



Assessing Inter-rater Agreement in the Coding of Narrative Data

Ken Miller

San Francisco State University

Greg Worthington

Chicago School of Professional Psychology



Focus: Inter-rater Agreement

- ◆ What is it?
- ◆ Why does it matter?
- ◆ The neglect of inter-rater agreement in the qualitative literature
- ◆ What are the S-O-P's for achieving and assessing inter-rater agreement?
- ◆ Strategies for improving level of agreement



Inter-rater Agreement: A Definition

- ◆ Definition for use with narrative or textual data:

The extent to which 2 or more coders agree with one another in their application of a code scheme to a body of narrative data.



Inter-rater Agreement: Rationale

- ◆ With multiple coders: to ensure that codes are understood and applied in the same way by all coders.
 - Especially important when reporting number or percent of participants who responded in a particular way.
 - Reduces the role of individual interpretive bias in the application of codes

Bias in Coding

- ◆ Qualitative researchers emphasize the impossibility of removing bias from the research process
- ◆ What we choose to study, the questions we ask and those we don't, how we probe for further information, all reflect inherent biases that can't be “removed” from the process without losing the process itself.
- ◆ However, it *is* both possible and desirable to minimize bias in the development and application of codes to the data.
 - Multiple perspectives increase the likelihood that codes reflect what is in the data, rather than what is in the idiosyncratic perception of a lone researcher.



Inter-rater Agreement: Rationale II

- ◆ For one researcher: to ensure that application of codes is not arbitrary, and that codebook can be effectively utilized by another individual with minimal variation.



A Case of Neglect

- ◆ Discussions of inter-rater agreement—its purpose and importance, and methods of achieving and assessing it, are sparse in the literature on qualitative research.
 - e.g., Miles & Huberman
 - e.g., Strauss & Corbin
 - e.g., Creswell



Standard Operating Procedure: A Reflection of this Neglect

- ◆ All too often, discussions of inter-rated agreement are completely omitted.
 - Dyregrov et al., 2000
- ◆ Alternatively, procedures are described, but the procedures themselves are problematic and not critically examined.
 - Carey, Morgan, & Oxtoby, 1995

Approaches to Assessing Inter-rater Agreement

- ◆ The Intercoder Agreement Formula (Miles & Huberman, 1994)

$$\text{Reliability} = \frac{\text{Agreements}}{\text{Agreements} + \text{Disagreements}}$$

See Miller & Keys (2001) for an illustration of this approach

Approaches to Assessing Inter-rater Agreement

- ◆ Inter-rater agreement index:

$$\frac{(\sum \text{Agreements}) \times (\# \text{ coders})}{\text{total coding instances}}$$

- ◆ Modified inter-rater agreement index, with partial agreements (e.g, $\frac{3}{4}$ agreement):

$$\frac{[(\sum \text{Agreements}) \times (\# \text{ coders})] + [\sum \text{partial agreements} \times .75 \times \# \text{coders}]}{\text{total coding instances}}$$

Example

◆ Example:

- 2 coders coding a single interview, using a set of 40 codes
- Coder A codes 30 segments of text, Coder B codes 28 segments. They agree on 24 coding instances.
- Agreement index: $\frac{24 \times 2}{30 + 28} = \frac{48}{58} = .83$

Cohen's Kappa

- ◆ The Inter-rater Agreement Index does not account for chance agreement.
- ◆ The Kappa coefficient, which can be calculated in Excel, SPSS and other programs, *purportedly* does take into account chance agreement.
 - The Kappa statistic is a measure of effect size that estimates the amount of agreement between 2 (or more) coders, above what might occur by chance alone.



Kappa: A Problematic “Gold Standard”

- ◆ Assumes independence of raters, an assumption that is rarely met.
- ◆ Treats all codes as if all they are equally likely to occur (i.e., be used) and equally clear in their meaning and usage.
 - This is rarely the case.
- ◆ Provides an omnibus statistic that can obscure significant differences in agreement on particular codes

Chance Agreement: How Concerned Should We Be?

- ◆ With a small number of codes and few coders, chance agreement is a significant factor. It can significantly inflate our estimate of agreement.
- ◆ e.g., with 10 codes and 2 coders, agreement due to chance is $1/10$ on any coding instance, *assuming independence of coders and same ease and likelihood of use among all codes.*
 - Given that these assumptions are rarely met, the estimation of chance is at best a crude estimate.

Chance Agreement: How Concerned Should We Be?

- ◆ However, with many codes and/or multiple coders, the effects of chance are minimized.
- ◆ With 40 codes and 2 coders, chance agreement is $1/40$.
- ◆ With 40 codes and 3 coders, chance agreement is $1/1,600$
- ◆ With 40 codes and 4 coders, chance agreement is $1/64,000$
- ◆ With 60 codes and 2 coders, chance agreement is $1/60$.
- ◆ With 60 codes and 3 coders, chance agreement is $1/3,600$
- ◆ With 60 codes and 4 coders, chance agreement is $1/216,000$



Summary

- ◆ The neglect of inter-rater agreement in narrative research plays into the hands of critics of qualitative methods, who view them as unscientific and lacking in rigor.
- ◆ Assessments of inter-rater agreement range from the very basic, to the more complex, and all are ultimately approximations.
- ◆ At this point, however, such approximations represent a step forward, for the information they do provide.
- ◆ Accounting for chance agreement is quite complex and represents an important area for further discussion. The importance of accounting for chance varies with the number of codes and the number of coders.



Strategies for Enhancing Reliability

- ◆ Allow a reasonable location margin
- ◆ Develop a detailed codebook
- ◆ Avoid vague and/or overlapping codes
- ◆ Find a good balance of detail and parsimony:
 - More is not always better