Measuring the Reliability of Qualitative Text Analysis Data

Klaus Krippendorff
University of Pennsylvania, kkrippendorff@asc.upenn.edu

Manuscript
Submitted to and published in *Quality and Quantity 38*: 787-800, 2004.

Measuring the Reliability of Qualitative Text Analysis Data

Klaus Krippendorff
The Annenberg School for Communication
University of Pennsylvania
3620 Walnut Street, Philadelphia, PA 19104.6220, USA
Tel.: USA.215.545.9356
kkrippendorff@asc.upenn.edu
2002.7.12

Abstract

This paper reports a new tool for assessing the reliability of text interpretations heretofore unavailable to qualitative research. It responds to a combination of two challenges, the problem of assessing the reliability of multiple interpretations -- a solution to this problem was anticipated earlier (Krippendorff, 1992) but not fully developed -- and the problem of identifying units of analysis within a continuum of text and similar representations (Krippendorff, 1995). The paper sketches the family of α-coefficients, which this paper extends, and then describes its new arrival. A computational example is included in the Appendix.

Keywords:

Reliability, Qualitative, Text Analysis, Unitizing, Multiple Interpretations, Krippendorff's Alpha

1 The Family of Alpha-Agreement Measures

In the last thirty some years α (alpha) has developed from a simple generalization of several agreement coefficients for two coders, notably Scott's (1956) π (pi) for nominal data, Spearman's ρ (rho) (Siegel, 1956:202-213) for ordinal data, and Pearson's (1901) and Tildesley's (1921) intraclass correlation r_{ii} for interval data into a whole family of agreement coefficients (Krippendorff, 1970, 1972, 1980, 1995, 2004). This development opened a space for consistent reliability assessments of

- Any number of observers or coders, not just two
- Incomplete data (unoccupied cells in a reliability data matrix)
- Small sample sizes, for which it corrects
- Data with any kind of metric: nominal, ordinal, interval, ratio, but also circular, polar, and specialized kinds
- Partitions and subsets of units of analysis, including individual units

- Situations in which data are unitized, not just coded. Coding of interval data has dominated the literature
- Multi-valued data, that is, multiple interpretations of single units of analysis, not just single-valued data

α enables various analyses, for example, calculating:

- Data reliability, the reproducibility of coding instructions, which is standard
- The reliability of individual coders
- The accuracy of coding processes relative to a trusted standard
- The reliability of decisions within a conceptual hierarchy of coding assignments
- The reliability of various data transformations, conditional reliabilities, reliability gains or losses due to the lumping of categories.

With D_o measuring the observed disagreement and D_e the expected disagreement, α's general form is:

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

Algebraically, when observed disagreement is absent, $D_o=0$ and $\alpha=1$, which indicates perfect reliability. When observed disagreement is merely chance, $D_o=D_e$ and $\alpha=0$, which signals the absence of reliability. This form reveals α to be a measure of how much the proportion of two disagreement measures of the reliability data deviate from the ideal of perfect agreement, $\alpha=1$.

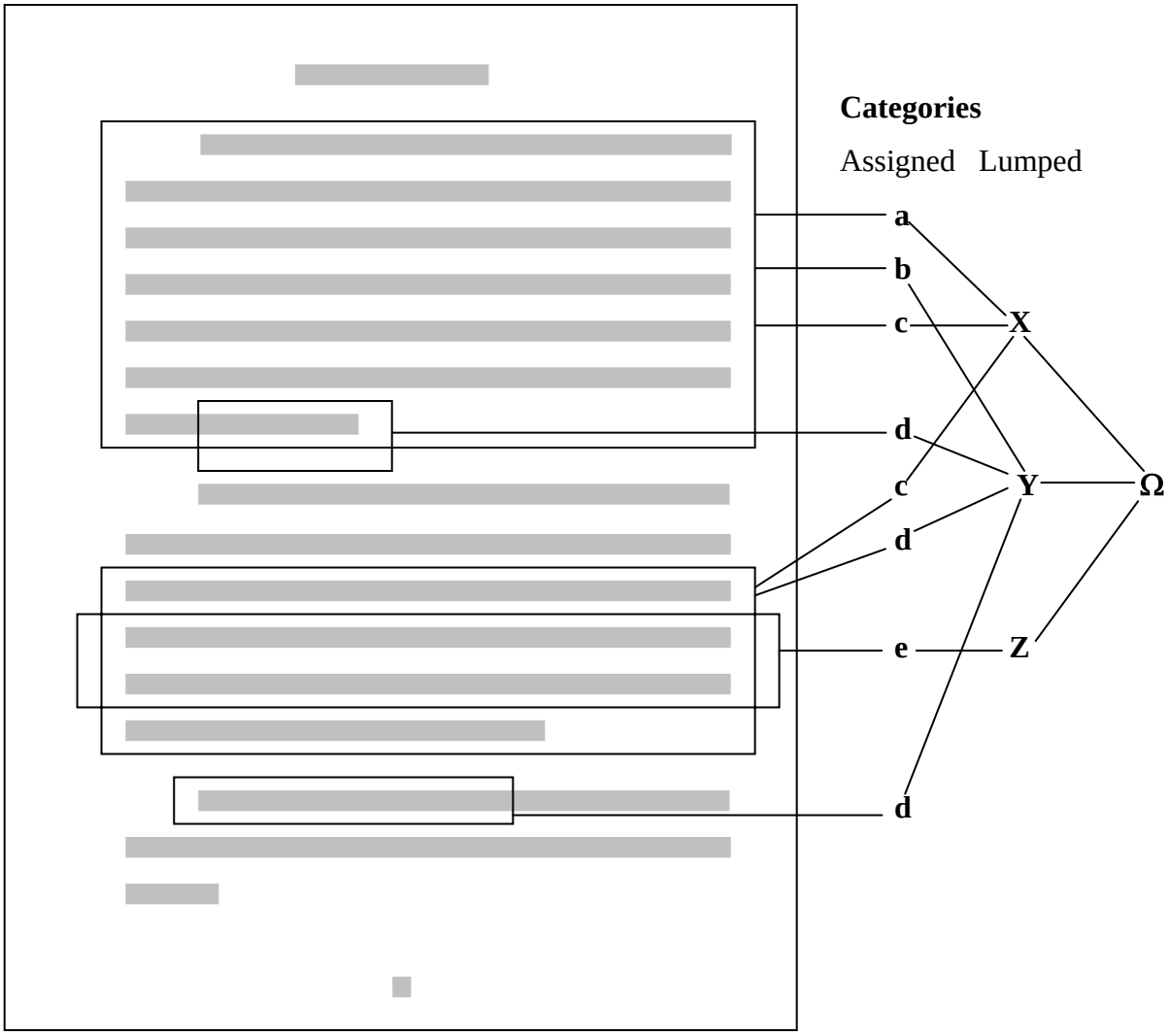
The literature shows divergent conceptualizations of agreement (Krippendorff, 1987). I will not discus these but need to alert the reader that agreement measures must not be confused with correlation coefficients or measures of association, a confusion that permeates the psychometric literature. A recent essay on reliability considerations (Brennan, 2001) attests to the almost exclusive reliance on correlations at the expense of agreements of interval data at the expense of other metrics, and on coding at the expense of unitizing and other data making processes. There are important differences in the assumptions underlying the calculations of expected disagreement, which defines the zero point of agreement coefficients. Common assumptions keep the α-family of agreement coefficients together and enable the researcher to apply uniform standards across numerous situations.

2 Qualitative Text Analysis Data

Researchers committed to qualitative analysis of text have criticized content analysts for relying on rigid definitions of textual units of analysis -- words, sentences, paragraphs, or less natural units like lines of text, 20 seconds of conversation – using one kind of unit for a whole body of text, just for being able to use available statistical techniques. I sympathize with this criticism. We are in need for ways of analyzing textual units of variable size, units that are natural to an intelligent reader and informative to the research question being pursued. The difficulty of analyzing such natural units of text has given rise to a qualitative research tradition that has essentially given up reliability concerns and focuses instead on issues of relevance to a particular contention or debate. In response, I am suggesting that the mathematical complexity of analyzing variably unitized text, while an unquestionable hurdle for replicating research, is no justification for creating the methodological schism between quantitative and qualitative approaches to analyzing textual matter. All text is qualitative to begin with. It is written to be read by intelligent and culturally competent individuals. Readers do not count, at least not to begin with. Content analysts as well as qualitative researchers interpret text, try to make sense of relevant parts of it by whatever means, and quote finite stretches from it in support of their conclusions.

A practice that literary scholars, journalist, and qualitative researchers share is to identify sections of text that they consider relevant, intend to use, revisit, or quote as a representative example of what they want to say. Students might underline relevant sections of a text. Literary scholars make notations on its margins, which amounts to a kind of categorization. Journalists might keep folders of written material for their story. Others collect quotations on index cards. Qualitative text analysis software, N*Vivo and Atlas-ti, for example, plays on the metaphor of clipping printed matter by enabling their users to highlight contiguous sections of text, assign different codes to these sections, and cut, paste, sort, list, and enumerate the highlighted portions in terms of user-assigned categories. Here, data consist of contiguous (textual) units of variable size and any number of categories or interpretations associated with each. Figure 1 offers a graphical example.

< Figure 1 About Here >



A Textual Continuum, Unitized and Categorized

Figure 1

Note that these textual units are contiguous. They may be assigned to more than one category. Units may overlap due to being assigned to different categories. Different categories are linked (become correlated) via units of text assigned in common. Several categories may be lumped into higher-order or less detailed categories.

Notwithstanding commercially motivated claims, qualitative text analysis software rarely implements theories of meaning or of reading. Such theories -- if one can call them theories -- are embodied in the intelligent and linguistically competent users of this software. It is these users who connect their own readings of text to the categories of their research question. It is a major epistemological mistake to assume that texts have inherent meanings or speak for themselves. It is equally problematic for qualitative text analysts to assume that their categorizations are self-evident, implying no need to question whether other readers/analysts would come up with same or similar categories and no need to tell the users of their findings what their coding criteria were. It is these naïve conceptions of text that obviate reliability considerations.

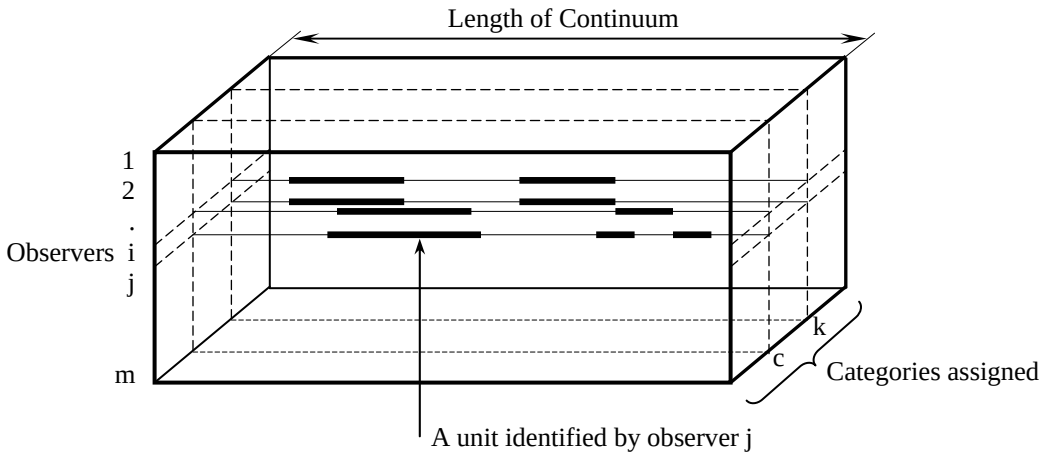
One feature that most qualitative text analysis software offers is to automatically extend a user's ad hoc coding to a body of text larger than they actually read. It amounts to operationalizing the theories by which analysts assign textual units to the categories of their analysis. This increases the efficiency of the coding process, but it still remains a single analyst's reading, stays entirely within the particular software, and says nothing about reliability. Most qualitative text analysis software is hermetically closed in this sense and makes it difficult, therefore, to assess how well it does. Yet, it is not impossible to develop coding instructions outside these computer aids that could be followed by several text analysts to yield comparable data that could shed light on the trustworthiness or reliability of the process.

I do not agree with methodologists of qualitative research who take the difficulty of measuring reliability as an excuse for being unconcerned with reliability considerations. Instead, I take this difficulty as a challenge.

3 Reliability Data for Qualitative Text Analysis

Let me be a bit more abstract in characterizing the data that qualitative text analysts typically generate. There is a continuum. This continuum may be a text, a video tape, or a period of time, anything that has an extension in a measurable dimension. In this continuum, several observers, analysts, coders, or readers introduce their own distinctions, ideally using common criteria. These

distinctions create contiguous units – highlighted text, stretches of a video recording, or intervals in time -- and irrelevant matter between them, which is left unattended. Relevant units are assigned to categories that are comparable across individual observers. The reliability data thus described can be depicted in a three dimensional data cube of observers-by-a continuum (of a certain length)-by-available categories, as in Figure 2.



Reliability Data Cube
Figure 2

In this purely graphical depiction, one may notice that the two observers, i and j, agree perfectly in category k but show disagreements in category c. In the latter, the two observers substantially agree on their first unit, i is merely a bit more conservative than j is. Regarding the other units, considerable uncertainty prevails. There seems to be no obvious pattern of agreement. The numerical representation of this example and the computations of the reliabilities are found in the Appendix.

4 Specification of Units and Gaps Between Them

4.1 Measures of Lengths

We consider the **continuum** as initially undifferentiated and known only by its beginning B and length L. Similarly, **units** and the **gaps** between units are located in this continuum by knowing their beginnings b and lengths ℓ .

The **unit for measuring** these lengths is *the smallest distinguishable length, duration, or number*, for example the characters in text, frames of film, or smallest division on a ruler. Lengths

are expressed in full integers, not in decimal points, not in units of varying size (like fractions of inches for small length and feet or miles for larger lengths).

4.2 Categories

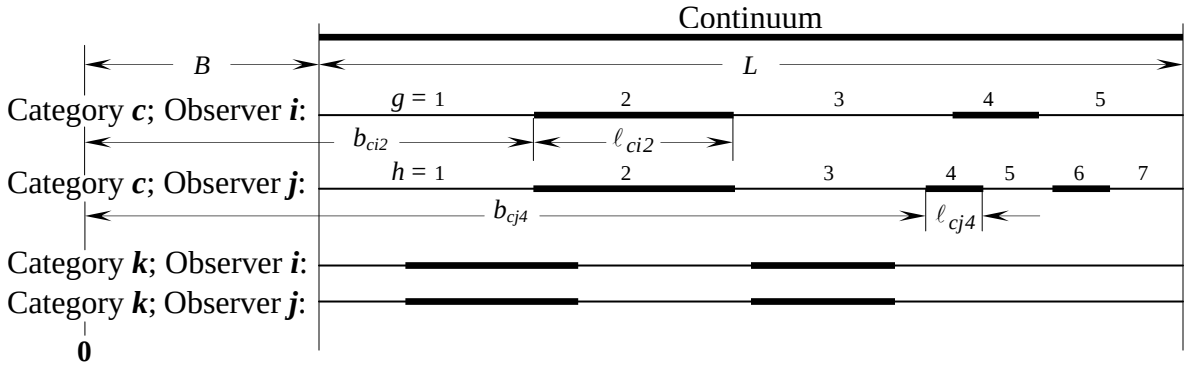
Usually, units are interpreted, assigned to categories, or coded. One unit may be assigned to any number of categories c or k. However, for any one observer, units of the same category may not overlap.

4.3 Observers, Coders, Unitizers, Analysts

There are any number m of observers, coders, unitizers, or analysts, at least two. They are generically referred to by i and j.

4.4 Numbering of Sections

The sections that an observer identifies as units or as gaps between units are consecutively numbered, separately for each observer and for each category. For category c, observer i's sections are referred to by g and observer j's sections are referred to by h. These are indicated by subscripts <cig>, <cjh>, <kig>, <kjh>, etc. of the beginnings b and lengths ℓ of sections in the continuum.



Specification of a Continuum and Location of Units Within It
Figure 3

4.5 Differentiation of Units and Gaps

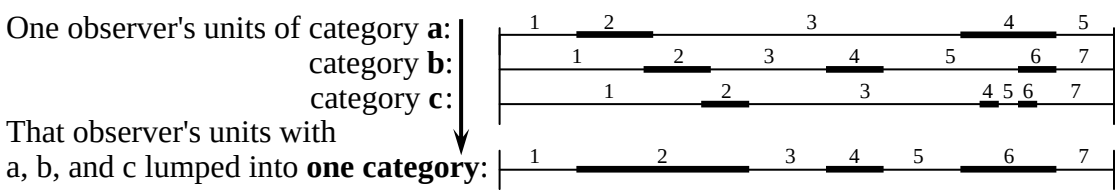
Whether a section of the continuum is a unit or a gap between two units, between the beginning of the continuum and one unit or one unit and the end of that continuum is indicated by a binary function v_{cig} of such sections:

$$v_{cig} = \begin{cases} 0 & \text{iff section <cig> is a gap} \\ 1 & \text{iff section <cig> is an identified unit} \end{cases}$$

5 Recoding of Categories f

Qualitative text analysts often apply higher-order categories to lower-order categories, creating conceptual hierarchies. Reliability may have to be evaluated on each level separately. But regardless of such multi-level categorization, reliability analysts may create their own hierarchies by applying a mapping, **c'=f(c)** to the lowest level categories, in effect allowing for:

- **Omitting categories** of units from the computed reliabilities and obtaining reliabilities that are conditional on the omitted ones
- **Lumping several categories into one** in which case the units of the lumped categories collapse into their set theoretical unions. For example:



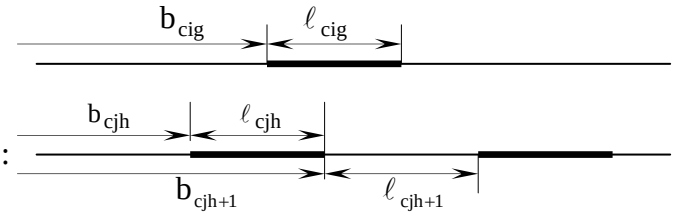
Collapsing of Units of Different Categories into One
Figure 4

6 Difference Function δ_{cigjh}^2

For reliability to be perfect, the units that different observers identify must occupy the same locations in the continuum and be assigned to identical categories. Disagreements sum deviations from this ideal by counting the pair wise differences between units and gaps, one pair at a time. Intuitively, such differences must be zero when units perfectly coincide. They must increase as the overlap between any two units lessens and reach their largest value when a unit does not overlap with any other unit. In terms of the following lengths:

Category c; Observer i; unit g:

Category c; Observer j; unit h; gap h+1:



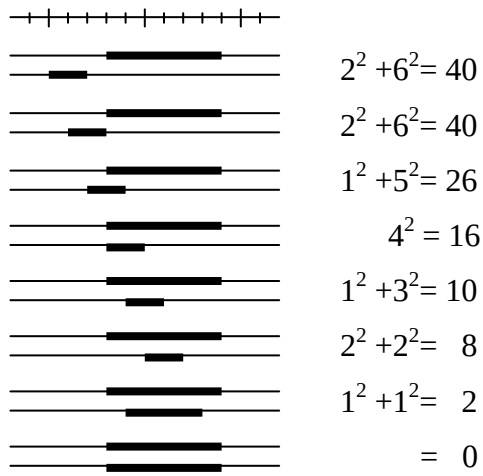
The squared difference δ^2_{cigjh} between any two observers' sections g and h, all of the same category c, is defined by this function:

$$\delta^2_{cigjh} = \begin{cases} (b_{cig} - b_{cjh})^2 + (b_{cig} + \ell_{cig} - b_{cjh} - \ell_{cjh})^2 & \text{iff } v_{cig}=v_{cjh}=1 \text{ and } -\ell_{cig} < b_{cig} - b_{cjh} < \ell_{cjh} \\ \ell^2_{cig} & \text{iff } v_{cig}=1,\ v_{cjh}=0 \text{ and } \ell_{cjh} - \ell_{cig} \geq b_{cig} - b_{cjh} \geq 0 \\ \ell^2_{cjh} & \text{iff } v_{cig}=0,\ v_{cjh}=1 \text{ and } \ell_{cjh} - \ell_{cig} \leq b_{cig} - b_{cjh} \leq 0 \\ 0 & \text{Otherwise} \end{cases}$$

The first condition pertains to pairs of overlapping units. Here, δ^2 sums the squares of the two non-overlapping lengths. The second condition applies when observer i's unit g is fully contained in observer j's gap h. The third condition is the converse of the second. It applies when observer i's gap g fully contains observer j's unit h. The fourth condition applies when two sections of the continuum overlap perfectly; are both gaps, not units; or have no relation with each other in the continuum.

To see how this function behaves in response to different degrees of overlap between two observers' units, consider these examples:

Units of Length:



Sliding Differences Between Units
Figure 5

7 Observed Disagreement D_{oc}

Just as in the definition of the observed Disagreement D_o of the established family of α-measures, the observed disagreement D_{oc} between all pairs of units of the same category c is obtained by comparing each observer's sections with all other observers' sections on the continuum, applying the above difference function. This systematic comparison gives rise to the measure of the observed disagreement:

$$D_{oc} = \frac{\sum_{i=1}^{m} \sum_{g} \sum_{j=1|j\neq i}^{m} \sum_{h} \delta^2_{cigjh}}{m(m-1)L^2}$$

wherein

m is the number of observers that unitize the continuum,

m(m–1) is the number of pairs of observers whose units are being compared,

L is the length of the continuum, and

$\delta^2_{\mathbf{cigjh}}$ is the difference between two sections <cig> and <cjh>, $\delta^2_{cigjh} = \delta^2_{cjhig}$.

$\mathbf{D_{oc}}$ is the observed disagreement in category c, which can be seen as an average difference.

Note that the sums in D_{oc} pair all observers i and j (but not with itself) and run through all pairs of sections in one category.

8 Expected Disagreement D_{ec}

The expected disagreement was difficult to derive. It amounts to a virtual generation of all possible unitizations, using only the actually identified units and gaps of a particular category, comparing each with each other, and applying the disagreement measure to each possible pair of unitizations. Just as the expected disagreement for coding is the disagreement without consideration of the units that were coded, the expected disagreement for unitizing is the disagreement without consideration of the location of the sections that should ideally match.

With the number of units of category c that all m observers have identified:

$N_c = \sum_{i=1}^{m} \sum_{g} v_{cig}$ = The total number of units of category c identified by all m observers

the expected disagreement for units in category c is:

$$D_{ec} = \frac{\frac{2}{L}\sum_{i=1}^{m}\sum_{g} v_{cig}\left[\frac{N_c - 1}{3}\left(2\ell_{cig}^3 - 3\ell_{cig}^2 + \ell_{cig}\right) + \ell_{cig}^2 \sum_{j=1}^{m}\sum_{h}\left(1 - v_{cjh}\right)\left(\ell_{cjh} - \ell_{cig} + 1\right) \text{ iff } \ell_{cjh} \geq \ell_{cig}\right]}{mL(mL-1) - \sum_{i=1}^{m}\sum_{g} v_{cig}\ell_{cig}(\ell_{cig} - 1)}$$

Its proof is lengthy and provided elsewhere (Krippendorff, 1995). Let me merely point out its principal components. The first double summation in its enumerator goes through all observers' units, which v_{cig} separates from the gaps between them. The first expression in the angular parenthesis accounts for the differences between one unit and all other units overlapping with that unit in all possible ways. The double summation in the angular parenthesis goes through all gaps between units, adding the differences due to that unit falling within all possible gaps in all possible ways. In the denominator, mL is the number of possible locations for a unit to occur in the continuum and mL(mL–1) is the number of pair comparisons of such units that the disagreement measure calculates virtually in this expression.

9 α-Agreement for One of Several Categories or Interpretations

With the observed and expected disagreements for one category c now in place and following the definition of α in our family of agreement measures, the α-agreement for a variably unitized continuum of one category c is:

$$\alpha_c = 1 - \frac{D_{oc}}{D_{ec}}$$

10 α-Agreement For Multiple Categories or Interpretations

To obtain the agreement for multiple categorizations/interpretations we aggregate the disagreements for all categories as in:

$$\alpha = 1 - \frac{\sum_c D_{oc}}{\sum_c D_{ec}}$$

The proof of this form follows the derivation of the reliability measure for unitizing (Krippendorff, 1995), which did not recognize multiple categorizations. Although high reliabilities in one category may compensate for low reliability in another categories, consider the effects that different kinds of confusions have on these disagreement measures. For example, suppose two observers agree on the location of a unit but assign it to different categories. This confusion would be

registered in the observed and expected disagreements for both categories. But suppose the two observers assign two units of the same length to the same category, but they do not occupy the same positions on the continuum. They overlap. This confusion would only minimally, if at all, affect the expected disagreement of this category but be registered by its observed disagreement -- the amount of this increase dependents on the degree to which these units overlap.

An obvious possibility of this form of α is to obtain agreement coefficients for various subsets of categories. If one category or interpretation turns out to be consistently unreliable or uncertain, researchers are informed by how much the reliability of data increases when the unreliable category is excluded from an analysis.

11 α-Agreement for Recoded Categories or Interpretations

When categories of units are recoded, α is computed for the transformed data, ignoring units whose categories are excluded and collapsing units whose categories are lumped into their set theoretical unions:

$$\alpha = 1 - \frac{\sum_{c'} D_{o,f(c)}}{\sum_{c'} D_{e,f(c)}}$$

In the extreme this recoding option enables analysts to calculate α-reliabilities for data in which all categories are collapsed into one, the common quality of units being identified as relevant. The difference between the reliability for all categories and for all categories collapsed into one (the reliability of identifying relevant matter) indicates how much categorizing adds to or distracts from the reliability of mere unitizing.

12 Summary

This paper suggests a computational solution to the problem of evaluating the reliability of variably unitized and multiply categorized or interpreted textual matter, video recordings, group interactions, and the like, all of which start out as undifferentiated continua until researchers draw distinctions within them. The proposal extends the family of α-agreement coefficients and brings to qualitative research standards that are acceptable elsewhere. Qualitative text analysts often consider the lack of reliability measures in their empirical domain as indicative of the fundamental difference between qualitative and quantitative research. This justification is no longer valid. Although the above proposal does not solve all problems of assessing the reliability of qualitative data, it shows its

possibility and its feasibility. I argue that even multiple interpretations of textual matter need to be reliable in the sense of being replicable by other researchers or described same or similarly by independent analysts of the same continuum. Actually, reliability considerations should not be entirely strange in qualitative research. Consider that qualitative researchers customarily accept some interpretations as valid and others as disagreeable, unacceptable, or without adequate ground. The above merely gives such judgments an explicit face and offers ways the trustworthiness of different data making processes may be compared. I would contend that addressing reliability questions is essential to improve the credibility of qualitative research.

The computation of these reliabilities are not simple indeed. When the volume of textual data is large and unitizing and coding is complex, reliabilities can no longer be calculated by hand. A computer program for calculating these is currently being developed. But the use of computer-aided text analysis software brings a computable precision to qualitative research that heretofore was unavailable to traditional qualitative researchers. The above solution can easily be incorporated in such software and made widely available. It would enable qualitative researchers to keep track of their unreliabilities as routinely as they currently use spell checkers. In my experience, information about the reliability of one's work is informative to the coder or analysts and the use of such quality checks encourages a more responsible use of qualitative data in the social sciences.

Appendix

A Numerical Example

The numerical values of the units depicted in the **reliability data cube** of Figure 2 are:

Continuum	**B**	**L**	
	150	300	
Sections	**b**	ℓ	**v**
ci1	150	75	0
ci2	225	70	1
ci3	295	75	0
ci4	370	30	1
ci5	400	50	0
cj1	150	70	0
cj2	220	80	1
cj3	300	55	0
cj4	355	20	1
cj5	375	25	0
cj6	400	20	1
cj7	420	30	0
ki1	150	30	0
ki2	180	60	1
ki3	240	60	0
ki4	300	50	1
ki5	350	100	0
kj1	150	30	0
kj2	180	60	1
kj3	240	60	0
kj4	300	50	1
kj5	350	100	0

The non-zero **differences** between the two observers' sections in category c are:

$$\delta^2_{ci2j2} = (225-220)^2 + (225+70-220-80)^2 = 5^2 + 5^2 = 50 = \delta^2_{cj2i2}$$

$$\delta^2_{ci4j4} = (370-355)^2 + (370+30-355-20)^2 = 15^2 + 25^2 = 850 = \delta^2_{cj4i4}$$

$$\delta^2_{ci5j6} = 20^2 = 400 = \delta^2_{cj6i5}$$

Evidently, the first pair of units, showing observer i as merely a bit more conservative than j is, contributes very little by comparison with the remaining three units, which are more scattered on the continuum. In category k all differences are zero:

$$\delta^2_{ki2j2} = 0 = \delta^2_{ki4j4}$$

The **observed disagreement** in category c is:

$$D_{oc} = \frac{\delta^2_{ci2j2} + \delta^2_{ci4j4} + \delta^2_{ci5j6} + \delta^2_{cj2i2} + \delta^2_{cj4i4} + \delta^2_{cj6i5}}{m(m-1)L^2} = \frac{2(50 + 850 + 400)}{2(2-1)300^2} = .0144$$

The observed disagreement in category k, D_{ok} = .0000, of course.

Calculating the **expected disagreement** with the above formula requires many more steps. In category c, with a total of N_c=2+3=5 identified units, the expected disagreement becomes:

$$D_{ec} = \frac{\frac{2}{300}\left\{\begin{array}{l} \left[\frac{5-1}{3}(2\cdot 70^3 - 3\cdot 70^2 + 70) + 70^2\begin{pmatrix} 75-70+1 \\ +75-70+1 \\ +70-70+1 \end{pmatrix}\right] \\ +\left[\frac{5-1}{3}(2\cdot 30^3 - 3\cdot 30^2 + 30) + 30^2\begin{pmatrix} 75-30+1 \\ +75-30+1 \\ +50-30+1 \\ +70-30+1 \\ +55-30+1 \\ +30-30+1 \end{pmatrix}\right] \\ +\left[\frac{5-1}{3}(2\cdot 80^3 - 3\cdot 80^2 + 80)\right] \\ +\left[\frac{5-1}{3}(2\cdot 20^3 - 3\cdot 20^2 + 20) + 20^2\begin{pmatrix} 75-20+1 \\ +75-20+1 \\ +50-20+1 \\ +70-20+1 \\ +55-20+1 \\ +25-20+1 \\ +30-20+1 \end{pmatrix}\right] \\ +\left[\frac{5-1}{3}(2\cdot 20^3 - 3\cdot 20^2 + 20) + 20^2\begin{pmatrix} 75-20+1 \\ +75-20+1 \\ +50-20+1 \\ +70-20+1 \\ +55-20+1 \\ +25-20+1 \\ +30-20+1 \end{pmatrix}\right] \end{array}\right\}}{2\cdot 300(2\cdot 300\text{-}1)\text{-}\begin{pmatrix} 70(70-1) \\ +30(30-1) \\ +80(80-1) \\ +20(20-1) \\ +20(20-1) \end{pmatrix}} = .0532$$

And in category k, with a total of N_k=2+2=4 identified units, the expected disagreement becomes:

$$D_{ek} = \frac{\frac{2}{300}\left\{ \begin{array}{l} + \left[\frac{4-1}{3}(2\cdot 60^3 - 3\cdot 60^2 + 60) + 60^2 \begin{pmatrix} 60-60+1 \\ +100-60+1 \\ +\ \ 60-60+1 \\ +100-60+1 \end{pmatrix} \right] \\ + \left[\frac{4-1}{3}(2\cdot 50^3 - 3\cdot 50^2 + 50) + 50^2 \begin{pmatrix} 60-50+1 \\ +100-50+1 \\ +\ \ 60-50+1 \\ +100-50+1 \end{pmatrix} \right] \\ + \left[\frac{4-1}{3}(2\cdot 60^3 - 3\cdot 60^2 + 60) + 60^2 \begin{pmatrix} 60-60+1 \\ +100-60+1 \\ +\ \ 60-60+1 \\ +100-60+1 \end{pmatrix} \right] \\ + \left[\frac{4-1}{3}(2\cdot 50^3 - 3\cdot 50^2 + 50) + 50^2 \begin{pmatrix} 60-50+1 \\ +100-50+1 \\ +\ \ 60-50+1 \\ +100-50+1 \end{pmatrix} \right] \end{array} \right\}}{2\cdot 300(2\cdot 300-1) - \begin{pmatrix} 60(60-1) \\ +50(50-1) \\ +60(60-1) \\ +50(50-1) \end{pmatrix}} = .0490$$

The **α-reliability for one category** – category c and k separately -- is:

$$\alpha_c = 1 - \frac{.0144}{.0532} = .7293$$

$$\alpha_k = 1 - \frac{.0000}{.0490} = 1.0000\,,$$

and the **α-reliability for categories** c and k jointly, is:

$$\alpha = 1 - \frac{.0144 + .0000}{.0532 + .0490} = .8591$$

References

Brennan, R. L. (2001). An Essay on the History and Future of Reliability from the Perspective of Replications. *Journal of Educational Measurement* 38,4: 295-317.

Gerbner, G. (with M. Brouwer, C. Clark, & K. Krippendorff) (1972). Violence in Television Drama: Trends and Symbolic Functions, Appendix B: Analytical procedures. In: G. A. Comstock & E. A. Rubinstein (eds.) *Television and Social Behavior*. Rockville, MD: National Institute of Mental Health, pp. 167-170.

Krippendorff, K. (1970). Bivariate Agreement Coefficients for Reliability Data. Chapter 8. In: E. R. Borgatta & G. W. Bohrnstedt (eds.) *Sociological Methodology 1970*, Vol. 2. San Francisco CA: Jossey Bass, Inc., pp. 139-150.

--- (1980). Reliability, Chapter 12. In: K. Krippendorff, *Content Analysis; An Introduction to its Methodology*. Beverly Hills, CA: Sage Publications, pp. 129-154.

--- (1987). Association, Agreement and Equity. *Quality & Quantity 21*: 109-123.

--- (1992). Recent Developments in Reliability Analysis. Paper presented to the International Communication Association (ICA) Conference in Miami, FL: May 25, 1992. Available as ED 352390 at ERIC Clearinghouse, Springfield, Virginia and through DIALOG, BRS, or SDC.

--- (1995). On the Reliability of Unitizing Continuous Data. Chapter 2. In: P. V. Marsden (ed.) *Sociological Methodology, 1995*, Vol. 25. Cambridge, MA: Blackwell, pp. 47-76.

--- (2004). Reliability, Chapter 11. In: K. Krippendorff, *Content Analysis; An Introduction to its Methodology, 2nd Edition*. Thousand Oaks, CA: Sage Publications, pp. 211-256.

Pearson, K. (1901). Mathematical Contributions to the Theory of Evolution IX: On the Principle of Homotyposis and its Relation to Heredity, to Variability of the Individual, and to that of Race. Part I: Homotyposis in the Vegetable Kingdom. *Philosophical Transactions of the Royal Society* (London) A 193: 358-479.

Scott, W. A. (1955). Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly* 19: 321-325.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Tildesley, M. L. (1921). A First Study of the Burmese Skull. *Biometrica* 13: 176-267.