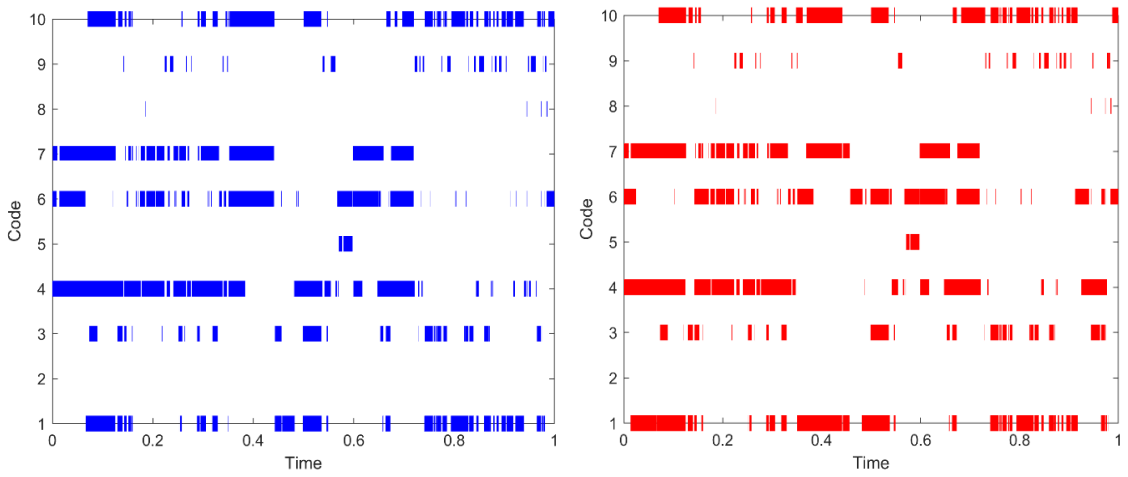
Figure 2. An example of eliminating the time overlapping issue

The limitations of a purely statistical approach to calculating IRR is that it is difficult to implement in practice, with a deep computational and statistical skillset required. Because our data can be analyzed visually, representing time-dependent data in images, we present here two novel additional methods for calculating IRR that have not been proposed in literature to our knowledge.

Novel Visual IRR Method 1: Image Comparison Approach

The Image Comparison Method we propose here leverages the ability to plot observational data in two dimensions simply characterized by codes occurring over a duration of time. These plots can be easily generated for both raters A and B in data processing software and exported as PNG files, as shown in Figure 3. These images are imported to an image processing software (we employ MATLAB software) where the images showed in Figure 3 are preprocessed for the sake of simplicity. The preprocessing step involves cleaning the axes and legends and converting the images into grayscale. Then, the images are compared and percent of similarity is found based on the structural similarity index (SSIM) [19]. Both the data visualization and the structural similarity index can be performed using image analysis software that is readily found in analysis software packages such as MATLAB. SSIM is generally used for measuring the image quality by comparing two images based on the computation of three parameters namely: Luminance, Contrast and Structure. As we are interested in finding the similarity between structures of the image, the luminance and contrast terms are ignored. Both images are imported in MATLAB as PNG files

with a white background. However, the white space of the images will interfere with the results; therefore, we must manipulate the white background before image comparison analysis. This manipulation is achieved using a dynamic control feature of SSIM where we define the range of



(a) (b)

Figure 3. Two-dimensional data visualization of observances by Rater A and B

colors that are compared. SSIM requires one of the images to be categorized as a reference image from which the similarity between structures is analyzed. The image with maximum number of data points is selected as reference. The modified SSIM result is then taken to be the p_a value that can be supplied to the Brenan-Prediger correction for the G-index calculation (Eq.6) that accounts for the expected (chance) error based solely on number of categories.

Novel Visual IRR Method 2: Scattered Points Approach with Chance Agreement Compensation

The scattered points approach leverages the ability of our data to be visualized in a three-dimensional space, with time start point and the code on the x- and y- axes, respectively, and the percentage duration of total time on the z-axis, as shown in Figure 4. As such, each occurrence of the observational data can be represented by a point in a three- dimensional space.

The scattered point approach is described by the flowchart in Figure 5. First, the coded data from both Rater A and Rater B are converted into three–dimensional data to envision a three-dimensional plot of each Rater's data. The dataset with maximum number of points (we can call it Data 1) is selected to be the reference data set, to correct for missing data. Systematically, a point is selected from Data 1 (called point t) and the Data 2 is searched for the points falling in the time vicinity of point t selected. The vicinity of the point is defined as per time start and percentage total time tolerance value defined by the user. This allows researchers to set thresholds to account for lag time between researchers. If a suitable matching point is found in Data 2, a proxy variable (k) is increased by one, which serves as a counting method to keep track of the points found in common. The process is repeated for each point in Data 1 until all points from Data 1 have been searched in Data 2. The percentage agreement is defined as the ratio of number of points found common (k) and total number of points in Data 1 (n).

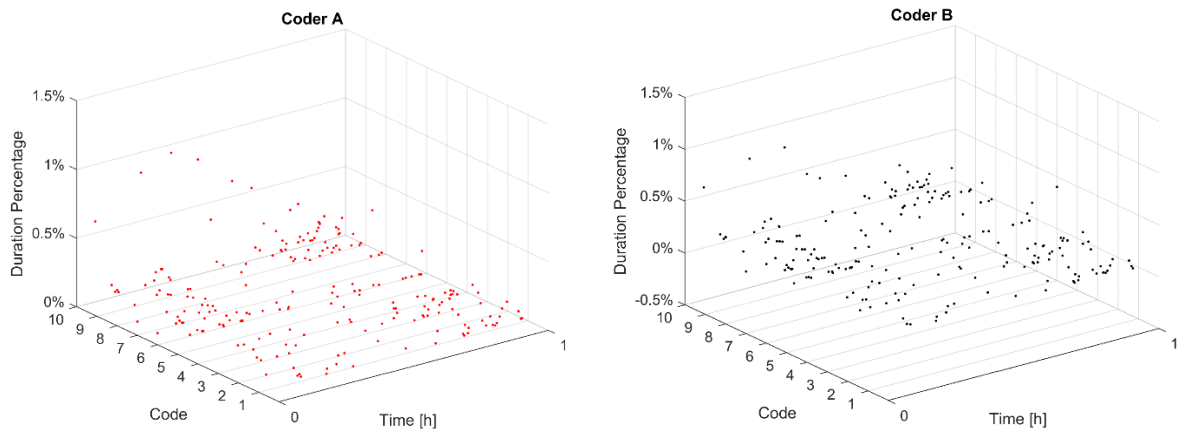


Figure 4. Three-Dimensional Visualization of Data

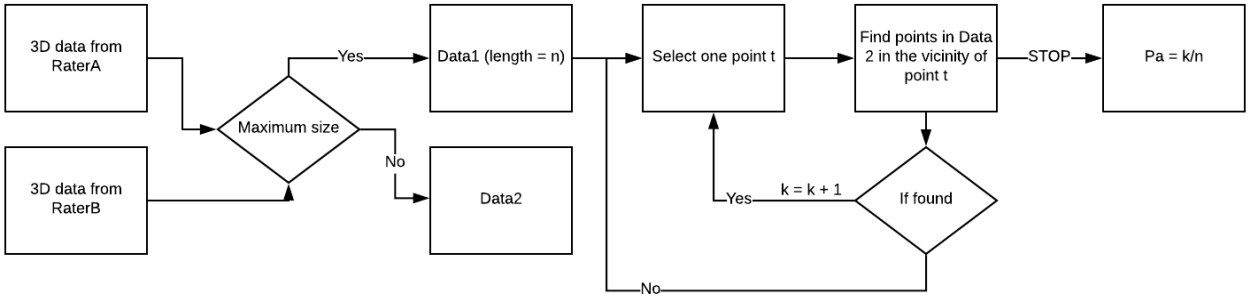


Figure 5. Algorithm Flow Chart for 3D Scattered Points IRR Approach

To account for the expected (chance) agreement, Brenan and Prediger's method [13] is implemented as it is independent of the data type and depends only the number of categories. IRR then can easily be calculated from the percent agreement calculated by the algorithm using the following equation

$$IRR = \frac{k/n - 1/q}{1 - 1/q} \tag{11}$$

Where k is the number of codes in agreement within the time threshold defined by the researcher, n is the total number of codes in Data 1 (Rater 1's 3-D scatterplot data), and q is the number of categories by which we can account for the expected (chance) agreement.

Results and Discussion: Comparison of Statistical and Proposed Visual Methods for IRR

We analyzed the five statistical IRR calculations in comparison to the two visual methods proposed in addition to the statistical methods for our data set (Table 4). The five statistical methods give approximately the same results for the given data, as expected, with most variation coming from the methods in which the expected (chance) agreement was corrected. The visual methods for calculating IRR (3D Scatter and Image Comparison) are very close to the statistical methods. The results suggest that the visual IRR methods can be implemented instead of spending

significant time cleaning overlapping data to create an appropriate agreement matrix for statistical methods. The scatterplot method has an additional benefit in that the researcher can set appropriate thresholds for how much lag time between rater coding can be considered still to be in agreement.

Table 4: Comparison of Calculated IRR Based on Statistical and Visual Methods

Method	Percent agreement (p_a)	Expected (chance) agreement correction (p_e)	IRR
Statistical IRR Methods			
Scott's Pi	0.7003	0.1446	0.6496
Cohen's Kappa	0.7003	0.1428	0.6504
Krippendorff alpha	0.7003	0.1446	0.6946
Brenan-Prediger	0.7003	0.0714	0.7190
Gwet AC	0.7003	0.0855	0.6772
Proposed Visual IRR Methods			
3D Scattered Points Approach	0.7808	0.0714	0.7643
Image Comparison Approach	0.83	0.0714	0.81469

After calculating interrater reliability, researchers must then use judgement to determine if the interrater reliability is high enough to signify that the coding scheme is reliable across raters, or if further honing of the coding schema is required. This "goodness" is somewhat flexible, and is context-dependent with regard to how critical perfect agreement is to the quality of research. In general, literature suggests that an interrater reliability (with respect to Cohens Kappa) of 0.41-0.6 as moderate, 0.61 -0.80 as substantial, and 0.81-1 as almost perfect (with 1 being the maximum)[20]. Other scholars [21] instead posit that Kappa values of 0.40-0.75 as fair to good, and above 0.75 as excellent. Since our data fall within the range of acceptability, we take the IRR calculations for both our proposed methods and our data as a whole to be satisfactory.

Conclusion and Future Work

In this paper, we presented two new image-based IRR calculation procedures, validating them on five statistical measures of interrater reliability using our data as a context. Our analyses confirm high interrater reliability in our data, and achieved high agreement between our visual methods and existing numerical IRR values. These data are promising, yet preliminary because they are applied to only our data sets. Future work includes applying these visual IRR methods to other time-resolved observational data to validate their transferability, both in the context of writing and in the research group's other research initiatives. It is easy to envision these methods for visually calculating IRR being applied to eye-tracking data, observational data, or other methods that are time-resolved or have overlapping qualitative categories.

References

[1] T. Eckes, “Introduction to Many-Facet Rasch Measurement,” p. 160, 2011.

[2] X. Zhao, J. S. Liu, and K. Deng, “Assumptions behind Intercoder Reliability Indices,” Ann. Int. Commun. Assoc., vol. 36, no. 1, pp. 419–480, 2013.

[3] K. Krippendorff, “Reliability in Content Analysis.,” Hum. Commun. Res., vol. 30, no. 3, pp. 411–433, 2004.

[4] J. Carletta, “Squibs and Discussions Assessing Agreement on Classification Tasks: The Kappa Statistic,” Comput. Linguist., 1993.

[5] C. G. P. Berdanier and N. M. Buswell, “Data Visualization for Time-Resolved Real-Time Engineering Writing Processes” 125th ASEE Annual Conference & Exposition, Salt Lake City, UT. 2018.

[6] K. L.Gwet, Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters, Fourth Edi. 2010.

[7] W. A. Scott, “Reliability of Content Analysis: The Case of Nominal Scale Coding,” Public Opin. Q., vol. 19, no. 3, p. 321, 1955.

[8] J. Cohen, “A coefficient of agreement for nominal scales,” Educ. Psychol. Meas., vol. 20, pp. 37–46, 1960.

[9] A. R. Feinstein and D. V. Cicchetti, “High agreement but low kappa: II. Resolving the paradoxes,” J. Clin. Epidemiol., vol. 43, no. 6, pp. 551–558, 1990.

[10] D. V Cicchetti and a R. Feinstien, “High agreement but low kappa: I. The problems of two paradoxes,” J. Clin. Epidemiol., vol. 43, no. 6, pp. 551–558, 1990.

[11] K. Krippendorff, “Estimating the reliability, systematic error and random error of interval data,” Educ. Psychol. Meas., vol. 30, pp. 61–70, 1970.

[12] J. W. Holley, “A note on the G index of agreement,” Educ. Psychol. Meas., vol. XXIV, no. 4, pp. 749–753, 1964.

[13] R. L. Brennan and D. J. Prediger, “Primary Identify Kappa 1960) Reliability Validity Kappa-Like,” Educ. Psychol. Meas., vol. 41, no. 1960, pp. 687–699, 1981.

[14] K. L. Gwet, “Computing inter-rater reliability and its variance in the presence of high agreement,” Br. J. Math. Stat. Psychol., vol. 61, no. 1, pp. 29–48, 2008.

[15] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, “A comparison of Cohen’s Kappa and Gwet’s AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples,” BMC Med. Res. Methodol., vol. 13, no. 1, pp. 1–7, 2013.

[16] W. A. M. Masufumi, and P. A. Gribble, “Intra- Rater and Inter - Rater Reliability of the Five Image - Based Criteria of the Foot,” IJSPT, vol. 9, no. 2, pp. 187–194, 2014.

[17] K. Mishima, H. Umeda, A. Nakano, R. Shiraishi, S. Hori, and Y. Ueyama, “Three-dimensional intra-rater and inter-rater reliability during a posed smile using a video-based motion analyzing system,” J. Cranio-Maxillofacial Surg., vol. 42, no. 5, pp. 428–431, 2014.

[18] C. G. P Berdanier and N. M. Trellinger, “Development of a Method to Study Real-Time Engineering Writing Processes. IEEE Frontiers in Education Conference, Oct 18-21, 2017, Indianapolis, IN. 2017.

[19] Z. Wang, A. C. Bovik, and H. R. Sheikh, “Image quality assessment: From error measurement to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, 2004.

[20] J. R. Landis and G. G. Koch, G. G. “The measurement of observer agreement for categorical data.” Biometrics, 159-174, 1977.

[21] J. Fleiss, “Measuring nominal scale agreement among many raters,” Psychol. Bull., vol. 76, pp. 378–382, 1971.