Day 3: Classical Content Analysis

Kenneth Benoit

CEU April 14-21 2011

April 18, 2011

Hand-coding: "Classic" content analysis

- Key feature: use of "human" coders to implement a pre-defined coding scheme, by reading and coding texts
- Human decision-making is the central feature of coding decisions, not a computer or other mechanized tool
- Alternatives are the purely statistical analysis of text as data, where human decisions are minimal or non-existent, and statistical methods are used to scale quantities from texts
- Other alternatives could be purely descriptive approaches to word frequency analysis

Hand-coding': "Classic" content analysis

- Validity is usually the objective, rather than reliability
- Another motivating factor could be ease of use, or the difficulty of implementing an automated procedure
- May be *computer-assisted*, especially for unitization
- Common "CATA" or "CACA" tools:
  - MaxQDA
  - T-Lab
  - Atlas-TI (formerly NUD*IST)
  - WordStat
  - TextPack
  - Diction
  - General Inquirer
  - many others

Components of manual coding approaches

Unitizing The systematic distinguishing of segments of text that are of interest to the analysis.

Sampling Choice (and justification of the choice) of text units to sample, from population of possible text units.

Coding Classifying each coded unit of text from the sample according to the pre-defined category scheme.

Summarizing Reducing the coded data to summary quantities of interest.

Inference and reporting The final steps wherein the analyzed results are used to generalize about social world, and communicating these results to others.

Reliability-Validity Tradeoffs

- Reliability refers to the dependability and replicability of the data generated by the text analysis method
- Validity is the quality of the data that leads us to accept it as "true," insofar as it measures what it is claimed to measure
- In text analysis, these two objectives frequently trade off with one another, since only human judgment can (ultimately) ensure validity, but human judgment is inherently unreliable
- Each concept has many variations, and in the case of reliability, several measures that can be applied
- Validity is the hardest to establish, since questions can always be raised about human judgment

Examples of tradeoffs

- Examples in coding text units:
  - Perfectly reliable procedure: Code all text units as pertaining to "Economic growth: positive"
  - Perfectly valid: Get a Nobel Prize laureate in economics to classify each text unit
- Examples in unitizing a text:
  - Perfectly reliable: Have a computer parse all texts into *n*-grams, such as words, pairs of adjacent words, etc. based on pre-defined rules (space is a delimiter, etc.)
  - Perfectly (?) valid: Have expertly trained humans parse the text into "quasi-sentences"

Reliability: Definitions

- Reliability in essence means getting the same answers each time an identical research procedure is conducted.
- In text analysis (and most other forms of empirical analysis), unreliable procedures yield results which are meaningless.
- Typically measures in terms of agreement between two human coders, when referring to hand-coded content analysis
- Computerized methods have largely removed this concern, inasmuch as they are mechanical procedures that yield the same results each time the procedure is repeated.

Types of reliability

Distinguished by the way the reliability data is obtained.

Type	**Test Design**	**Causes of Disagreements**	Strength
Stability	test-retest	intraobserver inconsistencies	weakest
Reproducibility	test-test	intraobserver inconsistencies + interobserver disagreements	medium
Accuracy	test-standard	intraobserver inconsistencies + interobserver disagreements + deviations from a standard	strongest

Reliability test designs

Test-retest The same text is reanalyzed/reread/reclassified, or the same measurement is repeatedly applied to the same set of texts. Goal is to establish inconsistencies. (Establishes *stability*)

Test-test Two or more individuals, working independently, apply the same analysis instructions to the same texts, to compare intraobserver differences. (Establishes *reproducibility*).

Test-standard The perfomance or one or more procedures is compared to a procedure that is taken to be correct. Deviations from a ("gold") standard are then recorded. (Establshes *accuracy*.) Typically used in coder training, or training of automated (computer-based) procedures.

Designing reliability checks in practice

- Repeating the procedure on the sample data
- Using independent tests from separate coders
- Can a "gold standard" be identified?
- Split-design tests
- Example: CMP
  - Same coders repeat own codings
  - Different coders code same test
  - The "reliability" coefficient reported in the dataset is correlation of category percentages obtained by a coder on the training document used by CMP versus the master "gold standard" version of the coding done by Andrea Volkens

Measures of agreement

- Percent agreement Very simple: (number of agreeing ratings) / (total ratings) * 100%
- Correlation
  - (usually) Pearson's r, aka product-moment correlation
  - Formula: $r_{AB} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{A_i - \bar{A}}{s_A} \right) \left(\frac{B_i - \bar{B}}{s_B} \right)$
  - May also be ordinal, such as Spearman's rho or Kendall's tau-b
  - Range is [0,1]
- Agreement measures
  - Take into account not only observed agreement, but also *agreement that would have occured by chance*
  - Cohen's κ is most common
  - Krippendorf's α is a generalization of Cohen's κ
  - Both range from [0,1]

Reliability data matrixes

Example here used binary data (from Krippendorff)

Article:	1	2	3	4	5	6	7	8	9	10
Coder A	1	1	0	0	0	0	0	0	0	0
Coder B	0	1	1	0	0	1	0	1	0	0

- A and B agree on 60% of the articles: 60% agreement
- Correlation is (approximately) 0.10
- Observed *dis*agreement: 4
- Expected *dis*agreement (by chance): 4.4211
- Krippendorff's $\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{4}{4.4211} = 0.095$
- Cohen's κ (nearly) identical

Computing reliability measures using R

- First, install the `irr` package:
  ```
  > library(irr)
  ```
- Enter the data from the previous example:
  ```
  > a = c(1,1,0,0,0,0,0,0,0,0)
  > b = c(0,1,1,0,0,1,0,1,0,0)
  ```
- Concatenate the two vectors to make the reliability data matrix:
  ```
  > rdm <- cbind(a,b)
  ```
- Now we can get the statistics we need. We will use commands from the `irr` library, namely `agree`, `kripp.alpha`, and `kappa2` Also the `cor` command will give us the correlation.

Computing reliability measures using R

```
> library(irr)
> a = c(1,1,0,0,0,0,0,0,0,0)
> b = c(0,1,1,0,0,1,0,1,0,0)
> rdm <- cbind(a,b)
> cor(a,b)
[1] 0.1020621
> agree(rdm)
 Percentage agreement (Tolerance=0)

 Subjects = 10
   Raters = 2
  %-agree = 60
> kripp.alpha(rdm)
 Krippendorff's alpha

 Subjects = 10
   Raters = 2
    alpha = 0.0952
> kappa2(rdm)
 Cohen's Kappa for 2 Raters (Weights: unweighted)

 Subjects = 10
   Raters = 2
    Kappa = 0.091

        z = 0.323
```

Computing reliability measures using R

- Variations can use **weights** (e.g. for kappa, alpha)
- Can **generalize to** *m* **raters** (not just two), by using
  - `kappam.fleiss`
  - `kripp.alpha`
- Can change **tolerance** for disagreement using `agree(ratings, tolerance=1`
- Can change **level of information** in the data, using α, by `kripp.alpha(x, method = c("nominal", "ordinal", "interval", "ratio"))`

Reliability and validity differences

- Reliability can be established through tests as a part of a research procedure; validity cannot be established through the same sort of (repetition) tests.
- Validity concerns substantive *truths*, whereas reliability is mainly procedural.
- Unreliability limits the chance of obtaining valid results, in the sense that procedures whose results cannot be trusted are less likely to be true.
- Reliability is no guarantee of validity, since reliable procedures can be consistently wrong, even when these procedures involve human judgment.

Additional (related) concepts

Generalizability The extent to which findings may be applied to cases other than those from which the research is immediately taken, for instance from a sample to a population. (We will subsume this under "external validity".)

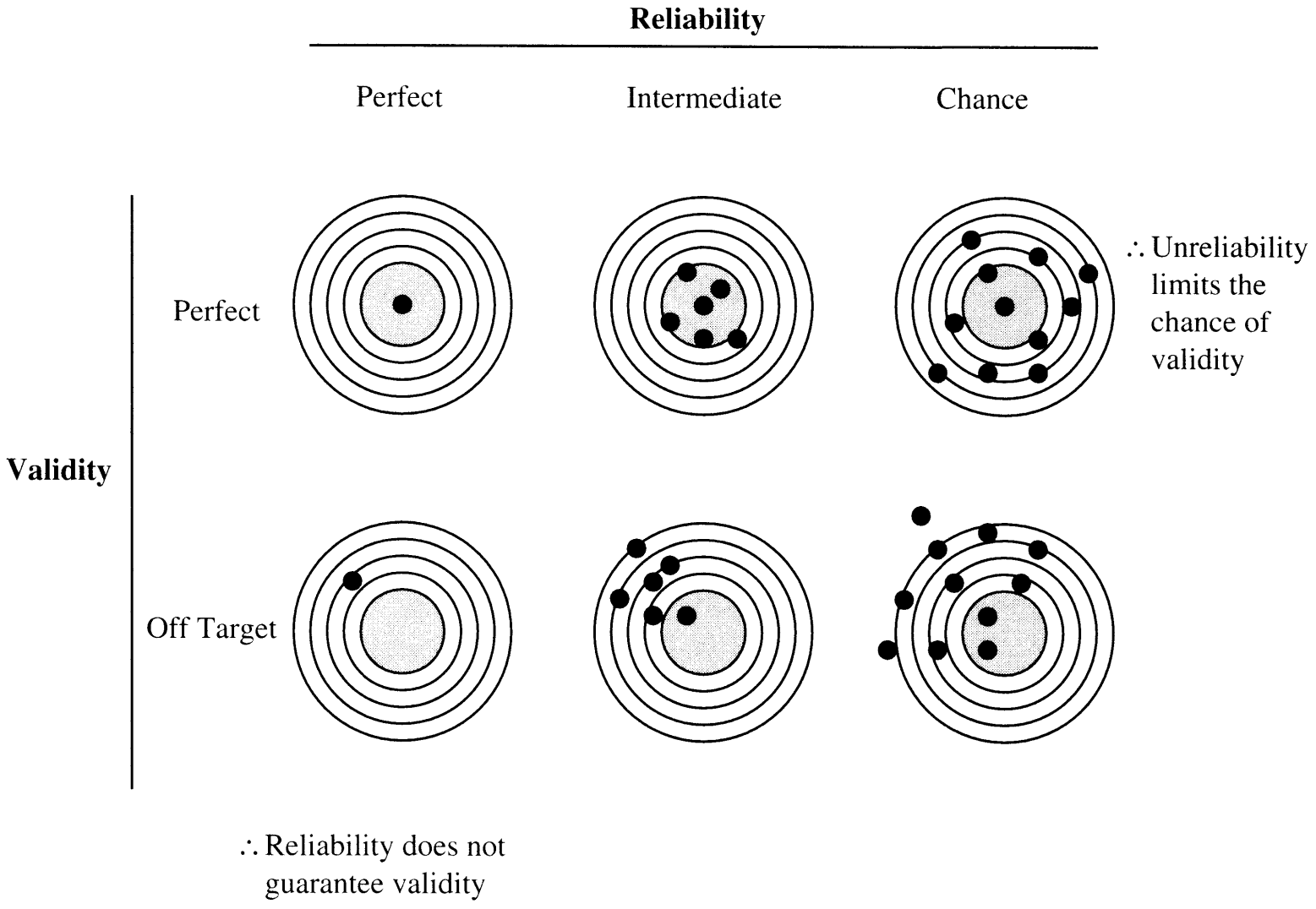
Precision The fineness of distinction or level of measurement. For instance, measuring time in morning/afternoon versus HH:MM:SS.

Accuracy The extent to which a measurement corresponds to the truth – usually determined by whether it is free from bias, but also affected by reliability.

These last two concepts also trade off with one another: highly precise measures are less likely to be accurate.

Interrelation of additional concepts

(From Krippendorff Figure 11.1)



Sampling Texts

- (Mainly we have already covered this on Days 3–4)
- In hand-coded schemes, sampling may be more deliberate
- For the Comparative Manifesto Project, the case study for this topic, "sampling" consists of selecting all texts of a particular class

Coding Text Units

- The key step in transforming raw texts into representations that can be analyzed
- Involves reducing and quantifying the data into discrete categories
- Requires a pre-defined scheme with rules for how these should be applied
- Question in designing the scheme is to maximize on the precision-accuracy-reliability frontier
- This can only be done through an iterative process of design, with *human-involved reliability tests at each step*
- The Big Problem: the dilemma of maintaining backwards-compatibility versus achieving optimal design

Summarizing

- Involves characterizing the coded text units using additional quantification
- Examples

  Category frequencies Coded category frequency measures, such as the proportion of times "economy" is mentioned in a speech, or the proportion of mentions of the environment

  Type/token measures Frequency tabulations of token types and their frequencies

  Range/variance Here we might be interested in the total number or the spread or variance of categories used in particular documents or by particular speakers
- May also involve scales or indexes constructed from summary information

Summarizing: Example

Democratic	Republican
iraq	consent
administration	ask
year	unanimous
health	bill
families	committee
program	senate
care	30
debt	2006
women	border
veterans	senator
help	vote
americans	law
country	hearing
children	authorized
new	further
education	states
funding	proceed
workers	order
programs	session
disaster	time

Top 20 Democratic and Republican words from the 2006 US Senate (source: Nicholas Beauchamp 2008)

Summarizing: Scale Example

- A very simple example comes from the CMP, using PER110 "European Union: Positive Mentions" and PER108 "European Union: Negative Mentions"
- The overall pro- versus anti- EU-ness can be assessed as PER110 - PER108. Theoretical range is $[-100, 100]$.
- A more complicated example is the CMP's famous "rile" index, which adds 26 categories of the "right" and subtracts from this the sum of 13 categories of the "left".

Inference and Reporting

- This involves drawing conclusions from the research, and these conclusions will depend on the *validity* established by the research design
- Reporting means communicating the results in a clear and relevant fashion. (This can be challenging – see for instance the Schonhardt-Bailey article.)
- No iron-clad rules here – use your discretion as applied to a particular case

Unitizing Texts

- Briefly read the CMP Coder Instructions in Appendix 2 of Mapping Policy Preferences II (on the web page for Day 2).
- To unitize the text on the next slide.

Unitize this

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. The fight against increases in the cost of living is the most important single issue in economic management.
People without jobs represent waste of productive effort: National supports a policy of full employment and the dignity of labour. We do not accept unemployment as a balancing factor in economic management.
Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.

A Test: How many of you said seven?

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. / The fight against increases in the cost of living is the most important single issue in economic management. / People without jobs represent waste of productive effort: / National supports a policy of full employment / and the dignity of labour. / We do not accept unemployment as a balancing factor in economic management. / Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.

Unitizing Texts

- What were our experiences unitizing the CMP reliability test document?
- What were your impressions of this unitization scheme?
- What alternatives exist?
  - physical distinctions: time, length, size, volume
  - syntactical distinctions: words, sentences, paragraphs, chapters, articles, etc.
  - categorical distinctions: units defined by membership in a class or category – references to a particular (pre-defined) topic
  - propositional distinctions: constructions from structure of the language, e.g. separating clauses. A version of this forms the basis for the CMP's "quasi-sentence" scheme
  - thematic distinctions
- Some methods exist for *assessing the reliability of unitization* but these are not simple to compute

And now try to *code* it

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. / *The fight against increases in the cost of living is the most important single issue in economic management.* / *People without jobs represent waste of productive effort:* / *National supports a policy of full employment* / *and the dignity of labour.* / *We do not accept unemployment as a balancing factor in economic management.* / *Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.*

And now try to *code* it

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. / *The fight against increases in the cost of living is the most important single issue in economic management.* / *People without jobs represent waste of productive effort:* / *National supports a policy of full employment* / *and the dignity of labour.* / *We do not accept unemployment as a balancing factor in economic management.* / *Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.*

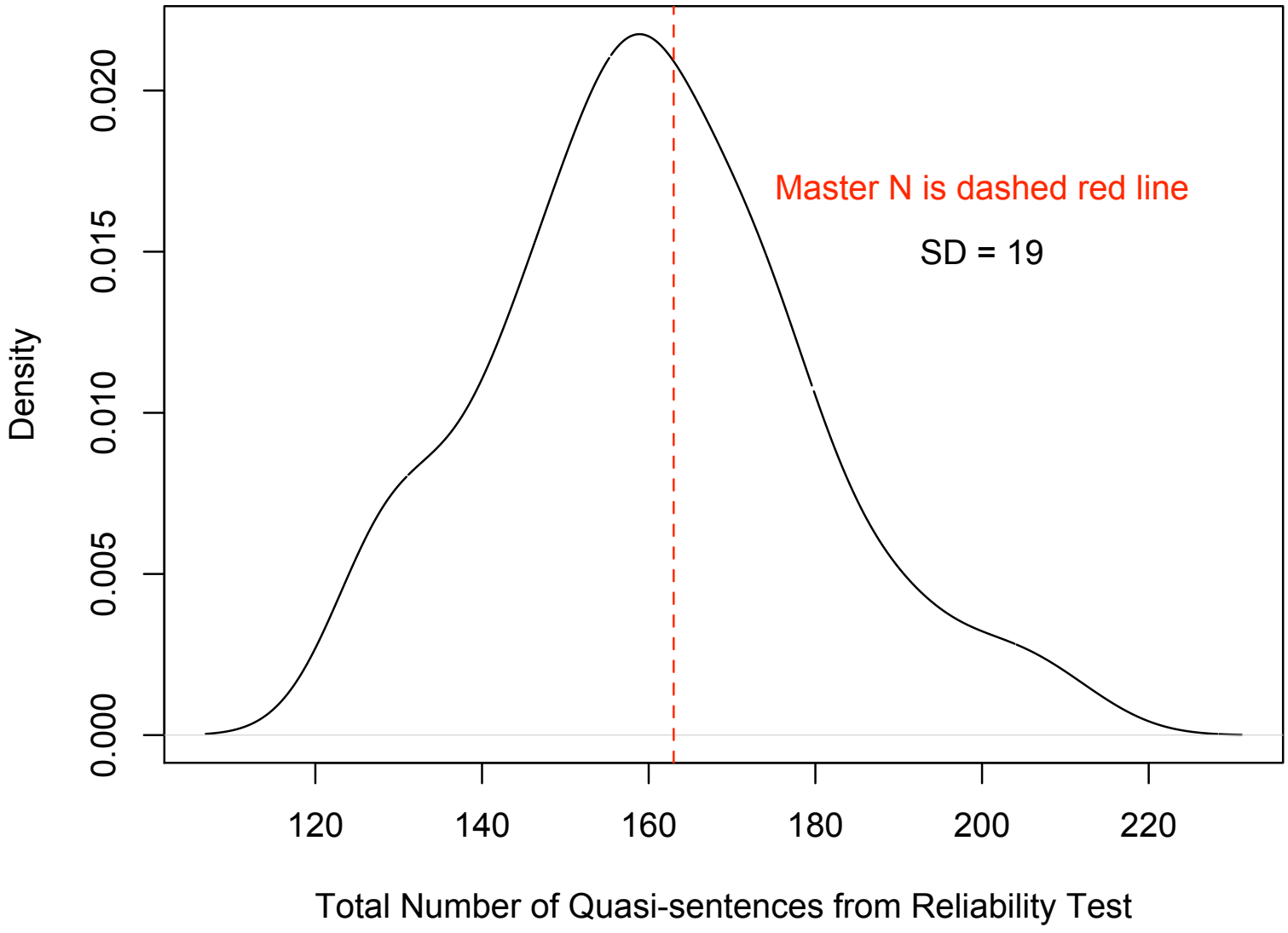
And the ("gold standard") answer is:

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. / The fight against increases in the cost of living is the most important single issue in economic management. /
People without jobs represent waste of productive effort: / National supports a policy of full employment / and the dignity of labour. / We do not accept unemployment as a balancing factor in economic management. /
Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour. /

414
414
410
408
701
701
405

414 "Economic Orthodoxy: Positive"
410 "Productivity: Positive"
408 "Economic Goals"
701 "Labour Groups: Positive"
405 "Corporatism: Positive"

Unitization empirical results from CMP tests

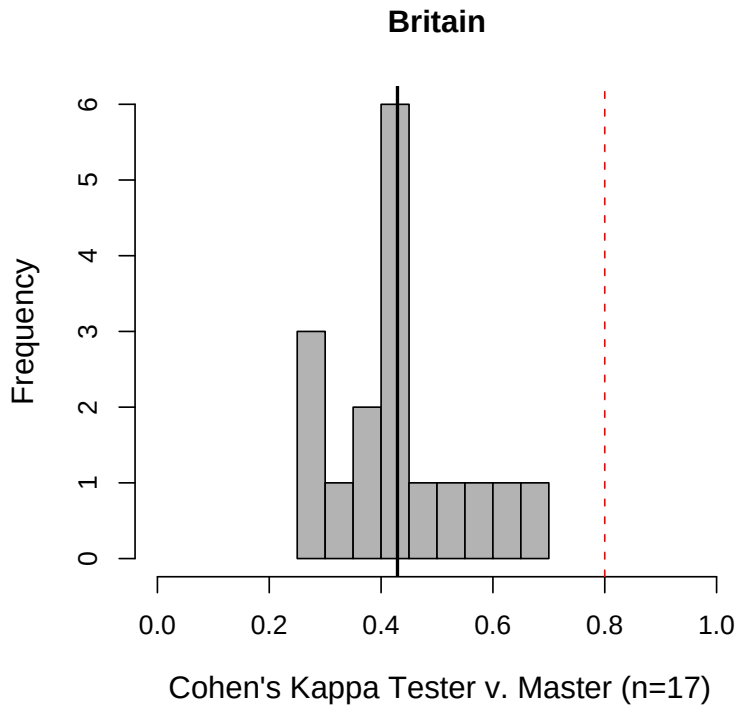


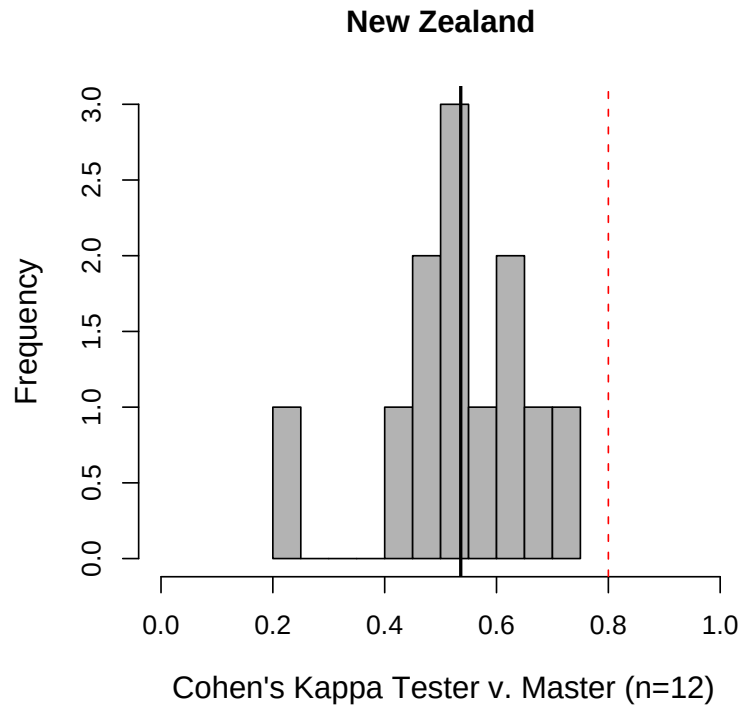
Empirical results from Mikhaylov and Benoit 2010

Caveats before I show you some compromising pictures:

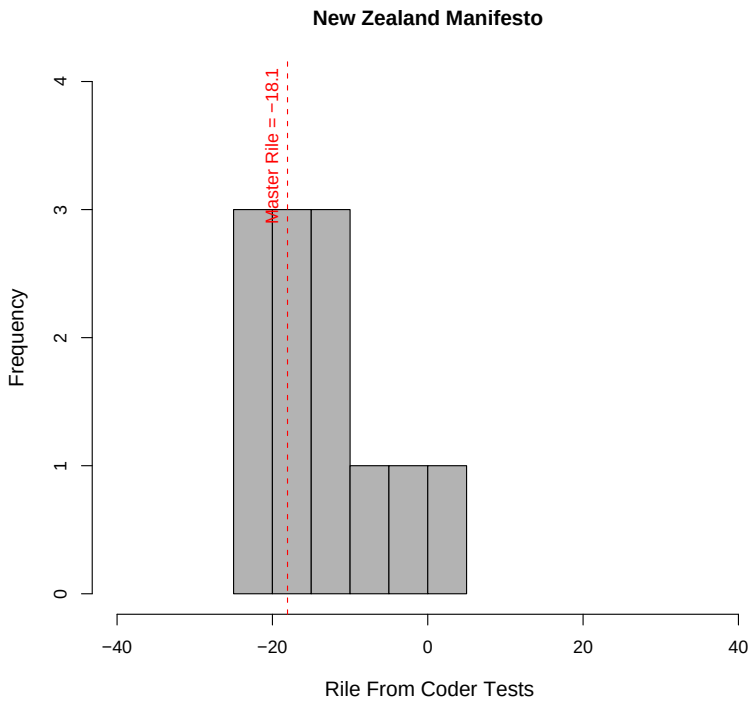
- We are not out to smear mud on the CMP! We actually like and respect the CMP and believe in the usefulness of their objective.
- *At the same time*, no research project should be immune from improvement
- There are weaknesses in the data and these are worth knowing
- The structure of the tests: Ask trained coders used by the CMP to code CMP manifestos to complete a recoding test online, for a test that was used as an example in the CMP coding instructions. Text was pre-unitized.

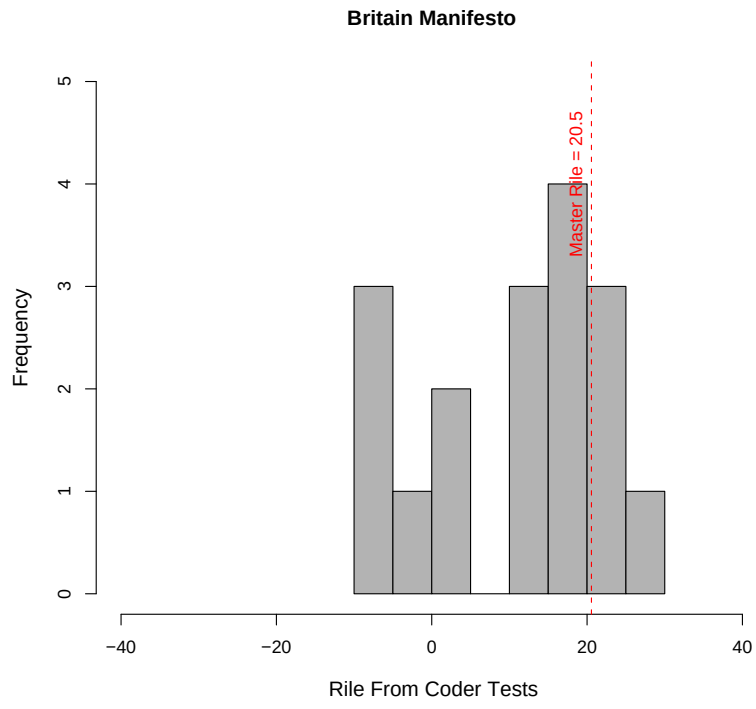
Empirical results from CMP reliability tests



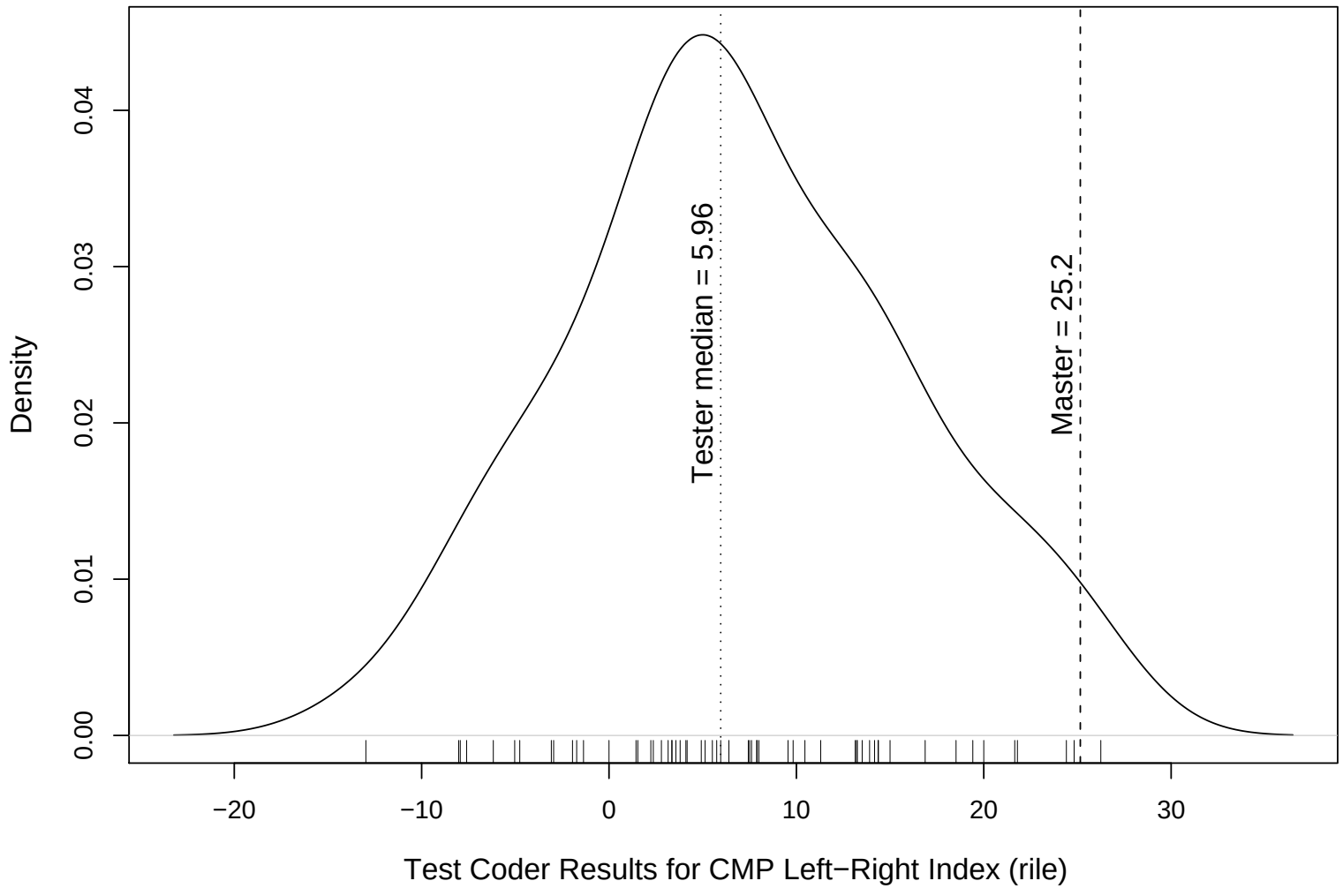


Empirical results from CMP reliability tests

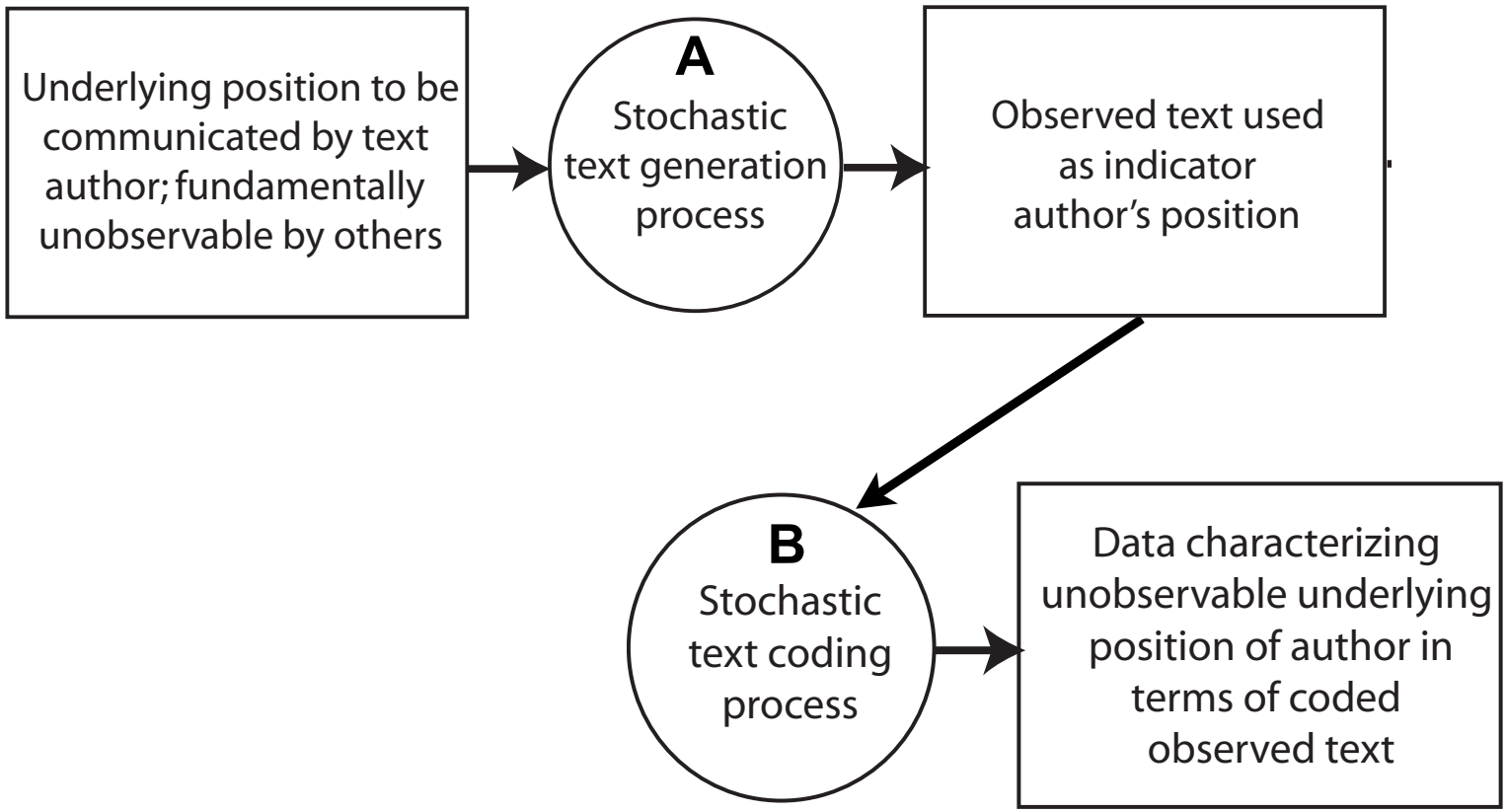




Empirical results from CMP reliability tests



The Big Picture



Scaling Issues

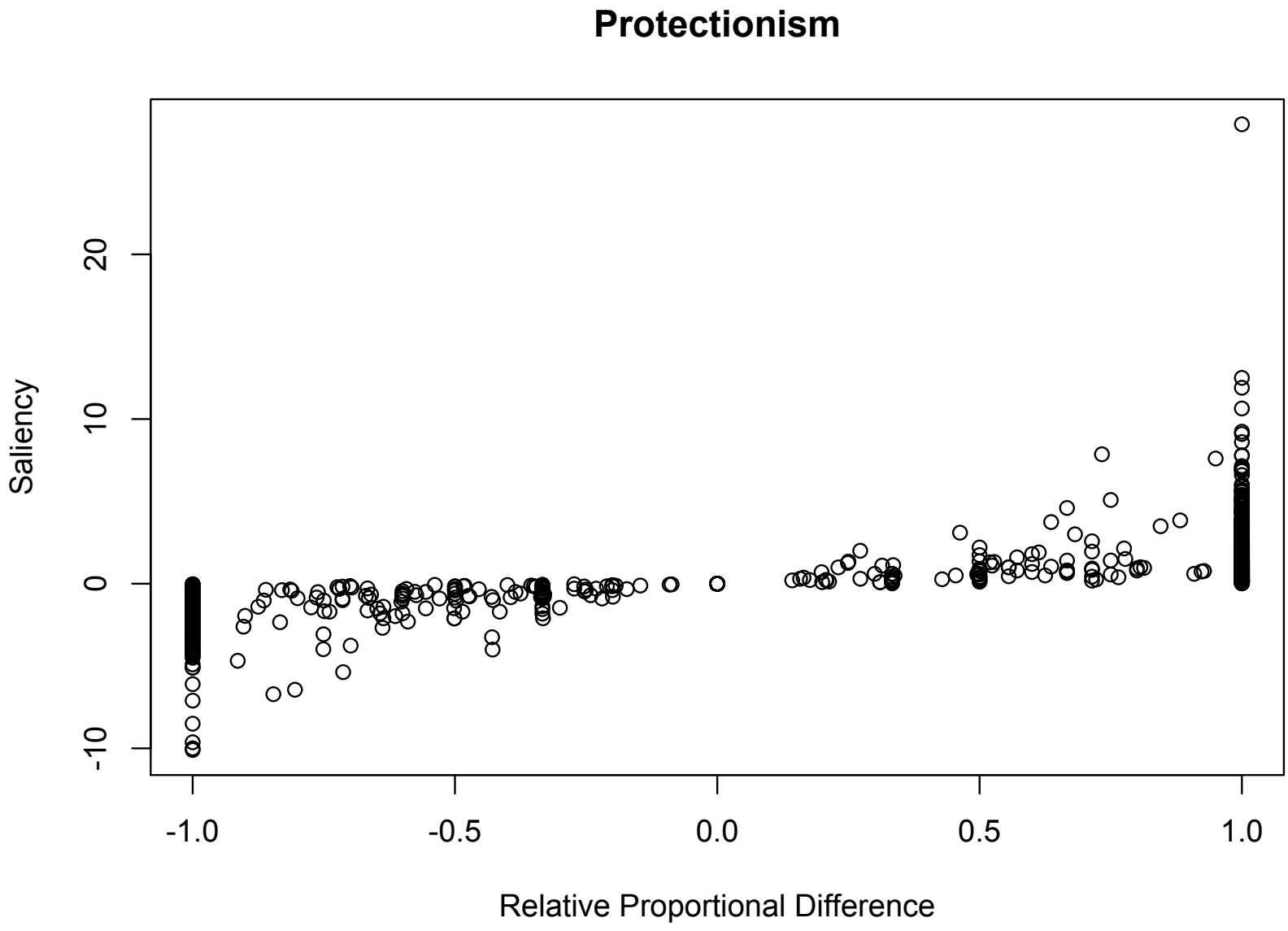
- Scaling becomes a major issue when we wish to construct quantities of interest from quantitative content analyses
- Simple example: Proportion of content of a given type (e.g. anti-Lisbon treaty)
- Complex example: Left-right policy positions (e.g. CMP "Rile")
- Are the metrics "natural"?
- Does the output metric resemble the input metric (if any)?
- What properties should the scale have, such as boundaries, type of increase, etc?
- How can uncertainty be characterized for the given scale?

Logit scale for left-right

- The Comparative Manifesto Project scales policy positions as absolute porportional difference, measured by proportion of "Right" mentions less proportion of "Left" mentions: $\frac{(R-L)}{N}$
- Problems:
  - Addition of irrelevant content shifts the scale toward zero
  - Assumes the additional mentions increase emphasis in a linear scale
- The alternative is to scale $\frac{(R-L)}{(R+L)}$ (Kim and Fording 2002; Laver and Garry 2000), but this too has problems:
  - Still linear shift in position for increase in repetition
  - Quickly maxes out at the extremes
- Lowe, Benoit, Mikhaylov and Laver (2010) propose using a logistic odds-ratio scale $\log \frac{R}{L}$

Comparing scales:

$\hat{\theta}^{(S)}$ v. $\hat{\theta}^{(R)}$



Comparing scales

Protectionism

distributions