Human Organization, Vol. 58, No. 3, 1999

0018-7259/99/030313-10$1.50/1

Measuring the Typicality of Text: Using Multiple Coders for More Than Just Reliability and Validity Checks

Gery W. Ryan

Social scientists often use agreement among multiple coders to check the reliability and validity of the analytic process. High degrees of intercoder agreement indicate that multiple coders are applying the codes in the same manner and are thus acting as "reliable" measurement instruments. Coders who independently mark the same text for a theme provide evidence that a theme has external "validity" and is not just a figment of the investigator's imagination. In this article, I extend the use of multiple coders. I use data taken from clinicians' descriptions of personal illness experiences to demonstrate how agreement and disagreement among coders can be used to measure core and peripheral features of abstract constructs and themes. I then show how such measures of multicoder agreement can be used to identify typical or exemplary examples from a corpus of text.

Key words: validity measures, reliability measures, intercoder agreement, text analysis, qualitative research methods

Browse through most social science journals and you will find ethnographic accounts interspersed with paraphrased or verbatim quotes taken from informants. A skeptic might ask two simple questions: "To what degree do the examples represent the data the investigator collected?"; and "To what degree do they represent the construct or constructs that the investigator is trying to describe?"

Both questions focus specifically on the investigator's synthesis and presentation of the data rather than on the investigator's data collection techniques. Such criticism has had a long history in qualitative research. In the late 1950s, Becker writes:

> Readers of qualitative research reports commonly and justifiably complain that they are told little or nothing about the evidence for conclusions or the operations by which the evidence has been assessed. A more adequate presentation of the data, of the research operations, and of the researcher's inferences may help to meet this problem (1958:659).

Thirty-five years later, Dey echoes the same frustrations. He says:

> Qualitative analysts have been notoriously reluctant to spell out even in the vaguest terms the decision-making process involved in their analysis. They have been reluctant to admit the possibility of error, preferring to present only such evidence as supports rather than contradicts their analysis. They have tended to rely instead . . . on the audience's trust in the integrity of the analyst. This makes even the most cursory scrutiny of the reliability of their procedures impossible. But the time spent on laying out our decision-making processes is not wasted if it gives our audience an insight into our analytic procedures and enhances their confidence in our results (1993:252-253).

How do practitioners of ethnography address such criticisms? Some investigators have advocated the use of multiple coders to better link abstract concepts with empirical data (Carmines and Zeller 1982; Miles and Huberman 1994). Investigators use agreement among multiple coders as proxies for the reliability and validity of the analysis process. High degrees of intercoder agreement indicate that multiple coders are applying the codes in the same manner and are thus acting as reliable measurement instruments. Coders who independently mark the same text for a theme provide evidence that a theme has external "validity" and is not just a figment of the investigator's imagination (Mitchell 1979).

Gery W. Ryan is an assistant professor in the Department of Anthropology at the University of Missouri-Columbia. *The research on which this article is based is part of a National Science Foundation Grant on "Methods for Conducting Systematic Text Analysis" (SRB-9811166). I thank Kathleen MacQueen, at the Centers for Disease Control, for steering me toward the problem of identifying and measuring typicality and for her helpful suggestions and H. Russell Bernard, for his insights into potential solutions to this problem. I also thank the clinical scholars at UCLA who collected and coded the texts and Jim Carey, Bobby Milstein, Thomas Weisner, Michael Schnegg, and two anonymous reviewers for their invaluable comments on earlier drafts of this paper.*

In this article, I extend the use of multiple coders. I use data taken from clinicians' descriptions of personal illness experiences to demonstrate how agreement and disagreement among coders can be used to measure core and peripheral features of abstract constructs and themes. I then show how such measures of multicoder agreement can be used to identify typical or exemplary examples from a corpus of text.

Clearly, I have oversimplified the complexity of the skeptic's problem in this brief introduction. In fairness, there are those who think that such concepts as reliability and validity are inappropriate for analyzing qualitative data (e.g., Lincoln and Guba 1985; Hammersley 1992; Denzin and Lincoln 1994). There are even some researchers who challenge the notion that multiple coders should be used at all in qualitative research (Morse 1994). In this article, I distance myself from the philosophical dilemmas of what "should" be done and present some examples of what "can" be done.

Data Collection

During the summer of 1996, I co-taught a qualitative data analysis course with Dr. Thomas Weisner to clinicians at the UCLA medical school. As part of the course, we wanted participants to actually collect, code, and analyze qualitative data in a systematic manner. One of the substantive topics we explored was clinicians' own past experiences with colds and flu. Such acute and frequently occurring illnesses were things that everyone had in common, and from past experience we knew they would provide rich and varied data. We also thought it would be interesting to see how clinicians would describe their own illness experiences.

As part of the first day's exercises, we asked the participants to complete an open-ended questionnaire. One section included the following instructions: "In a couple of short paragraphs, please describe the last time *you* had a cold or the flu." In the spirit of Spradley's (1979) *grand tour questions*, we were intentionally vague about what to include in the descriptions and did not prompt participants for particular types of answers.

We collected the short responses from 23 clinicians in the class. For example, one clinician wrote:

> [For] two weeks I was covering the practice of a very busy physician in another city. I had to leave my family and was very, very busy. For the first couple of days I really did not feel very well and on the third day, I got up in the morning and immediately started vomiting. I did my best to pull myself together and went off to work—I had to. All this time I really thought my symptoms were all from stress and not an illness. Then several of the nurses who were working for me called in sick with a vomiting illness—I bet. I had the same thing and couldn't even recognize it.

Another described her illness thus:

> I felt extremely tired the evening before staying home. I was sneezing a lot and didn't do any reading or work on the computer. Instead, I went to bed early. The next morning I just wanted to sleep since I had not slept well because my nose was clogged and I can't sleep and breathe through my mouth comfortably. I decided to call in sick even though I might have been able to work because there were no pressing issues at work that day.

In all, the text totaled 1,554 words (about five single-spaced pages). We made copies of all 23 descriptions and gave them to the participants to read in the next class. While reading, we asked participants to keep a running list of themes and ideas that they noticed. When they finished, we led a group discussion about the themes they identified. From the discussion, we decided to focus further on three themes: 1) respondents' perceptions of signs and symptoms; 2) their descriptions of how the illness interrupted their daily activities; and 3) the criteria they used to select treatments.

For the next class assignment, we asked participants to "Read each illness description and mark blocks of continuous text where informants mention any of the three themes." The instructions about what counted as "blocks of continuous text" were left intentionally vague, so each coder had to decide whether to mark whole sentences or just key phrases. We did, however, explicitly tell them that they could mark text units with multiple themes.

For convenience, we had the clinicians code the illness descriptions with paper and pencil. We used a variety of simple marking conventions: Signs and symptoms were underlined with a straight line; interruptions of daily routine were marked with a wiggly line; and decision criteria were marked with "<<" at the beginning of the block and with ">>" at the end. We also had participants record the exact time they started coding and the exact time they finished. Below, I look at 10 participants who completed the coding task. On average, it took participants only 20 minutes to read five pages and code for three themes.

Data Management

Assessing intercoder agreement requires that each coder's marking behavior be placed in a standard format. With free-flowing texts, the trick is to consider the data as a long list of words that can be converted into a simple matrix. Each word in the text represents a single row in the matrix and is described by its structural relationship to a particular sentence, paragraph, and informant number. Each word can also be tied to the themes marked by each coder.[1]

Describing text in a matrix format does not reduce in any way the potential for interpretive analysis, nor does it make text analysis a mechanical procedure. Interpretation occurs when investigators identify new themes (thus creating new columns in a matrix) and when investigators associate text passages with these themes (thus assigning values to the appropriate cells). The fact that I can "back translate" the matrix and reproduce exactly what each coder did when he or she marked up the text indicates that data are not lost in the translation procedure. Describing text as a matrix, how-

ever, allows for comparisons across words, sentences, themes, questions, domains, informants, ethnic groups, and coders (Miles and Huberman 1994).

In order to move from our paper coding to a matrix of words, I went through a number of steps. First, I made 10 identical word-processing files of the text—one for each coder. Then I used a set of macros in my word processor to transfer each coder's marking into his or her own electronic file (Ryan 1996). (Of course, it would have been easier if each coder had marked the text on the computer in the first place, but this was not feasible at the time.) The macros insert coding tags at the beginning and end of specific blocks of text. Figure 1 shows some examples of the marking conventions. This procedure is further described by Truex (1993).

Once the tags were embedded in the file, I wrote another program to convert each coder's text file into a matrix (in this case, a database). Each matrix (or database) had 1,554 rows/records (one for every word in the text) and eight columns/fields. The first five columns were filled with variables that characterized each word and were constant for all coders. These included the word identification number (1-1154), the sentence number (1-107), the respondent number (1-23), the number of times each word form appeared in the text (between 1 and 129 times), and whether each word belonged to a list of common words that included particles, prepositions, and pronouns (0-1). (Word frequency counts and common-word lists are standard techniques in content analysis [Krippendorf 1980; Weber 1990] and have been reviewed by Ryan and Weisner [1996] and Bernard and Ryan [1998].) The remaining three columns were filled with 1s and 0s to indicate whether or not the coder had marked the particular word as pertaining to one of the three respective themes.

Since I wanted to analyze multicoder agreements for each theme separately, I merged the 10 coders' matrices into three theme-oriented matrices—one for signs/symptoms, one for interruption of daily routine, and one for decision criteria. As before, each matrix had 1,554 rows and the first 6 columns contained structural data describing each word. The next 10 columns contain 1s and 0s to indicate whether or not each of the 10 coders marked the word for the theme of interest.

Figure 1. Coding of Signs and Symptoms by Coders 1 and 2 for One Illness Description

Coder	Coded Text for Informant 1
1	>>S/SII I was uncomfortable with fevers associated with myalgrams, headaches, and a sore throat. IIS/S<< >>IntII I was unable to do my usual activities such as exercising, keeping up with routine paperwork. I could not enjoy social events and went to work but was miserable. IIInt<< It lasted about ten days and I recovered.
2	I was uncomfortable with fevers associated with >>S/SII myalgrams, headaches, IIS/S<< and a >>S/SII sore throat. IIS/S<< I was unable to >>IntII do my usual activities such as exercising, keeping up with routine paperwork. I could not enjoy social events and IIInt<< >>DecII went to work IIDec<< but was >>S/SII miserable. IIS/S << It lasted about ten days and I recovered.

Intercoder Pairs

One way to describe the central and peripheral aspect of a theme is to examine the agreement between *pairs* of coders. Since the texts were coded by 10 coders, I selected a pair of coders (1 and 2) at random and divided their data into 4 parts for each theme. Figure 2 shows the partial results for the signs/symptom (S/S) theme. All the text that Coder 1 marked as S/S, but Coder 2 did not, appears in the column headed "1 Only." All text that both Coder 1 *and* Coder 2 marked as S/S appears in the column headed "1 and 2." All text that Coder 2 marked but Coder 1 did not appears in the column headed "2 Only." The last column is filled with the remaining text (the text not marked by either Coder 1 nor 2 as S/S). The numbers in Figure 2 indicate how the original text was segmented. For example, the first phrase in column 1, "*I was uncomfortable with fevers associated with. . .,*[1]" is connected to the first phrase in column 2, "[1]*...myalgrams, headaches,...,*[2]" is connected to the second phrase in column 1, "[2]*...and a...*[3]" (Note that the word *myalgrams* was actually a typographical error in the text that the clinicians coded. The word used in the informant's original narrative was *myalgia* and referred to muscle pain. I will discuss the significance of this error below.)

Intercoder agreement (the text in the column labeled "1 and 2") shows us the core features of a theme. Core features are similar to the bull's eye of a target. They represent the central tendency or typical examples of abstract constructs. In Figure 1, core signs and symptoms include *myalgrams, headaches, sore throat, tired, contested/congestion, cough/coughing*, and *lower energy levels.*

In contrast, intercoder disagreement (the text in the columns labeled "1 Only" and "2 Only") shows the theme's peripheral features. Here, peripheral features are equivalent to the outer rings of a target. They may still be considered to be part of the construct but are less typical. In a sense, peripheral features represent the "edges" of a theme. In Figure 1, peripheral features include subthemes associated with general discomfort and hassle (*uncomfortable, miserable, could not control, needed to nap*, and *difficulty working all day*) and time (*two weeks, after the, continued to, at times, several times during the day*, and *all day*). Similar words and phrases are relatively uncommon in the core features, suggesting that they may be systematic differences rather than the results of one or another coder forgetting to code a particular phrase.

Figure 2. Intercoder Agreement and Disagreement about Signs and Symptoms for Three Illness Descriptions with All Words Presented

Case ID	Coders: 1 Only	1 & 2	2 Only	Neither 1 Nor 2
1	I was uncomfortable with fevers associated with...[1] [2]...and a...[3]	[1]...myalgrams, headaches,...[2] [3]...sore throat...[4]	[5]...miserable. ...[6]	[4]...I was unable to do my usual activities such as exercising, keeping up with routine paperwork. I could not enjoy social events and went to work but was...[5] [6]...lasted about ten days and I recovered.
2	[3]...I ...[4] [6]...I was...[7] [8]...After the...[9] [10]...I had an...[11] [12]...for two weeks which was embarrassing at times....[13] [14]...Although I...[15] [16]...I could not control....[17]	[4]...felt tired,...[5] [7]...very congested....[8] [9]...congestion cleared,...[10] [11]...annoying dry cough ...[12] [15]...felt fine,...[16] [17]...the coughing during the play....[18]	[1]...a cold...[2] [19]...continued to cough.	I had...[1] [2]...of moderate severity....[3] [5]...but I had to work....[6] [13]...I had tickets to a play that I had looked forward to for two months....[14] [18]... I bought two kinds of cough drops and cough syrup but still...[19]
3	[1]...usual, more...[2] [3]...needed to nap several times during day. Had a ...[4] [5]...and progressive difficulty working all day....[6]	Energy level felt lower than...[1] [2]...easily tired,...[3] [4]...productive cough...[5]		[6]...Felt like probable bronchitis and arranged for antibiotics. Recovery noted after several weeks.

Since there is more text in the column "1 Only" than there is in "2 Only," we can see that Coder 1 has a tendency to mark more text than does Coder 2. It turns out that Coder 1 always marked the entire sentence, while Coder 2 marked specific phrases. These two approaches have different advantages. Marking phrases provides a narrower and more concise summary of the theme, while marking sentences broadens the theme to less obvious aspects of signs and symptoms, such as associations with discomfort and time.

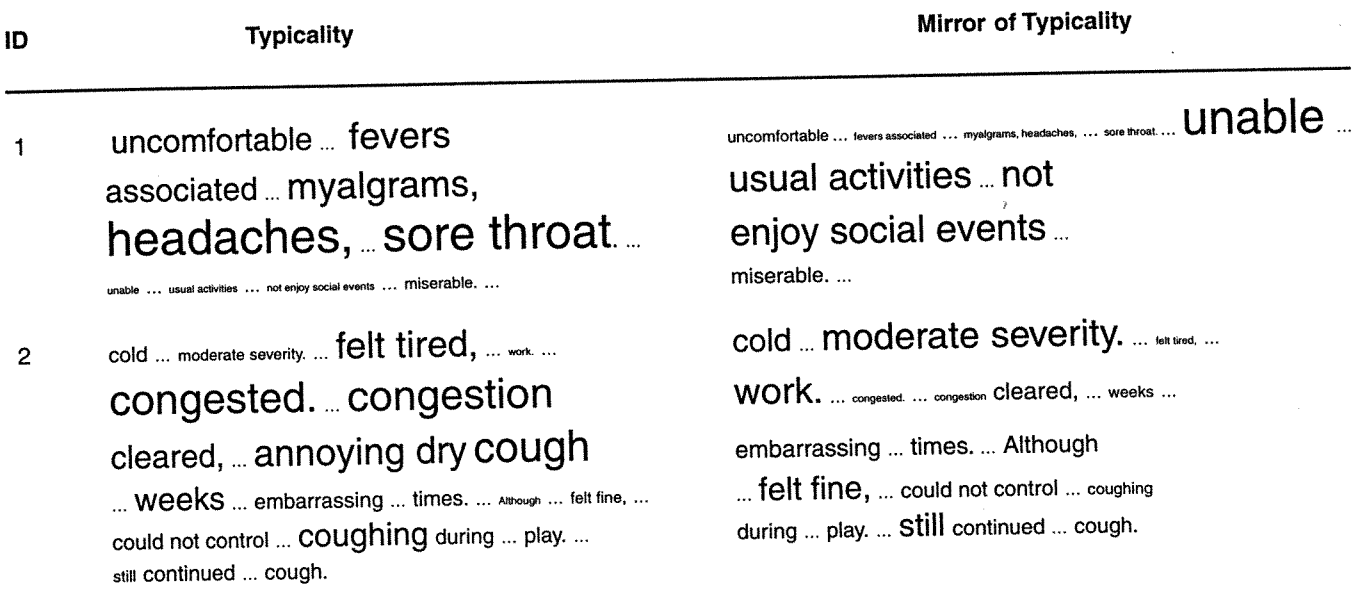
Coders also agreed about what text should *not* be marked as pertaining to signs and symptoms. The text in the column marked "Neither 1 Nor 2" is indicative of the extreme boundaries of the theme. For instance, neither of these two coders felt that the phrase "of moderate severity" pertained to signs and symptoms, yet one of the coders felt that "a cold" did. MacQueen et al. (1998) refer to such boundaries as "exclusion criteria" and find them useful for helping coders decide whether to mark particular segments of text. The "left over" text is also a good place to look for additional themes. For example, much of the last column pertains to the interruption of daily routine and treatments. It is also a good place to look for text that both coders may have missed inadvertently.

Multiple Coders

With only two coders, the typicality of responses pertaining to any theme is difficult to assess. With multiple coders, however, the task is easier. First, I calculated the *intercoder word frequency*—i.e., the number of times that the 10 coders marked each word. The numbers ranged from 0 (no coders marked the word) to 10 (all coders marked the word). I assume that the more coders who identify words and phrases as pertaining to a given theme, the more representative the text.

I used a database program to identify those words where at least one coder had marked them as pertaining to the theme (intercoder word frequency >0) and that did not belong to the common word list. Once I identified the key text for the theme, I formatted the output based on the intercoder word frequencies. I printed the text marked by all 10 coders in 20-point font, the text marked by 9 coders in 18-point font, the text marked by 8 coders in 16-point font, and so on. The bigger the font, the more typical the text. Think of large fonts as being closer to the conceptual bull's eye. (Instead of font size, I could have used variations in color or background shading.)

Figure 3. Intercoder Agreement as Represented by Font Size for Two Illness Descriptions



I also created a mirror image of the typicality output. I printed the text marked by one coder in 20-point font, the text marked by two coders in 18-point font, the text marked by three coders in 16-point font, and so on. In this case, the larger the font, the more peripheral the text is to the theme.

Figure 3 shows the two display techniques side by side. In the figure's second example, the core concepts related to signs and symptoms are *tired, congested,* and *cough* followed by concepts related to *felt, congestion, cleared, annoying,* and, to some extent, *weeks.* The mirror image shows that *moderate severity, embarrassing, times,* and *felt fine* occupy more peripheral position within the theme. By juxtaposing core and periphery concepts, I have a more sophisticated manner for describing abstract themes.

The techniques described above are useful for displaying the typicality of words in context but are not very practical for describing general patterns for large corpuses of text, such as techniques. In such cases, an even more concentrated format is needed. I combined basic techniques from content analysis to list the words that coders agree belong to a given theme. Figure 4 shows the most typical words associated with signs and symptoms. The first column of words are those that all 10 coders agreed belonged to the S/S theme. These words are ranked according to how often each one appeared in the text. For example, the word *throat* occurred 14 times in the text, and in at least one occurrence, all 10 coders agreed the word pertained to the S/S theme. Unlike classic content analysis that associates high-frequency words with theme salience, this technique identifies words that are pertinent to a theme but that may have low frequencies. For example, coders always associated *shaking* and *sweats* with signs and symptoms, even though both words only occurred once in the illness descriptions. Also, note that all 10 coders recognized the typographical error, *myalgrams,* as a sign or symptom. Each coder probably assumed that the misspelled word was a sign or symptom simply by the context in which it was found.

The set of words on which all informants agreed tends to be related to physiological indicators. These words occupy the core part of the sign/symptom construct. The second set has a number of words related to severity (e.g., *mild, degrees, horrible, several, annoying, minimal*), suggesting that evaluations of illnesses might be an important subtheme. Sets further to the right indicate peripheral subthemes. For example, words related to time (e.g., *days, ago, weeks, before, end, night, morning, followed, next*), and behaviors (e.g., *slept, nap, work, concentration*) are found in the last three columns.

Measuring the Typicality of Quotes

The analysis techniques described above are useful for depicting the range and central tendency of an abstract theme. Researchers, however, typically want to use typical examples and quotes in their descriptions. They need to address two problems: 1) How does the investigator identify the most typical quotes? and 2) How can a critic or a reviewer be sure that the selected quotes or examples are indeed representative of the text being analyzed?

Drawing from the previous analysis, I calculated three measures of sentence typicality. Figure 5 shows the sentences that are the most representative of signs and symptoms and their respective measures of typicality. I calculated a *raw agreement score* by counting the number of coders that had marked at least one word in a sentence as pertaining to the S/S theme. The raw agreement score consists of integers from 0-10 and is simple to explain. From a practical perspective, however, this score was not very helpful for our purposes as

Figure 4. Word Frequency Ranked by Intercoder Agreement (Common Words Eliminated)

Intercoder Agreements									
10		9		8		7		6	
Freq.	Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.	Word
24	feel(ing, felt)	14	last(ed, ing)	33	day(s)	17	work	8	because
14	throat	8	not	6	slept	10	home	7	ago
12	sore	7	developed	3	lot	9	so	6	flu
11	cough(ing)	4	mild	2	second	9	weeks	6	next
11	fever(s, ish)	4	started	1	associated	5	symptoms	4	just
7	congest(ed,ion)	4	well	1	clogged	3	before	3	didn't
6	tired	3	bedridden(be,d)	1	extremely	3	began	3	during
4	ach(es,ing)	3	degrees	1	immediately	3	several	3	few
4	nasal	3	did	1	nap	2	end	3	morning
3	chills	3	usual	1	needed	2	Friday	2	got
2	myalgias	2	clear(ed)			2	night	1	typical
3	nose	2	horrible			2	really		
3	over	2	pretty			2	staying		
2	fatigue(d)	2	severe			2	still		
2	headache(s)	1	alternately			2	third		
2	level	1	annoying			2	times		
2	low(er)	1	appeared			1	concentration		
2	malaise	1	body			1	difficulty		
2	productive	1	light			1	evening		
2	rhinorrhea	1	lymphadenopatry			1	experienced		
2	sneezing	1	minimal			1	followed		
2	vomiting	1	muscles			1	high		
2	whole					1	progressive		
1	discharge					1	spent		
1	drip					1	subsequently		
1	dry					1	uncomfortable		
1	easily					1	waning		
1	energy					1	weak		
1	generalized								
1	grade								
1	headedness								
1	lethargic								
1	myalgrams								
1	nonproductive								
1	post								
1	runny								
1	shaking								
1	sweats								
1	than								

it produced a lot of ties and tended to give higher scores to longer sentences. Note that in Figure 5, all the sentences had a score of 10.

Next, I calculated a *total agreement score* by counting the number of coders who had marked each word and then I summed the counts across all the words in a sentence. The minimum score was 0 and the maximum score was the number of coders (in this case, 10) multiplied by the number of words in the longest sentence (in this case, 42). The range of score values was much higher than the raw agreement scores, thus allowing for finer distinctions between sentences. Although longer sentences still had a better chance of scoring higher than did shorter sentences, high scores identified sentences containing the most signs and symptoms.

Finally, I calculated a *weighted agreement score* by taking the total agreement score for a sentence and dividing it by the number of words in the sentence. Scores ranged from 0-10 (the total number of coders) and can be considered a measure of a sentence's potency with regard to a theme. Those sentences that score high tend to be extremely pithy with

Figure 5. Sentences Related to Signs and Symptoms Sorted by Weighted Typicality Scores

		Typicality				
		Total		Weighted		
ID	Raw	Score	Rank	Score	Rank	Sentence
11	10	40	36	10.0	1	Sore throat, rhinorrhea, myalgias.
10	10	44	34	9.8	2	Minimal sore throat, nasal discharge.
17	10	72	25	9.0	3	My nose was congested and I was sneezing.
22	10	116	10	8.9	4	I had a headache, postnasal drip, low-grade fever, and felt horrible.
13	10	79	22	8.8	5	I felt fatigued and experienced generalized malaise and myalgias.
8	10	52	33	8.7	6	Subsequently congestion and mild cough appeared.
4	10	137	6	8.6	7	Energy level felt lower than usual, more easily tired, needed to nap several times during day.
17	10	42	35	8.4	8	I felt lethargic and tired.
12	10	134	7	8.4	9	I started coughing and then developed some congestion; a sore throat as well as a fever.
2	10	107	13	8.2	10	I was uncomfortable with fevers associated with myalgias, headaches, and a sore throat.
19	10	171	4	8.1	11	On Friday night, I developed a fever of 103° and spent that night in bed alternately with chills and sweats.
6	10	341	1	8.1	12	It began for me with a mild sore throat and a mild fever but by the second day my fever was up to 101°F; by the third day up to 103°F and my sore throat was pretty severe.
3	10	32	40	8.0	13	I was very congested.
21	10	104	15	8.0	14	I was congested, achy, and feverish, with the symptoms lasting about 4-5 days.
9	10	198	2	7.9	15	Five weeks ago, I developed sore throat that lasted 2 days followed by about 10 days of nasal congestion, clear rhinorrhea, and mild nonproductive cough.

regard to the S/S theme. For example, the first sentence in Figure 5, "Sore throat, rhinorrhea, myalgias" has a weighted agreement score of 10 because all 10 coders marked all four words as pertaining to the S/S theme.

By ranking the sentences according to total agreement scores, I can identify those sentences that coders not only agreed pertained to the theme but that also contained the most typical words. By ranking the sentences according to the weighted agreement scores, I identify typical but pithy sentences that pertain to the theme. In general, the two types of scores are quite similar. In fact, I find that the scores on all 107 sentences are correlated at $r = 0.77$ ($p < 0.001$), and the correlation of the ranks is 0.85 ($p<0.001$).

Discussion

Multicoder agreement serves a variety of functions in the analysis of text. They include:

Multicoder Agreement as Reliability

Agreement between coders tells investigators the degree to which coders can be treated as "reliable" measurement instruments (e.g., Carey et al. 1996). High degrees of multicoder reliability means that multiple coders are applying the codes in the same manner. If one coder marks half the data and another coder marks the other half, investigators need to know that both coders are performing more or less the same tasks. Normally, the reliability test is done by having both coders independently code a sample of the entire text.

Multicoder reliability is particularly important if the coded data will be analyzed statistically. If coders disagree, then the coded data are inaccurate. Discrepancies between coders is considered error and affects analysis calculations. One of the advantages of using content analysis and word dictionaries (Stone et al. 1966) is that such techniques are 100% reliable because they always mark exactly the same text. Of course, content analysis cannot judge the more subtle meaning of statements, as human coders can.

Multicoder agreement as reliability is also important for text retrieval tasks. After coding, researchers usually want to search through their texts and find examples of a particular code. An investigator who uses a single coder to mark themes relies on the coder's ability not to miss examples. Having multiple coders mark a text increases the likelihood of finding *all* the examples in a text that pertain to a given theme.

Multicoder Agreement as Validity

Mitchell (1979) noted that most qualitative analyses use multicoder agreement to measure construct validity rather than measurement reliability. Demonstrating that multiple coders can pick the same text as pertaining to a theme shows that the theme is not just a figment of the primary investigator's imagination. Validity could be further increased if informants (rather than investigators) acted as coders.[2] In the example described above, the coders are informants. As an investigator, I am not "interpreting" the informants' illness descriptions, rather I am reporting their own understandings of the abstract idea of signs and symptoms.

Multicoder Agreement as Construct Definition

Theme identification and definition is part of the inductive research process. It begins when investigators try to define the themes that they find are emerging from their texts. After reading over the corpus, team members discuss what constructs or themes they want to examine and what kind of things "count" as a particular construct. This is the process of building a codebook or theme list. In the example above, we held a group discussion where we first identified emerging themes from the texts and then chose to code for signs and symptoms, interruption of routine, and decision criteria.

At this point in the research process, the central task is to identify and discuss diverging interpretations of codes. When developing or refining codebook definitions, the differences among researchers' interpretations are not measured systematically, nor should they be. The idea for the investigative unit is to come to some agreement as to what "counts as the construct." In most cases, the constructs are "fuzzy" and require typical examples rather than strictly logical definitions.

In this article, I have suggested two additional ways to use multicoder agreement. First, multicoder agreement measures can be used as a tool to systematically describe the range, central tendency, and distribution of responses within a theme. They are particularly useful for identifying gradations of core and peripheral structures within abstract constructs. Second, multicoder agreement is a measurement device for identifying and ranking typical quotes from informants. Critics and reviewers often want to know to what degree the quotes and examples used by investigators are representative of informants' texts. Multicoder agreement measurements provide an answer.

Addendum

By far the most common responses to reviews of this article have been, "You only used 23 short paragraphs. How do I apply this to my much larger corpus of text?" and "But I can't afford to use 10 coders. How many are enough?" These are fair, but tough questions. Below, I offer some general thoughts on these matters.

How Do I Apply This to a Large Project?

First, these techniques are meant to identify the fuzzy boundaries of abstract themes and to describe core and periphery features of such concepts. They are not designed for managing text, nor for doing standard search-and-retrieve kinds of tasks. Furthermore, investigators should only apply these techniques to selected themes and need not apply them to every theme in a large codebook.

A typical large project might include five interviews with 150 informants over a year-long period. Each interview is tape recorded and transcribed. The transcripts are relatively short and average just over 10 single-spaced pages. The resulting corpus is over 7,500 pages long. After a preliminary reading of the data, the research team develops a codebook with 100 themes. The cost to have one coder (let alone multiple coders) carefully read over all the text and identify all 100 codes would be staggering. So, how can we reduce the workload while doing as little damage to the analysis as possible?

The cost of coding can be reduced by changing the number of codes, the amount of text to be coded, and the number of coders. In the case above, I would first select a manageable subset of themes from the codebook. The maximum number of themes that can be adequately described in a journal is probably closer to 10 or 12 than it is to 100. Likewise, most codebooks are hierarchical and have a smaller set of superordinate categories. Investigators should probably examine superordinate categories first and subordinate categories later, as more detail is needed.

Next, I would reduce the amount of text that needs to be coded. This can be done a priori by eliminating sections of text where the theme is unlikely to be mentioned. Depending on the specificity of a theme, investigators can often narrow a search to a very small portion of the original corpus. This tactic is particularly useful in semistructured interviews that are organized around general topics and experiences. For themes that are found throughout the corpus, researchers can identify a subsample of texts. The subsample can be selected randomly or purposefully, depending on the objectives of the investigators. For those interested in finding the full range of a theme, a purposeful sample drawn from diverse informants and circumstances might be advisable (see Patton [1990] for other types of nonrandom sampling techniques). With either of these techniques (or some combination of them), investigators should be able to identify a smaller set of data ranging from 50-200 pages.

This smaller sample can be further reduced by having investigators rapidly "scan" the corpus for particular themes. A "scanner" who is familiar with a theme quickly reads through the material and identifies any paragraph that contains the construct, using as loose a construct definition as possible. Whenever the theme is found, the paragraph is cut from the corpus and stored in a separate document of hits. To ensure that the first scanner does not bias the sample, a second scanner reads the remaining text (e.g., the "left over" column in Figure 2) and looks for theme occurrences that the

first scanner might have missed. These paragraphs are also pulled and stored in the document of hits. The analysis techniques described above can then be applied to the smaller sample of "hits."

How Many Coders Are Enough?

The answer seems to depend on: 1) the ability of the coder to identify themes; 2) the core/periphery dispersion of the theme; 3) the number of times that any given theme appears in the text; and 4) the levels of specificity investigators wish to achieve.

The last two constraints are similar to the sampling problems Bernard and Killworth (1993) solved for time-allocation research. They showed that the rarer an event's occurrence in a population, the more you have to sample to ensure that you will find it with any confidence. They also showed that optimal sample size depends on whether you want to be sure of identifying at least one occurrence of an event or whether you want to know the frequency of an event's occurrence in the population within a particular confidence interval.

In text analysis (unlike time allocation), the population is known—it's the entire corpus of text that has been collected. The unknowns are the rate of a theme's occurrence and each coder's ability to identify the theme when it occurs. It stands to reason that if a theme occurs a lot, a single coder is likely to find at least one example of it, even if the coder is not very good at identifying themes. If the theme occurs rarely, however, the likelihood of finding a single example decreases. It decreases even more if the coder isn't very good. Investigators usually are willing to miss a few examples of a theme that occurs a lot, but they can't afford to miss any examples of a theme that occurs rarely. It makes sense, therefore, that the rarer a theme's occurrence and the more important it is to find all occurrences, the more coders you want to look for it.

The number of coders needed to identify aspects of core/peripheral structures in abstract concepts depends on the levels of distinctions an investigator wants to make. I visualize themes and abstract constructs as targets made up of concentric circles. The more coders you add, the more circles you have in the target. With a single coder, you cannot distinguish between core and peripheral features of a theme. Figure 2 shows what can be learned about core/peripheral features with just 2 coders; Figure 4 shows the kinds of distinctions that can be made with 10 coders.

In hindsight, I could have used multiple coders to calculate the probability of any one coder associating a single word with a particular theme. Figure 4 shows that any single coder would probably have associated any of the words in the first column with the S/S theme. The single coder would have had a 90% probability of identifying those words in the second column; an 80% probability for those in the third column; and so forth. Investigators interested in confidently identifying and describing the peripheral aspects of a theme would probably want to use multiple coders.

Likewise, since all the quotes shown in Figure 5 were marked by all the coders, I can assume that any single coder would have found them. Of course, a single coder would have also identified quotes that were less typical. With a single coder, however, there would be no way to separate less typical quotes from the more typical ones. Increasing the number of coders would not help find more core quotes; it would allow investigators to distinguish among them in a replicable manner. It seems reasonable to assume that the less well defined a construct, the more coders are needed to describe it in detail.

Another promising technique for calculating the number of coders is consensus analysis (Romney et al. 1986). Consensus analysis is a formal analysis technique that uses the agreement found among respondents (in this case, respondents) to calculate a culturally appropriate answer key. The technique first checks to see if there is intercoder agreement. If there is group agreement, the algorithm identifies those coders who agree most with the rest—the expert coders, so to speak. Consensus analysis, then, calculates an answer key by weighing more heavily the answers of the experts. If investigators can estimate the average coder accuracy, consensus analysis provides a means to calculate the number of coders needed to achieve a given confidence level. Romney et al. (1986:325-327) suggest that the higher the average agreement among coders, the fewer coders (in some cases, as few as four) are needed to make accurate statements about a set of data.[3]

I find it helpful to recognize the limitations and advantages of single- and multiple-coder research. It seems plausible that for some tasks, investigators can rely on a single coder and that for other tasks they should use multiple coders. Ultimately, it is the investigator's responsibility to identify the goals of the research and to determine what kind of coding is required.

Notes

[1]Semantic network analysts (e.g., Osgood 1959; Danowski 1993; Jang and Barnett 1994) have long taken free-flowing texts and turned them into word-by-word matrices. Classical content analysts (Krippendorf 1980; Weber 1990) have typically had coders assign themes to fixed units of text which they then converted into unit-by-theme matrices. I have combined these two approaches to create word-by-theme matrices.

[2]I thank Roy D'Andrade for this suggestion.

[3]Schnegg and Bernard (personal communication) report that they used consensus analysis to identify typical coders in their study of students' perceptions of anthropology (Schnegg and Bernard 1996). In their study, they asked 21 graduate students to describe what they liked about anthropology. Twelve of the same graduate students then read each of the 21 descriptions and coded them for 20 different codes. Unlike the example presented above, coders didn't mark blocks of text, but simply indicated whether or not the description contained each of the codes. This coding procedure produced a description-by-code matrix filled with 1s and 0s for each of the 21 coders. Schnegg and Bernard then submitted all 12 of the coders' matrices to consensus analysis. Consensus analy-

sis showed that there was sufficient agreement among the coders to warrant a consensus. The analysis identified the "best" coders in the group (those who agreed most with everyone else) and determined the "group's" answer key. Other than this one example, I know of no published studies that use consensus analysis on textual data.

References Cited

Becker, Howard S.
1958 Problems of Inference and Proof in Participant Observation. American Sociological Review 23:652-660.

Bernard, H. Russell, and Peter D. Killworth
1993 Sampling in Time Allocation Research. Ethnology 32:207-215.

Bernard, H. Russell, and Gery W. Ryan
1998 Qualitative and Quantitative Methods of Text Analysis. *In* Handbook of Research Methods in Cultural Anthropology. H. Russell Bernard, ed. Pp. 595-646. Walnut Creek, Calif.: AltaMira Press.

Carey, James W., Mark Morgan, and Margaret J. Oxtoby
1996 Intercoder Agreement in Analysis of Responses to Open-Ended Interview Questions: Examples from Tuberculosis Research. Cultural Anthropology Methods Journal 8:1-5.

Carmines, Edward G., and Richard A. Zeller
1982 Reliability and Validity Assessment. Beverly Hills, Calif.: Sage Publications.

Danowski, James
1993 Network Analysis of Message Content. *In* Progress in Communication Science, XII. W. D. Richards and G. A. Barnett, eds. Pp. 197-221. Norwood, N.J.: Ablex.

Denzin, Norman, and Yvonna S. Lincoln
1994 Introduction: Entering the Field of Qualitative Research. *In* Handbook of Qualitative Research. Norman Denzin and Yvonna S. Lincoln, eds. Pp. 1-18. Thousand Oaks, Calif.: Sage Publications.

Dey, Ian
1993 Qualitative Data Analysis: A User-Friendly Guide for Social Scientists. London: Routledge and Kegan Paul.

Hammersley, Martyn
1992 What's Wrong with Ethnography? London: Routledge.

Jang, H.Y., and G. Barnett
1994 Cultural Differences in Organizational Communication: A Semantic Network Analysis. Bulletin de Méthodologie Sociologique 44 (September):31-59.

Krippendorff, Klaus
1980 Content Analysis: An Introduction to Its Methodology. Beverly Hills, Calif.: Sage Publications.

Lincoln, Yvonna S., and Egon G. Guba
1985 Naturalistic Inquiry. Newbury Park, Calif.: Sage Publications.

MacQueen, Kathleen M., Eleanor McLellan, Kelly Kay, and Bobby Milstein
1998 Codebook Development for Team-Based Qualitative Research. Cultural Anthropology Methods Journal 10:31-36.

Miles, Matthew B., and A. Michael Huberman
1994 Qualitative Data Analysis: An Expanded Sourcebook. 2nd ed. Thousand Oaks, Calif.: Sage Publications.

Mitchell, Sandra K.
1979 Interobserver Agreement, Reliability, and Generalizability of Data Collected in Observational Studies. Psychological Bulletin 86:376-390.

Morse, Janice M.
1994 Designing Funded Qualitative Research. *In* Handbook of Qualitative Research. Norman Denzin and Yvonna S. Lincoln, eds. Pp. 220-235. Thousand Oaks, Calif.: Sage Publications.

Osgood, Charles
1959 The Representational Model and Relevant Research Methods. *In* Trends in Content Analysis. Ithiels de Sola Pool, ed. Pp. 33-88. Urbana, Ill.: University of Illinois Press.

Patton, Michael Q.
1990 Qualitative Evaluation and Research Methods. Thousand Oaks, Calif.: Sage Publications.

Romney, A. Kimball, Susan C. Weller, and William Batchelder
1986 Culture as Consensus: A Theory of Culture and Informant Accuracy. American Anthropologist 88:313-339.

Ryan, Gery
1996 Fieldnote Searcher, 1.0. Los Angeles: Fieldwork and Qualitative Data Laboratory, University of California, Los Angeles.

Ryan, Gery, and Thomas Weisner
1996 Analyzing Words in Brief Descriptions: Fathers and Mothers Describe Their Children. Cultural Anthropology Methods Journal 8:13-16.

Schnegg, Michael, and H. Russell Bernard
1996 Words as Actors: A Method for Doing Semantic Network Analysis. Cultural Anthropology Methods Journal 8:7-10.

Spradley, James P.
1979 The Ethnographic Interview. New York: Holt, Rinehart, and Winston.

Stone, Philip J., M. S. Dunphy, and D. M. Ogilvie
1966 The General Inquirer: A Computer Approach to Content Analysis. Cambridge: MIT Press.

Truex, Gregory F.
1993 Tagging and Typing: Notes on Codes in Anthropology. Cultural Anthropology Methods Journal 5:3-5.

Weber, Robert Philip
1990 Basic Content Analysis. 2nd ed. Newbury Park, Calif.: Sage Publications.