Appendix
Inter-Coder Agreement in ATLAS.ti [i]

Klaus Krippendorff
Klaus.krippendorff@asc.upenn.edu

What is reliability?

Reliability is the ability to rely on something, here on data generated by coding within ATLAS.ti for representing textual phenomena, their readings, of analytical interest, and only these.

Concerns with the reliability of data are **motivated** by the experience that unreliable data reduce the chance of their analysis to lead to valid conclusions; introduce uncertainty for researchers to know what they are analyzing, and make it difficult for other scholars, critics, and stakeholders of said phenomena to interpret or build on the published findings.

Reliability can be experienced only when the use of data did not lead to failures. Prior to such experiences, the reliability of data **needs to be inferred** from observable conditions that are known to reduce failures.

Three **kinds of reliabilities** can be distinguished by the sources of unreliability they respectively capture:

- **Stability of one coder** declines when that coder confuses given codes, use them inconsistently over time, or is unable to repeat the process of generating data
- **Replicability of the coding instructions** declines not only with intra-coder instabilities but also with inter-coder disagreements among several coders who interpret and apply them independent of each other to the same set of phenomena. Replicability has to be immune to all irrelevant influences, whether unequal coder qualifications, unlike literary competences, different recording instruments or variations in the circumstances of replication, and times
- **Accuracy of the coded data** refers to the correspondence of coding by one or more coders with an accepted standard. It declines with intra-coder instabilities, inter-coder disagreements, and shared coder preferences, biases, or prejudices.

Evidence of replicability is stronger than evidence of intra-coder stabilities but weaker than evidence of accuracy. However, standards for the coding of data are rarely available and when they are, coding efforts would be mute, except for testing small subsamples of reliability data. Therefore, replicability is the reliability measure of choice.

What data are needed to infer replicability?

Data that give rise to inter-coder agreements from which the replicability of a population of data can be inferred

- Have to **replicate** the very coding process on a sample of phenomena to be converted into reliable data, using different coders who apply identical coding instructions to the same set of phenomena of analytical interest

- Must be **informed by written coding instructions**, and only these
- The **sample** (volume of textual matter including videos) of phenomena to be coded must be large enough to represent the diversity of the phenomena of analytical interest
- **Coder qualifications** must be **sufficiently common** for coders to be freely replaceable.
- The **number of coders** employed needs to embrace various analysts' and diverse stakeholders' ability to understand the phenomena studied through the coding instructions applied. Two coders may not suffice
- **Coders must work independent of each other** and not communicate about their coding task
- Any **preparatory training** that coders received and the qualifications for which they were selected need to be **communicable for replication elsewhere.**

Deviations from these conditions tend to pollute the reliability data and inflate the observable inter-coder agreement, leading to mistaken assurances of their replicability. For example, selecting coders among close acquaintances, with a stake in the outcome of a study, receiving thorough but undocumented training, allowing them to discuss how to interpret given coding instructions, or settling emerging uncertainties by consensus, yield deceptively higher inter-coder agreements which is no longer indicative of the replicability of the generated data.

Perfect replicability means that data have the potential of leading to valid answers of given research questions, analysts can use the communicable coding instructions in reverse, to decode what their data represent, and the stakeholders in a research project can critically evaluate the analysis, talk about, or respond to the phenomena studied and build on published findings.

Reliability data in ATLAS.ti:

After a principal investigator has developed suitable **coding instructions** in writing, without references to the textual matter to which they are to be applied, independently working coders need to apply these coding instructions to the same set of phenomena and return comparable data to the principal investigator. Although ATLAS.ti cannot prevent coders from introducing new codes, they will have to be ignored except as suggestions to improve the coding instructions for subsequent coding efforts.

Coding instructions must specify

- **The set of** relevant and logically or conceptually separate **semantic domains** with definitions and examples made readily available to coders.

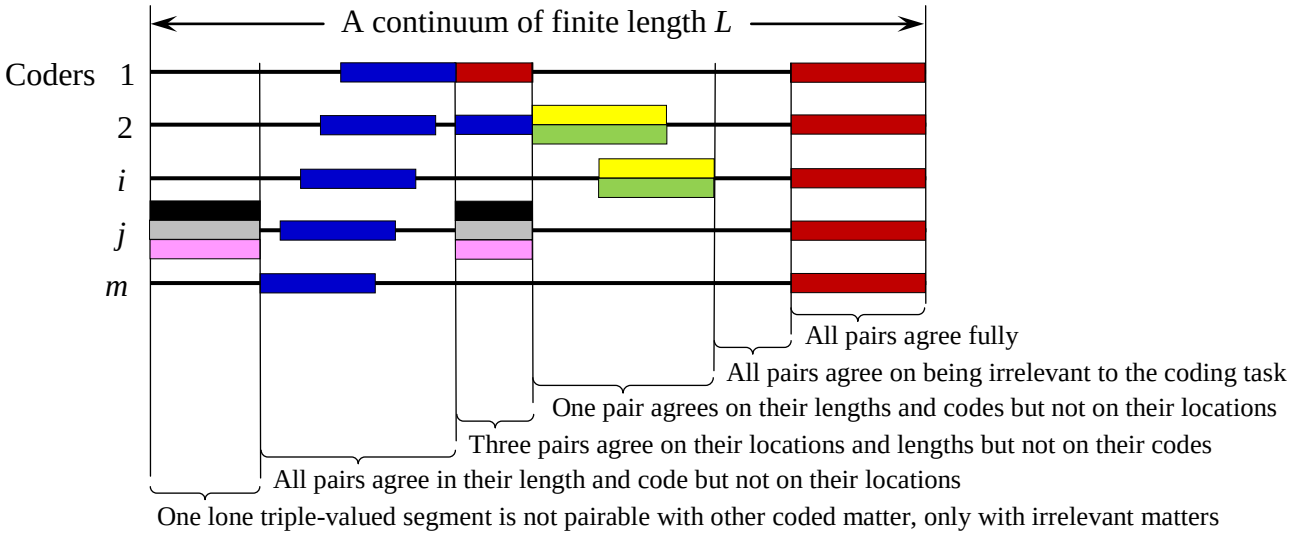
  A semantic domain names a space of distinct concepts with shared meanings, e.g., "colors," "mental illnesses," "emotions," "gender issues," or "personalities." Semantic domains may be named abstractly but are always context dependent. The concept of color is different when applied to the sky, a dress, a national flag, an ethnic group, or the state of a drunk. The gender of nouns is unlike the gender of living organisms. The contexts of semantic domains need to be preserved when coding texts. A single quote typically invokes several connected semantic domains. For example:

  - "[ ] says [ ] to [ ] intending to [ ] but causing [ ]" defines the semantic domains of speakers, utterances, addressees, and intended and unintended consequences.
  - "[ ] diagnoses [ ] to have [ ]."

If the first semantic domain concerns medical professionals, the second concerns patients and the third illnesses. If the first is a car mechanic, the other two relate to cars.

Most semantic domains concern attributes of objects, actions, people, or abstract ideas.

- **Each semantic domain contains a set of mutually exclusive codes** (at least one) by name with definitions and examples made readily available to coders
- **Coders highlight or identify segments of a given textual continuum**, e.g., quotes, propositions, or paragraphs and, following the written coding instructions, assign one code from each applicable semantic domain to them. For example, with colors identifying codes from separate semantic domains:



Definitions of terms

Coders: 1, 2, …, *i*, …, *j*, … *m*

Segments:

Coder *i*'s segments: $S_{i1}, S_{i2}, \ldots S_{ig}, S_{ig+1}, \ldots, S_{i\text{last for } i}$

Coder *j*'s segments: $S_{j1}, S_{j2}, \ldots S_{ih}, S_{ih+1}, \ldots, S_{j\text{last for } j}$

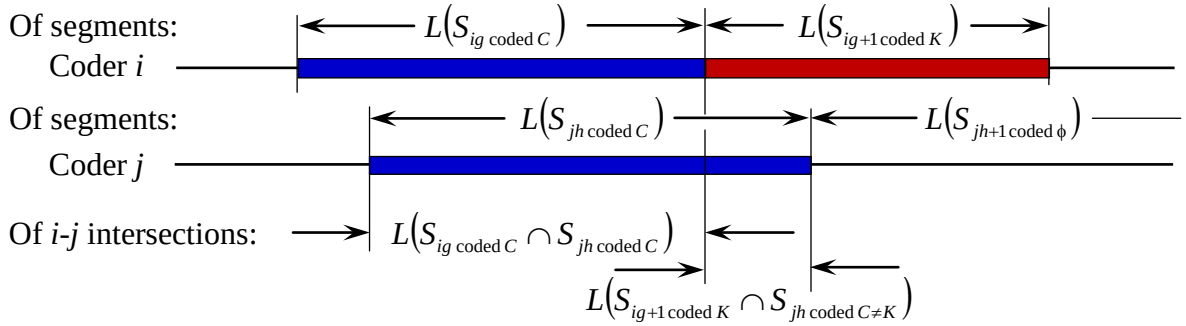
Coding of segments:

multi-valued sets C of codes $c \in C$ or K of codes $k \in K$

single-valued codes $C=c$ or $K=k$

uncoded matter, designated by $C=\phi$ or $K=\phi$

Lengths (in terms of the number of characters for texts or number of seconds for videos):



Of the continuum: $L = \sum_{g=1}^{\text{last for } i} L(S_{ig}) = \sum_{h=1}^{\text{last for } j} L(S_{jh}) L$

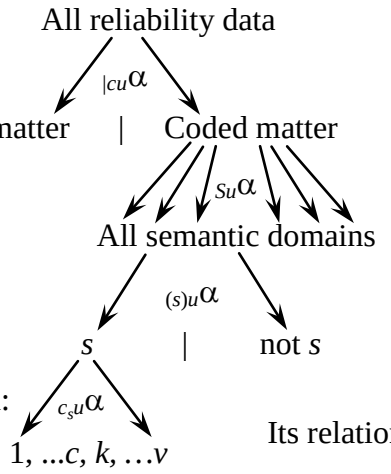
Differences:

With the number of elements in sets referred to as its cardinality $|C|$, differences

Between two sets C and K of codes: $\Delta_{CK} = |C||K - |C \cap K|^2$

Between two single codes c and k: $\Delta_{ck} =_{nominal} \delta^2_{ck} = \begin{cases} 0 \text{ iff } c = k \\ 1 \text{ iff } c \neq k \end{cases}$

ATLAS.ti provides four measures of inter-coder agreement (version 8.4):



(1) On relevant/irrelevant matter: A binary measure

(2) On all semantic domains: A multi-valued measure

(3) On one semantic domains: A binary measure

(4) On coding within that domain: A single-valued measure
Its relation to $_{C_s u}\alpha$ is discussed in (4) and (5)

$_{|cu}\alpha$ indicates the extent to which coders agree on the relevance of texts for the research project,

$_{Su}\alpha$ indicates the extent to which coders agree on the presence or absence of sematic domains,

$_{(s)u}\alpha$ indicates the degree to which coders identify a particular semantic domain *s*,

$_{c_s u}\alpha$ indicates the agreement on coding within a semantic domain *s*. When coders pollute the reliability data by intra-coder confusions, the multi-valued $_{C_s u}\alpha$ is computed. It inflates $_{c_s u}\alpha$ but can serve as an invitation to explore the source of that confusion, see (5).

Note: In version 8.3 only the first three measures are available, currently referred to as Alpha binary, and Cu-alpha / cu-alpha.

The binary **agreement on distinguishing relevant + from irrelevant ϕ matter**

(Simplified by assuming all relevant matter as finely grained):

$_{|cu}\alpha$'s coincidences are:

$$\ell_{\phi+} = \frac{1}{m-1}\sum_i \sum_{j\neq i}^{m} \sum_{g,h} L\left(S_{ig \text{ coded } \phi} \cap S_{jh \text{ coded } \neq\phi}\right)$$

$$\ell_{\phi\cdot} = \frac{1}{m-1}\sum_i^m \sum_g L\left(S_{ig \text{ coded } \phi}\right)$$

$$\ell_{\cdot+} = \frac{1}{m-1}\sum_i^m \sum_g L\left(S_{ig \text{ coded } \neq\phi}\right)$$

Its coincidence matrix is:

	ϕ	+	
ϕ	$\ell_{\phi\phi}$	$\ell_{\phi+}$	$\ell_{\phi\cdot}$
+	$\ell_{+\phi}$	ℓ_{++}	$\ell_{\phi\cdot}$
	$\ell_{\cdot\phi}$	$\ell_{\cdot+}$	$\ell_{\cdot\cdot} = mL$

Its disagreements are: $_{|cu}D_o = \frac{\ell_{\phi+} + \ell_{+\phi}}{\ell_{\cdot\cdot}}$ and $_{|cu}D_e = \frac{\ell_{\phi\cdot}\ell_{\cdot+} + \ell_{+\cdot}\ell_{\cdot\phi}}{\ell_{\cdot\cdot}\left(\ell_{\cdot\cdot}-1\right)}$

The binary $_{|cu}\alpha$**-agreement** is: $_{|cu}\alpha = 1 - \frac{_{|cu}D_o}{_{|cu}D_e} = 1 - \left(\ell_{\cdot\cdot}-1\right)\frac{\ell_{\phi+}}{\ell_{\phi\cdot}\ell_{\cdot+}}$

Where: *u* stands for unitizing, *c* for single-valued coding, and | for its binary nature.

(1) The **agreement on recognizing diverse semantic domains:**

Any segment may be described in terms of several semantic domains. Being logically or conceptually independent of each other, joint descriptions constitute multi-valued coding.

$_{Su}\alpha$ assesses the agreement on recognizing semantic domains (not their distinct codes) within relevant matter. Accordingly, *C* and *K* are the sets of semantic domains applicable to segments of the textual continuum.

Its coincidences are: $\ell_{CK} = \frac{1}{m-1}\sum_i \sum_{j\neq i}^{m} \sum_{g,h} L\left(S_{ig \text{ coded } C\neq\phi} \cap S_{jh \text{ coded } K\neq\phi}\right)$

Its coincidence matrix is:

	1	2	…	*K*	*v*	
1	ℓ_{11}	ℓ_{12}	…	ℓ_{1K}	ℓ_{1v}	$\ell_{1\cdot}$
2	ℓ_{21}	ℓ_{22}	…	ℓ_{2K}	ℓ_{2v}	$\ell_{2\cdot}$
C	ℓ_{C1}	ℓ_{C2}	…	ℓ_{CK}	ℓ_{Cv}	$\ell_{C\cdot} = \sum_{K=1}^{v} \ell_{CK}$
:	:	:	:::	:	:	:
v	ℓ_{v1}	ℓ_{v2}	…	ℓ_{vK}	ℓ_{vv}	$\ell_{v\cdot}$
	$\ell_{\cdot1}$	$\ell_{\cdot2}$	…	$\ell_{\cdot K}$	$\ell_{\cdot v}$	$\ell_{\cdot\cdot} = \sum_{C=1}^{v}\sum_{K=1}^{v} \ell_{CK}$

Matching coincidences occupy its diagonal: $\ell_{11}, \ell_{22}, \ldots \ell_{CC}, \ldots \ell_{KK}, \ldots \ell_{vv}$.

Mismatching coincidences are located in its off-diagonal triangles: $\ell_{CK} = \ell_{KC}$.

Differences between two sets C and K are: $\Delta_{CK} = |C||K - |C \cap K|^2$

Its disagreements are: $$_{Su}D_o = \frac{\sum_C \sum_K \square_{CK} \Delta_{CK}}{\sum_C \sum_K \square_{CK} |C||K|}$$

and: $$_{Su}D_e = \frac{\sum_C \square_C \cdot \sum_K \square_{\cdot K} \Delta_{CK}}{(\sum_C \square_C \cdot |C|)^2 - \sum_i^m \sum_g \left(L(S_{ig \text{ coded } C \neq \phi})\right)^2 |C|}$$.

The **$_{Su}\alpha$-agreement** is: $$_{Su}\alpha = 1 - \frac{_{Su}D_o}{_{Su}D_e}$$

Where the capital S stands for multi-valued sets of references uses of semantic domains.

(2) The **agreement of identifying the applicability of any one semantic domain *s***:

A chosen semantic domain s may or may not occur in a set of semantic domains, $s \in C$ of the coincidences generated in (2): $$\square_{CK} = \frac{1}{m-1} \sum_i \sum_{j \neq i}^m \sum_{g.h} L\left(S_{ig \text{ coded } C \neq \phi} \cap S_{jh \text{ coded } K \neq \phi}\right)$$

Its disagreements are: $$_{(s)u}D_o = \frac{\sum_s \sum_K \sum_{k \in K} \square_{CK} \text{ iff } s \in C \text{ and } k \notin C}{\sum_s \sum_K \sum_{k \in K} \square_{CK} \text{ iff } s \in C}$$

and from (2): $$_{Su}D_e = \frac{\sum_C \square_C \cdot \sum_K \square_{\cdot K} \Delta_{CK}}{(\sum_C \square_C \cdot |C|)^2 - \sum_i^m \sum_g \left(L(S_{ig \text{ coded } C})\right)^2 |C|}$$

The **$_{(s)u}\alpha$-agreement** is: $$_{(s)u}\alpha = 1 - \frac{_{(s)u}D_o}{_{Su}D_e}$$

Where (s) denotes the semantic domain singled out for attention.

(3) The **agreement on the mutually exclusive codes *c* and *k* of a chosen semantic domain *s***:

$_{cSu}\alpha$, which is to assess the agreement on the single-valued coding within a chosen semantic domain s is not applicable when coding violates the mutual exclusiveness of codes within that domain. This may occur when coding instructions are ambiguous, or coders feel unable to distinguish between codes. Instead, ATLAS.ti computes the multi-valued $_{CSu}\alpha$-agreement, following (2), but applied to segments described in terms of the semantic domain s.

$_{CSu}\alpha = {}_{cSu}\alpha$ when data are coded correctly. $_{CSu}\alpha$ inflates $_{cSu}\alpha$ when reliability data are infested by intra-coder confusions. To signal that inflation, the computed $_{CSu}\alpha$-value is shown in red, not to be cited or acted upon. It merely invites to explore the source of that confusion, facilitated by (5).

$_{C_s u}\alpha$'s coincidences within *s* are: $\square_{C_s K_s} = \frac{1}{m-1}\sum_i^m \sum_{j\neq i}^m \sum_{g,h} L\left(S_{ig \text{ coded } C_s \neq \phi} \cap S_{jh \text{ coded } K_s \neq \phi}\right)$

Its disagreements are: $_{C_s u}D_o = \frac{\sum_{C_s}\sum_{K_s} \square_{C_s K_s} \Delta_{C_s K_s}}{\sum_{C_s}\sum_{K_s} \square_{C_s K_s} |C_s||K_s|}$

and: $_{C_s u}D_e = \frac{\sum_{C_s} \square_{C_s \cdot} \sum_{K_s} \square_{\cdot K_s} \Delta_{C_s K_s}}{(\sum_{C_s} \square_{C_s \cdot} |C_s|)^2 - \sum_i^m \sum_g \left(L(S_{ig \text{ coded } C_s \neq \phi})\right)^2 |C_s|}$

The ***$_{C_s u}\alpha$*-agreement** is: $_{C_s u}\alpha = 1 - \frac{_{C_s u}D_o}{_{C_s u}D_e} \geq {}_{c_s u}\alpha$ – For testing the equality $_{c_s u}\alpha = {}_{C_s u}\alpha$, see (5).

Where c_s stands for single-valued coding of segments within the semantic domain *s*

Users who wish to explore the reasons for $_{C_s u}\alpha$'s value may examine its coincidence matrix:

	1st	2nd	…	*K*th	*v*th
1st set of codes by names	ℓ_{11}	ℓ_{12}	…	ℓ_{1K}	ℓ_{1v}
2nd set of codes by names	ℓ_{21}	ℓ_{22}	…	ℓ_{2K}	ℓ_{2v}
*C*th set of codes by names	ℓ_{C1}	ℓ_{C2}	…	ℓ_{CK}	ℓ_{Cv}
:	:	:	:::	:	:
*v*th set of codes by names	ℓ_{v1}	ℓ_{v2}	…	ℓ_{vK}	ℓ_{vv}

Note: When coding does not violate the mutual exclusiveness of codes in a semantic domain, as tested in (4), $_{c_s u}\alpha = {}_{C_s u}\alpha$ is accurate and rows and columns are defined by single codes.

(4) Test whether all segments of the domain *s* are coded by single-valued sets, $|C_s| = 1$ (not yet available in version 8.3)

- When this test is positive: ${}_{C_s u}\alpha = {}_{c_s u}\alpha$ and its values are listed in black.
- When this test fails, the computed ${}_{C_s u}\alpha$ inflates the correct ${}_{c_s u}\alpha$ and is highlighted red.
- For each coder *i* who fails this test a confusion matrix is available for inspection, containing the sum of all of *i*'s confusions:

$$\wp_{ck} = \sum_g \frac{1}{|C_s|(|C_s|-1)} \sum_{c \in C_s} \sum_{k \in C_s,\ k \neq c} L\left(S_{ig \text{ coded } C_s \neq \phi}\right) \text{ if } |C_s| > 1$$

This matrix does not show where *i* correctly uses single codes. Its entries are the very confusions which are missing from ${}_{C_s u}\alpha$'s observed disagreement, hence its inflation.

	1st	2nd	…	*k*th	*v*th
1st code by name	$\wp_{11}$	$\wp_{12}$	…	$\wp_{1k}$	$\wp_{1v}$
2nd code by name	$\wp_{21}$	$\wp_{22}$	…	$\wp_{2k}$	$\wp_{2v}$
*c*th code by name	$\wp_{c1}$	$\wp_{c2}$	…	$\wp_{ck}$	$\wp_{cv}$
:	:	:	:::	:	:
*v*th code by name	$\wp_{v1}$	$\wp_{v2}$	…	$\wp_{vk}$	$\wp_{vv}$

Accounting for the effects of this confusion amounts to an increase in the numerator of CsuDo by what it falsely takes to be a perfect match between two multi-valued sets |Cs| > 1.

For one $|C_s| > 1$ in ${}_{C_s u}\alpha$, $\square_{C_s C_s} \Delta_{CC} = 0$

For one $|C_s| > 1$ in ${}_{c_s u}\alpha$: $\sum_{c \in C_s} \sum_{k \in C_s} \square_{C_s C_s \text{ nominal}} \delta^2_{ck} = \square_{C_s C_s} |C_s|(|C_s|-1)$ iff $|C_s| > 1$

So, we can **estimate** ${}_{c_s u}\alpha$ after users resolved the confusion by adding the above to the numerators of ${}_{C_s u}D_o$ and ${}_{C_s u}D_e$:

$${}_{c_s u}D_o \approx \frac{\sum_{C_s} \sum_{K_s} \square_{C_s K_s} \left(\Delta_{C_s K_s} + (|C_s \cap K_s|(|C_s \cap K_s|-1) \text{ iff } |C_s|>1 \text{ and } |K_s|>1)\right)}{\sum_{C_s} \sum_{K_s} \square_{C_s K_s} |C_s||K_s|}$$

and:

$${}_{c_s u}D_e \approx \frac{\sum_{C_s} \square_{C_s} \cdot \sum_{K_s} \square_{\cdot K_s} \left(\Delta_{C_s K_s} + (|C_s \cap K_s|(|C_s \cap K_s|-1) \text{ iff } |C_s|>1 \text{ and } |K_s|>1)\right)}{\left(\sum_{C_s} \square_{C_s} \cdot |C_s|\right)^2 - \sum_i^m \sum_g \left(L(S_{ig \text{ coded } C_s \neq \phi})\right)^2 |C_s|}$$

Whereby the **estimated ${}_{c_s u}\alpha$-agreement** becomes: ${}_{c_s u}\alpha \approx 1 - \frac{{}_{c_s u}D_o}{{}_{c_s u}D_e} \leq {}_{C_s u}\alpha$

To reiterate, this $_{c_{SU}}\alpha$ can only be a numerical *estimate*. The true $_{c_{SU}}\alpha$ depends on how you resolve the confusion, which likely influences the frequencies of the originally confused codes and other coincidences. The fact that this type of intra-coder confusion occurred, however infrequent it may be, means that it creates data for which many analytical methods become inapplicable when they are not corrected. This is only partly related to the numerical differences in computed agreement coefficients. Therefore, it is worthwhile to correct this type of confusion as it will clarify potential problems with the coding system.

---

[i] Developed from Chapter 12 in Klaus Krippendorff (2018). *Content Analysis; An Introduction to Its Methodology, 4th Edition*. Thousand Oaks, CA: Sage. Partly implemented in Klaus Krippendorff; Yann Mathet; Stéphane Bouvry & Antoine Widlöcher (2016). On the Reliability of Unitizing Textual Continua: Further Developments. *Quality & Quantity 50*, 6: 2347-2364. Online since 2015.9.15 at https://link.springer.com/article/10.1007/s11135-015-0266-1