Concept Mapping as an Alternative Approach for the Analysis of Open-Ended Survey Responses

KRISTIN M. JACKSON
WILLIAM M. K. TROCHIM
Cornell University

This article presents concept mapping as an alternative method to existing code-based and word-based text analysis techniques for one type of qualitative text data—open-ended survey questions. It is argued that the concept mapping method offers a unique blending of the strengths of these approaches while minimizing some of their weaknesses. This method appears to be especially well suited for the type of text generated by open-ended questions as well for organizational research questions that are exploratory in nature and aimed at scale or interview question development and/or developing conceptual coding schemes. A detailed example of concept mapping on open-ended survey data is presented. Reliability and validity issues associated with concept mapping are also discussed.

Qualitative text data in the form of brief, open-ended survey responses are often elicited in organizational research to gather new information about an experience or topic, to explain or clarify quantitative findings, and to explore different dimensions of respondents' experiences (Sproull, 1988). For example, they can provide details in the employees' "own words" as to why they feel stress on the job, why there may be resistance to an organizational change effort, or why employee perceptions have changed toward an organization policy. The appeal of this type of data is that it can provide a somewhat rich description of respondent reality at a relatively low cost to the researcher. In comparison to interviews or focus groups, open-ended survey questions can offer greater anonymity to respondents and often elicit more honest responses (Erickson & Kaplan, 2000). They can also capture diversity in responses and provide alternative explanations to those that closed-ended survey questions are able to capture

Authors' Note: The authors wish to thank Randall Peterson for his thoughtful suggestions on this manuscript and Alex Susskind for granting us access to the subject pool. This article was previously presented at the American Evaluation Association Conference, Honolulu, Hawaii, November 2000. Correspondence concerning this article should be addressed to Kristin M. Jackson, Johnson Graduate School of Management, Cornell University, 301 Sage Hall, Ithaca, NY 14853-6201; e-mail: kmj13@cornell.edu.

Organizational Research Methods, Vol. 5 No. 4, October 2002 307-336
DOI: 10.1177/109442802237114

(Miles & Huberman, 1994; Pothas, Andries, & DeWet, 2001; Tashakkori & Teddlie, 1998). Open-ended questions are used in organizational research to explore, explain, and/or reconfirm existing ideas.

However, the drawbacks are that open-ended survey data are often time-consuming to analyze, some respondents do not answer the questions, and coding decisions made by researchers can pose threats to the reliably and validity of the results (Krippendorff, 1980; Seidel & Kelle, 1995). Depending on the method chosen for analysis, there are different trade-offs that limit the type of inference we can draw and the strength of theory we can build from this type of data (Fine & Elsbach, 2000). In this article, we present concept mapping as an alternative method to existing text analysis techniques that is particularly well suited to the type of text generated by open-ended questions as well as to the exploratory nature of these types of questions. By blending the strengths of existing text analysis techniques and coupling them with the use of advanced multivariate statistical methods, concept mapping offers organizational researchers a way to code and represent meaning in text data based on respondent input with considerable savings in analysis time and improvement in analytic rigor. Concept mapping can be used to develop coding schemes and/or reexamine existing theoretical coding schemes, to develop follow-up interview questions and closed-ended scale items and to represent the diversity and dimensionality in meaning through analysis of the entire sample as well as assessment of subgroup differences. Concept mapping offers organizational researchers a chance to make better use of open-ended text data.

Characteristics and Analysis of Open-Ended Survey Question Text

Open-ended survey responses are extremely useful in helping to explain or gain insight into organizational issues but at the same time to generate both an interesting and challenging type of text to analyze. This type of text contains characteristics of shorter "free list" types of text as well as more "narrative" characteristics of longer text documents. The limited response length of the survey format forces respondents to express themselves in more of a concise "list" format while at the same time giving them the opportunity to "vent" or explain themselves in a short narrative form. Responses typically vary from a few phrases to a couple of paragraphs and represent a wide variety of concepts with varying frequency and detail—a "free list in context" type of text.

The analysis of this type of text poses several challenges. The "free list in context" nature of the data can make it difficult to choose an appropriate methodology. There has been considerable debate about which methods give the greatest reliability and validity in representing content in text (Gerbner, Holsti, Krippendorff, Paisley, & Stone, 1969; Pool, 1959). Open-ended survey responses are challenging because brief responses (as compared to interview transcripts or journals) are typically sparse, and the removal of context from concepts is problematic for coder understanding. The survey format does not allow the opportunity for immediate follow-up questions to improve understanding. Also, some respondents are more willing or able to express their answers, respondents typically produce many different kinds of responses, and responses can generate frequent or infrequent mention of topics that may have different importance to the respondents (Geer, 1991; Rea & Parker, 1997; Sproull, 1988).

This type of data makes standardization and reduction into codes very difficult, can make the reporting of frequencies or co-occurrences less meaningful, and requires careful justification of analysis decisions.

Ryan and Bernard (2000) have suggested that for analyzing free-flowing text, there are two broad methodological approaches that can be classified as (a) *words* as units of analysis (e.g., keywords in context [KWIC], semantic networks, cognitive maps) versus (b) *codes* as units of analysis (grounded theory, traditional content analysis, schema analysis, etc.). This distinction—between word-based and code-based methodologies—is the starting point for the methodological considerations here.

The central contention of this article is that concept mapping methodology is particularly well suited for open-ended survey text data because it combines the strengths of word-based and code-based methodologies while mitigating some of their weaknesses. As described here, concept mapping is a type of participatory text analysis that directly involves respondents or their proxies in the coding of the text. It is a multistep, hybrid method that uses original intact respondent statements as units of analysis, solicits the actual survey respondents or respondent proxies who use pile sorting to "code" the data, aggregates quantitatively across individual conceptual schemes, and enables data structure to emerge through use of multidimensional scaling and cluster analysis of the aggregated individual coding data. Because it is based on the coding schemes of the original survey respondents (or their proxies), it avoids some of the problems associated with researcher-generated coding schemes. Depending on decisions made at each step, the analysis can vary in the degree to which it is grounded in existing theory. The result is a visual representation—a map—of thematic clusters. This article discusses word-based and code-based approaches to text analysis and argues that concept mapping offers a unique blending of the strengths of each. A detailed example of the use of concept mapping on open-ended text response data is presented.

Background

Word-Based Analysis Methods

Methods using words as units of analysis have been applied in organizational research primarily in inductive qualitative studies seeking to allow data structure to emerge or to validate a thematic content analysis (e.g., Jehn, 1995). They have several strengths (see Ryan & Bernard [2000] and Mohammed, Klimoski, & Rentsch [2000] for a detailed discussion of different methods). Because they use the natural meaning embedded in language structures to represent meaning in text (Carley & Kaufer, 1993; Carley & Palmquist, 1992), they can be used to analyze both dense and sparse types of text. These methods typically employ computer-assisted coding, which has the advantages of time-saving automation, improved reliability of coding, and expanded possibilities for units of analysis—for example, the ability to map (and quantify) the relational patterns among symbols along a series of dimensions (Carley, 1993; Stone, 1997).

For example, semantic network representations count the co-occurrence of word units to identify clusters of concepts as well as the attributes (strength, direction) of relationships between them (Doerfel & Barnett, 1999; Roberts, 1997; Young, 1996).

Consider the following statements: “Tom loves working with Tim. Tim likes working with Tom but loves working with Joe.” There is a strength difference between the words *like* and *love*, and there is a difference in direction between the strengths in coworker preference. Text mapping techniques can capture these attributes (Carley, 1993). They can also map the relationships of concepts both within a respondent’s statement and between respondents along a series of dimensions (e.g., grammatical patterns or centrality) (Carley, 1997). Cognitive mapping is another word-based technique that aims to elicit individuals’ judgments about relationships between a set of important concepts about a topic in a map form that represents a mental model (Axelrod, 1976). This is useful for comparing cognitive structures about a topic between respondents (Carley, 1993).

These methods have great strength in that they use words (created by the respondents) for units of analysis, capture relationships between concepts, and allow structure in the data to emerge based on co-occurrences of words or relational similarities rather than imposing researcher bias in the form of preconceived thematic categories. Word analysis techniques often are able to represent relationships that thematic code methods cannot.

However, although convenience, reliability, and number of coding options offered by these computer-aided analyses are improved, there are two common validity criticisms. The first is that computers are unable to interpret meaning in symbols, so they do not add validity to inference (Shapiro, 1997). They do not add an understanding or explanation of the word unit in its social or psychological context—a human, often the researcher, is still required to interpret map outputs. Consider the example of coding a statement that reads, “The employee received the urgent memo and put it in the trash.” Human judgment might thematically classify or deduce, for example, that either the memo was not addressed to the recipient or the recipient did not think the memo was urgent or important. Word-unit analysis can only identify concepts or actions (e.g., received the memo, put in trash) and/or the direction of the action (e.g. memo to employee, memo to trash). These methods are useful in identifying similarities in responses between individuals but less useful in drawing conclusions about the context of the concepts or about the sample’s responses as a whole. The second criticism is that even when the analysis does identify these relationships, they continue to be based on an initial researcher judgment in selecting concepts for analysis, choosing frequency cutoffs for selection, or creating exception dictionaries for the computer to run analyses (Ryan & Bernard, 2000).

Code-Based Analysis Methods

Code-based analyses, or thematic coding methods, are often used for reducing text data into manageable summary categories or themes for making inference about a sample (Krippendorff, 1980; Weber, 1990). These methods are most commonly used with denser types of text, such as in-depth interview transcripts or employee journals, in which richer context can lead to the identification of reoccurring themes or metaphors. Although they differ in the end result they produce (e.g., grounded theory approaches seek to build theory through systematic inquiry techniques to discover themes, whereas content analysis seeks to test theory with preestablished themes [Denzin & Lincoln, 2000; Ryan & Bernard, 2000]), they share strength in making clear links between theory and data and in drawing conclusions across (rather than between, as with word-based approaches) subjects or text blocks in a sample. Because open-ended survey responses are typically a sparse, list-like type of text, content analysis has typically been applied to it over other types of code-based approaches. Therefore, we will focus on criticisms of content analysis in this context.

Content analysis has been criticized for three main reasons: (a) It relies on researcher-driven classification schemes; (b) it allows interdependence between coders; and (c) as a methodology, it offers weak reliability and validity assessments (Kelle & Laurie, 1998; Krippendorff, 1980; Weber, 1990). Preconceived categorical coding schemes have been criticized for two reasons. First, relying on coding schemes that are created a priori or through a process of induction by the researcher can create a biased method of classification that forces meaning into a framework that may or may not accurately represent the respondent's meaning. Second, because meaning is not interpreted uniformly across individuals, training coders to understand and agree on the meaning of preestablished categories often leads to intercoder discussion about certain units or categories to increase interreliability (to force "fit" between the data and the theoretical framework) of the analysis. In many contexts, content analysis will be an overly deterministic approach to finding structure in open-ended survey responses. Finally, results from this type of analysis are often reported in frequency tables, cross-tabulations, or correlations. The tendency for sporadic mention and wide variety of concepts typically generated by open-ended responses makes the validity of this kind of reporting suspect. Nonexhaustive categorical coding schemes pose a common threat to validity in content analysis (Seidel & Kelle, 1995). This problem can be compounded by the fact that respondents who are more interested in the topic of an open-ended question are more likely to answer than those who are not as interested (Geer, 1991). Therefore, frequency counts may overrepresent the interested or disgruntled and leave a proportion of the sample with different impressions of reality underrepresented in the results. If coding categories are not exhaustive or statements are coded into a category that is only semirepresentative of the respondent's reality, frequency counts and cross-tabs may underrepresent or overrepresent the distribution of meaning in the sample. It has been suggested that one way to avoid this is to calculate frequencies on the basis of the number of respondents rather than the number of comments (Kraut, 1996). However, this does not overcome the issue of preconceived and/or nonexhaustive coding schemes.

Concept Mapping as a Methodological Blend of Word-Based and Code-Based Approaches

The "free list in context" nature of open-ended survey responses makes it difficult to choose between the two approaches. On one hand, the free list characteristics of the data lend themselves nicely to word-based approaches. They can easily recognize reoccurring words or patterns of words. On the other hand, retaining the context of those concepts and a desire to analyze the responses as a set representing the whole sample make code-based analyses more appropriate.

Given the mixed strengths and weaknesses in thematic and word-mapping approaches, there are likely to be benefits from using both in concert. This could perhaps most easily be accomplished by analyzing the same data twice, once with each approach. But there are likely to be efficiencies, and perhaps even new synergies, from combining features of both approaches into new integrated methods for text analysis. This article argues that concept mapping is such an integrated approach.

There are several specific methodologies that share the name *concept mapping*, but they differ considerably both methodologically and in terms of results. One form of concept mapping (Novak, 1998; Novak & Gowin, 1997) widely used in education is essentially an informal process whereby an individual draws a picture of all the ideas related to some general theme or question and shows how these are related. The resulting map usually has each idea in a separate box or oval with lines connecting related ideas and often labeled with "connective" terms (e.g., *leads to, results from, is a part of*, etc.). This has been done in "free form," where the respondents record whatever comes to their minds, and also in a more "fixed form," where respondents construct meaning among a given set of concepts (Novak, 1998). The cognitive mapping approach described above (Carley & Kaufer, 1993) is a more statistical variant of this type of concept mapping. These methods are aimed at representing the mental models of individuals.

Another form of concept mapping (Trochim, 1989) is a more formal group process tool that includes a sequence of structured group activities linked to a series of multivariate statistical analyses that process the group input and generate maps. Instead of representing the mental models of individual respondents, it depicts an aggregate representation of the text (across respondents) in the form of thematic clusters as generated by respondents. The process typically involves participants in brainstorming a large set of statements relevant to the topic of interest and then in individually sorting these statements into piles based on conceptual similarity (a free or single-pile sort technique (Weller & Romney, 1988). The individual sort matrices are aggregated simply by adding them together. The analysis includes a two-dimensional multidimensional scaling (MDS) of the sort data and a hierarchical cluster analysis of the MDS coordinates. The resulting maps represent a "structured conceptualization" or a multidimensional graphic representation of the group's set of ideas. Each idea is represented as a dot or point, with ideas that are more similar (as determined by the multivariate analysis of the participants' input) located more proximally. Ideas (i.e., points on the map) are clustered statistically into larger categories that are overlaid on the base maps. Thus, the methods referred to as concept mapping range from informal, individual-oriented approaches to formalized, statistical group processes.

This article concentrates solely on the latter form of more formalized group-oriented concept mapping, and for the remainder of this article the term *concept mapping* will be used to refer only to this variant. Although it has typically been used in group process or evaluation applications, it has potential to analyze and represent meaning in open-ended survey responses. It is similar to word-based approaches in that it allows for visual representation of conceptual similarities through statistical mapping, but different in that it retains context by using intact respondent statements as unit of analysis instead of words. It is similar to code-based approaches because it allows human judgment to cluster these similarities thematically, but different in that it uses statistical analysis based on respondent judgments (rather than being researcher-driven) as a basis for those decisions. The role that theory plays in informing (or biasing, as it may be) the concept mapping analysis depends on decisions made by the researcher at each stage of the analysis (e.g., in creating units, choosing sorters, and finishing the cluster analysis solution).

The Concept Mapping Analysis: An Example

In considering why this method is a good match for the “free list in context” type of data, it is useful to discuss each step of the analysis through an extended example. There are five steps in the concept mapping process: (a) Create units of analysis, (b) sort units of analysis into piles of similar concepts, (c) run the MDS analysis of the pile-sort data, (d) run the cluster analyses on the MDS coordinates to decide on a final cluster solution, and (e) label the clusters. The decisions made at each stage of the analysis (e.g., about how to unitize, how to choose sorters, whom to include in the cluster replay analysis) have reliability and validity implications. After the example is presented, the reliability and validity issues associated with the concept mapping analysis will be discussed.

Content for the concept mapping analysis is generated by the survey responses. The data presented here were gathered from an open-ended question at the end of a two-page Likert-type scale questionnaire about group process. The closed-ended questions were primarily about group conflict, group knowledge, and expectations about the group outcomes. The open-ended question was intentionally placed at the end of the questionnaire to make sure the respondents had thought about the way their groups worked together. The open-ended question was intended to explore what different types or categories of group norms were operating in a sample of 22 work teams at the time of measurement. The intent of the analysis was to explore what categories or themes would emerge from the sample as a whole, not to assess which particular norms were operating in specific individual teams. These data represent the team members’ answers to the following question, which was part of a larger survey:

> What are the norms in your team? (e.g. Group norms have been defined as “written or unwritten patterns of beliefs, attitudes, communication, and behaviors that become established among team members.”)

Participants

The survey sample consisted of 22 teams with 76 respondents (a 74% response rate) from an undergraduate hotel administration course at Cornell University. Each class member was assigned to a work team of 4 or 5 people at the beginning of the semester. Each team was then given the task of conceptualizing, opening, and managing a restaurant using a restaurant management computer simulation program. The teams worked together on decision tasks such as marketing, setting menus, facilities upgrades, staffing, and so on. Final grades were based on restaurant success, a business plan, and teammate evaluations. The responses to this survey were gathered 1 month into the semester, after the teams had completed several group assignments and established some degree of working history with or understanding of each other. Respondents were instructed that they were to answer all questions on the survey with their group in mind and were given class time to complete it.

Procedure

Step 1: Creating Units of Analysis. The list-like format of open-ended survey question text lends itself to relatively easy creation of units of analysis. A unit of analysis consists of a sentence or phrase containing only one concept—units can often be lifted intact from the response because respondents tend to express one idea for each concern or opinion they list. Otherwise, unitizing is done by breaking sentences into single-concept phrases. In this way, the context of each concept is retained and is readily available to the sorters. It is important that each unit only contain one concept so that it can be considered distinct from other units—for similar reasons that double-barreled survey questions pose problems. There are two options for making unitizing decisions: They can be made (a) by two or more researchers together (researchers can also unitize decisions separately, then perform an interrater reliability check) or (b) by a group of respondents (typically three to four) who work together to create units. The result of the unitizing process is a set of single-concept statements that are placed on cards for sorting. The benefit to having the researcher do the unitizing is that involving participants can be both time-consuming and costly if it is necessary to pay them. The drawback is that the way units are created may not reflect the original intent of the respondents. But with this type of text, creating units of analysis is relatively easy. If trade-off decisions have to be made concerning the amount of access to respondents, it is recommended that respondents be involved in the sorting and cluster-solution stages of the analysis over the unitizing process.

In this example, the researchers did the unitizing. The respondents' answers to the group norms question were, on average, a short paragraph of one to three sentences and contained different ideas ranging from "don't know" to statements about communication, group roles, personalities, and so on. Each answer was broken down into separate statements containing one idea about a group norm. For example, one response was, "We have a solid belief that we all want to do well on this project and will work as hard as possible to achieve a good grade and learn a lot from this project."

This response was broken down into three separate statements: (a) We have a solid belief that we all want to do well on this project, (b) we will work as hard as possible to achieve a good grade, and (c) to learn a lot from this project. This was done for the entire data set and resulted in 156 statements. To ensure that each unit of analysis would be considered independently of the others, each statement was given a random number generated by a random number function and placed on a 2- by 4-inch card.

Step 2: Sorting. The next step in the concept mapping process is to have a group of at least 10 sorters code these units by sorting them into piles of similar statements.[1] Sorters are given instructions to put each card in a pile with other cards that contain statements they think are similar. There is no limit to the number of piles they can create. Their only limitation is that they cannot create a "miscellaneous" pile. Any statement they do not judge to be similar to any other statement should be left in its own pile. This improves the fidelity of the data by excluding the possibility of a "junk" cluster after the final analysis. Finally, they were asked to give each pile a name that they thought most accurately represented the statements in it.

In general, it is recommended that the original respondents do the sorting to ensure maximum representativeness of the structure that emerges from the MDS analysis. Using the original respondents eliminates the possibility that the researcher will impose his or her interpretation of meaning on the data (as in thematic coding schemes or concept selection in word-analysis methods). However, there are times and circumstances that may make using respondents very difficult. For example, there may be limited access to a sample (e.g., permission to administer the survey only); the respondents may have very limited time to spare (e.g., CEOs); or using the original respondents may be a source of contamination for a follow-up survey (e.g., prime them to discuss certain issues that will be measured again in the future). In such cases, proxy sorters can be acceptably substituted based on careful consideration of the following criteria: (a) how their backgrounds and experiences are similar/different to the respondents and how that might influence their interpretation of units, (b) any theoretical background/understanding underlying the research topic that they have in common with the respondents and how a deeper/lesser understanding of that theory may influence interpretation, and (c) the degree to which existing theoretical frameworks can provide a basis for comparison in gauging the degree of difference between respondent content and proxy sorter groupings. Obviously, using the original respondents will allow for less bias from existing theory or research frameworks. When the original respondents cannot be used, it is important that it be made publicly known, that the proxy sorters be carefully selected, and that caution be used in drawing inference from the final maps—as the case would be with content analysis.

In this example, graduate students were used as proxy sorters instead of the original respondents. This trade-off was made to eliminate contamination of the respondent sample for a second time measurement about group norms. The graduate student sorters were selected as appropriate based on their familiarity with the content of the research question—in this case, teamwork and group processes. Each had also been involved in multiple classroom team experiences. Even though the reality of the survey respondents was not technically available to the sorters, the general social experience was. In addition, both the sample and the proxy sorters had similar "theoretical" understandings about what group norms are from taking courses in the same school. Finally, because there is an extensive literature about group norms, if there had been a major mismatch between the content of the respondents' statements and the way the proxy sorters grouped them, it would have stood out to the researchers. For example, previous research has shown that groups generally develop norms that govern their "task"-related interaction (e.g., work strategies) as well as their "social"-related interaction (e.g., how they handle personality clashes) (Guzzo & Shea, 1992). If the proxy sorters were consistently placing statements exclusively about personality clashes in clusters labeled *work strategies*, there would be reason to further investigate why the proxies were doing this. Each of the 10 sorters was given a packet with the stack of 156 cards, Post-it Notes to put labels on their piles, and rubber bands to bind their piles and labels.

Step 3: The Multidimensional Scaling Analysis. Using respondents or respondent proxies to code the data allows structure to emerge from the MDS analysis based on aggregated individual understanding (in the form of similarity judgments) of original responses. The first step is to create a matrix for each sorter. In this example, a 156 × 156 binary square matrix (rows and columns represent statements) was created for each coder. Cell values represented whether (1) or not (0) a pair of statements was sorted by that coder into the same pile. The second step is to aggregate the similarity judgments of the sorters by adding all 10 of the individual matrices together. From that aggregated matrix, MDS created coordinate estimates and a two-dimensional map[2] of distances between the statements based on the aggregate sorts of the 10 coders as shown in Figure 1. Each statement on the map is represented by a point (accompanied by the statement number). The distance between the points represents the estimates from MDS of how similar the statements are judged to be by the sorters. Points that are farther apart on the map were sorted together less often than those that are closer

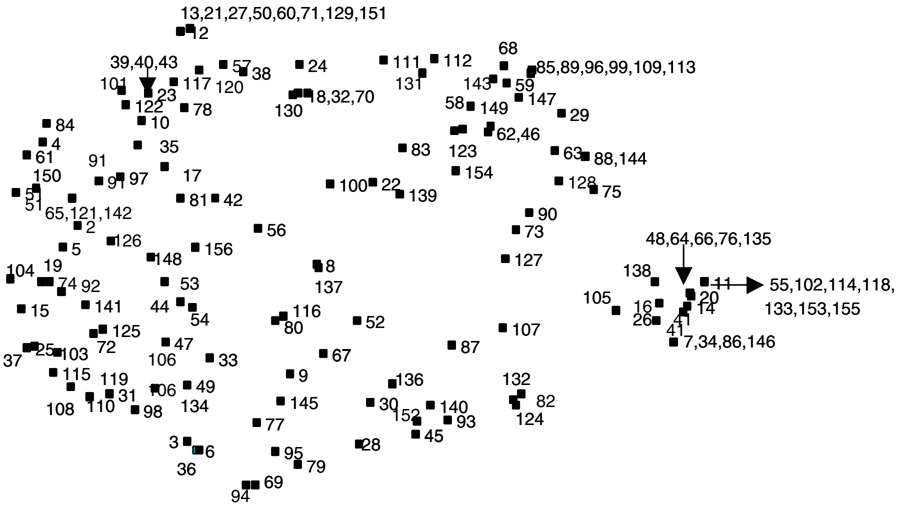

Figure 1: Multidimensional Scaling Point Map of Statements
Note. Similar statements are closer together.

together. The position of each point on the map (e.g., top, bottom, right, left) is not important—only the distance or spatial relationship between the points.

Step 4: Choosing a Final Cluster Solution. The next step in this analysis is to determine the appropriate number of clusters that represent a final solution for the data. In this example, hierarchical agglomerative cluster analysis using Ward's algorithm was used on the MDS map coordinates to determine how the statements cluster together based on similarity. This type of cluster analysis is most helpful in identifying categories when the structure of categories is not already known (Afifi & Clark, 1996). A 20-to-8 cluster replay analysis (Concept-Systems, 1999) was done to decide on the appropriate cluster solution. This analysis begins with each statement as its own cluster and tracks the merging of the statements into clusters up to a 20-cluster solution. The output from this analysis generates two decision tools[3]: (a) a list of the statements in the 20-cluster solution with their bridging values,[4] presented in Table 1; and (b) the merging of clusters for each cluster solution (a list version of a dendogram), presented in Table 2. The two decision tools together provide a statistical basis to guide human judgment about the goodness of fit for the final cluster solution.

Each proposed cluster solution is then examined to determine how appropriate the merging or splitting of statement groups is. A final cluster solution is chosen by examining all of the cluster solutions within a certain range[5] to determine how appropriate the merging or splitting of statement groups is. It is important to note that the central decision being made here is on the number of clusters to select—the hierarchical cluster tree structure is entirely determined by the analysis and is not the subject of researcher discretion or judgment. The reason such judgment is required with cluster analysis is that there is no sensible mathematical criterion that can be used to select the number of clusters. This is because the "best" number of clusters depends on the level of specificity desired and the context at hand, factors that can only be judged subjec-

Table 1
20-to-8 Cluster Replay Solution Output

Cluster	*Statement*
Number	*(With Statement Number)*
1	(153) We have not had any opportunity to establish group norms.
	(155) We haven’t spent that much time together.
	(64) We haven’t actually worked together yet (on the project) but through a few interactions.
	(102) We haven’t done that much together.
	(114) We have done little group work so far.
	(118) We do not have any established yet.
	(133) Did not meet enough to figure out.
	(20) Not sure yet.
	(48) We have not met enough to determine.
	(76) I’m not sure I can answer yet?
	(1) Don’t know yet.
	(11) Don’t know yet.
	(55) Don’t know yet.
	(66) I don’t know yet.
	(135) Don’t know yet.
2	(14) None yet.
	(41) None yet.
	(7) N/A
	(34) N/A
	(86) N/A
	(146) N/A
	(16) We haven’t had that much time to work together.
	(26) There is not really a “group” goal yet.
	(138) Our group has limited work time; I’ll get back to ya!
	(105) We haven’t discussed anything regarding beliefs, attitudes.
3	(120) To communicate our beliefs until everyone is satisfied.
	(23) Open.
	(39) Openness.
	(40) Open.
	(43) Openness.
	(101) To listen to each other.
	(78) Everyone is not afraid to voice their opinion.
	(57) Discussion.
	(117) Norms include: group discussions.
	(122) Hearing one another’s opinions.
	(35) Consideration for opinion of others.
	(38) Make sure everyone understands what is going on.
	(10) We are able to discuss our ideas, nobody holds back.
4	(12) No lack of communication.
	(13) Communicative.
	(21) So far, we seem to communicate well.
	(27) Everyone communicates everything.
	(60) Communicating among the group about our tasks.
	(71) Communication.
	(129) Communicative.
	(151) The norms of my team: our strong communication skills.
	(50) We want to have a channel of open communication.
5	(32) We let everyone have a say in decisions.
	(70) Everyone should agree on a final decision.
	(18) We are pretty much a democratic group.
	(130) We must make key decisions as a group.
	(24) Majority rules.

(continued)

Table 1 continued

Cluster	*Statement*
Number	*(With Statement Number)*
6	(112) Everyone should contribute.
	(111) We review work all together.
	(131) We are all contributing members of the team.
7	(85) Splitting up the work fairly.
	(89) Dividing up the work as fairly as possible.
	(113) Work should be distributed evenly.
	(96) We divide work.
	(99) Norms include: divvying up the work.
	(109) Everyone agrees to split up work evenly.
	(68) Each person must contribute.
	(59) We are all expected to contribute equally.
	(143) To all contribute equally.
	(29) Dividing up the tasks as fairly as possible.
	(147) Doing your portion of the work.
	(149) Make sure to be a part of the group.
8	(58) Complete all work that is assigned to you.
	(46) Do what's asked of you.
	(83) That everyone will help when they are needed.
	(62) The group norm: take responsibility.
	(100) Jeremy and I seem to argue a lot (compromising in the end) and take charge.
	(154) Responsibility.
	(123) Just that everyone works hard.
	(22) We are all expected to do our work on time.
	(139) The group norm: do your work.
9	(90) There will probably end up a certain structure—like a leader, secretary, caller, etc.
	(88) There is one member that volunteers to collect everyone's papers and e-mail lists.
	(144) Larissa—organizer/recorder. Lisa & Christopher—implementers. Me (Drew)—idea person
	(128) The girl offers to do the grunt work and contributes but isn't too forceful.
	(73) Who the leaders are.
	(127) Probably will have one leader that leads w/o meaning to, someone who is an organizer.
	(75) Only a few take initiative to write or do stuff.
	(63) One person or two assign the work and the others follow.
10	(91) Everyone respects each other's ideas.
	(65) We are always respectful (well . . . so far).
	(121) Respect.
	(142) Mutual respect for each other.
	(150) Cooperative.
	(97) Actually, everyone seems to work very well together.
	(126) We all seem willing to work together.
	(2) To work together.
	(51) To always help each other w/patience.
	(5) Norms include: lots of interactive helping.
11	(4) Must compromise.
	(84) So far, we are all willing to compromise.
	(61) Compromising is sometimes hard.
12	(17) If everyone does not agree, then there should be no unnecessary hostility.
	(42) No one is better than another person.
	(81) So far, we are all willing to work together.
13	(156) Good meshing of attitudes.
	(54) We all seem willing to meet.
	(148) I think a norm in our team is honesty.

Table 1 continued

Cluster	*Statement*
Number	*(With Statement Number)*
	(56) Everyone is open-minded.
	(53) Showing up to meetings.
	(44) We are all expected to do a quality job.
14	(92) Attendance.
	(74) To be there.
	(19) We all believe in well thought out decisions.
	(104) To help get all the work done.
	(15) We all know we have a job to do.
15	(25) We all believe in hard work.
	(115) The desire to achieve.
	(72) Hard work.
	(37) Must work hard.
	(103) Get the job done.
	(141) We all want what is best for the group.
	(125) We try to get things done as quickly as possible.
16	(31) The desire to be the best.
	(110) Positive attitudes.
	(119) The norms of my team: our positive attitude.
	(98) Everyone keeps a positive attitude.
	(33) My group norm is, and I am very happy about it, professionalism.
	(47) The group norm: do a good job.
	(49) Joking.
	(134) Humor is a key.
	(106) I think a norm in our team is fun.
	(108) Positive outlook.
17	(69) Our team wants to learn.
	(36) Wants to exceed.
	(94) We want to learn a lot from this project.
	(6) We have a solid belief that we all want to do well on this project.
	(3) We will work as hard as possible to achieve a good grade.
	(79) We want to run an upscale restaurant.
	(95) We all want to do our work well, and get it behind us.
	(28) Our team and project will have our own style and attitude.
18	(8) Our group seems fairly diverse with a multitude of attitudes and backgrounds.
	(137) We are a good blend of different people.
	(116) Seemingly compatible.
	(80) We get along well.
	(52) Our group members have these characteristics: quiet, unsure, persuadable, leader, intelligence, optimistic.
19	(9) Kindness.
	(77) Our team seems to contain dedicated individuals who want to do well in the class.
	(30) Must be calm.
	(67) Very outgoing.
	(136) Must be personable.
	(145) Regular meetings will be held.
20	(124) I think there are cliques starting already.
	(93) Male-based.
	(82) Everyone has their own personal objectives it seems like.
	(152) Intelligent.
	(140) For some reason we all agree we are the team with the lowest cumulative IQ.
	(45) We are all transfers.
	(107) 2 kids offer very little input and say “whatever” a lot.
	(132) We seem to be divided on most issues.
	(87) Some keep talking and do not listen to others when they talk.

Table 2
Cluster Replay Solutions: From 20 to 8

At Cluster Solution	*Clusters Merged*
19	14, 15
18	10, 11
17	5, 6
16	12, 13
15	1, 2
14	18, 19
13	8, 9
12	3, 4
11	14, 15, 16
10	10, 11, 12, 13
9	18, 19, 20
8	5, 6, 7

tively. So this issue of cluster number selection illustrates how concept mapping is a blending of human judgment based on the more objective mathematical algorithm of cluster analysis.

It was decided by the researchers[6] that a 15-cluster solution was most appropriate. Original respondents or proxies were not used primarily because of resource constraints and because the purpose of the analysis was merely to create a heuristic-like representation of how the class described the norms of their team. The final cluster solution map is represented in Figure 2. This decision was based on the desirability of not splitting Clusters 1 and 2 (the "don't know" clusters). All previous splits were deemed reasonable.

Step 5: Labeling the Clusters. The final step in the analysis is to identify the sort-pile label (i.e., the labels each sorter assigns to the piles of sort cards) that best represents each cluster. A centroid analysis is used to select a label for each cluster from the pile names generated by the sorters. A *centroid* is defined as "the point whose coordinates are the means of all the observations in that cluster" (Afifi & Clark, 1996, p. 392).

Three steps are involved in the computation. First, a centroid is computed for each of the clusters on the map. For each cluster, this is the average x and the average y value of the MDS coordinates for each point in the cluster. Second, a centroid value is computed for every sort-pile label for every sorter. For each sort-pile label, this is the average x and the average y value of the MDS coordinates for each statement point that the sorter placed in that pile. Finally, for each cluster, the Euclidean distance is computed between the cluster's centroid and the centroid of each pile label. The pile label with the smallest Euclidean distance is considered the best fitting one. The closest 10 pile labels constitute a "top-10" list of pile names that offers the best choice and the 9 most reasonable alternative choices. It is then up to the decision makers to examine the list of possible pile labels and decide if any of them is more appropriate to the statements in the pile than the label that was statistically chosen by the software. If none of the pile labels completely captures the theme of the cluster, a label can also be manually entered. This decision process is also indicative of the blending of objective statistical algorithm and human judgment involved that makes concept mapping a blend between word-based and code-based approaches.

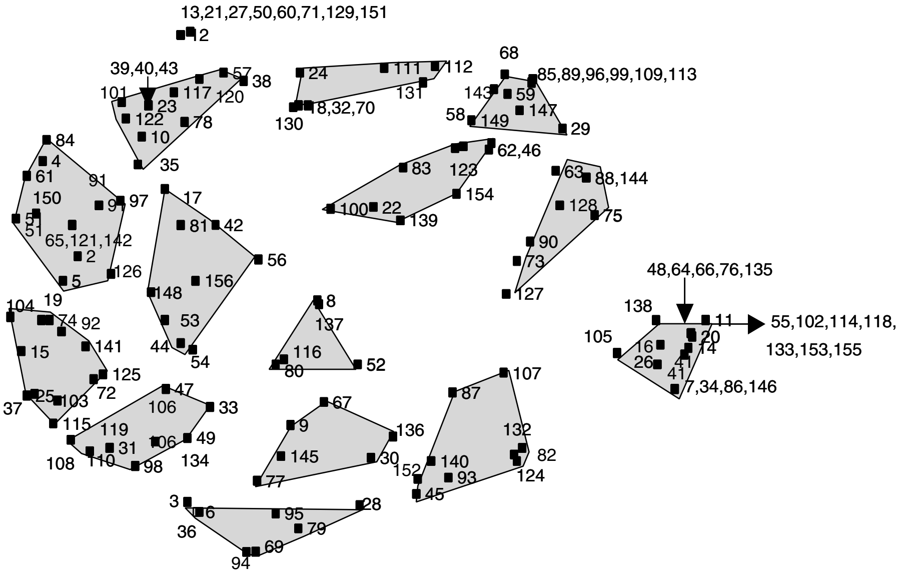

Figure 2: Final Cluster Solution With Statement Points

The decision about what to label each cluster was made by the researchers. The resulting 15 categories of group norms that emerged from this analysis were Role Definition, All Do Their Part, Equal Distribution of Work, Consensual Decision Making, Communication, Open to Ideas/Listening, Respect/Compromise, Cooperation, Work Hard, Compatibility, Keeping Positive Attitude, Personality Characteristics, Achievement, Team Characteristics, and Don't Know. The final map with labels is presented in Figure 3, and the corresponding cluster statements are presented in Table 3. In interpreting the final map, keep in mind that each statement on the map is represented as a point that is included in a cluster. The proximity of the clusters represents how similar the statements in them were judged to be by the coders/sorters. Clusters that are farther apart on the map contain, in general, statements that were sorted together less often than those that are closer together. The position of each cluster on the map (e.g., top, bottom, right, left) is not meaningful—only the distance or spatial relationship between them. The breadth or tightness (i.e., shape and size) of a cluster generally represents whether it is a broader or narrower conceptual area.

Interpretation

These results can be interpreted in several ways. The most basic interpretation is that through this analysis, there has emerged a theory-based representation of 15 categories, including the classification of text content within these categories, which represents the range of norms in the teams of this sample. Similar to traditional content analysis, the concepts have been coded into themes or categories (based on the clusters). Similar to word-mapping analyses, the concepts' positions on the maps represent relationships of similarity—both at the cluster and the unit-of-analysis level.

(Text continues on page 327)

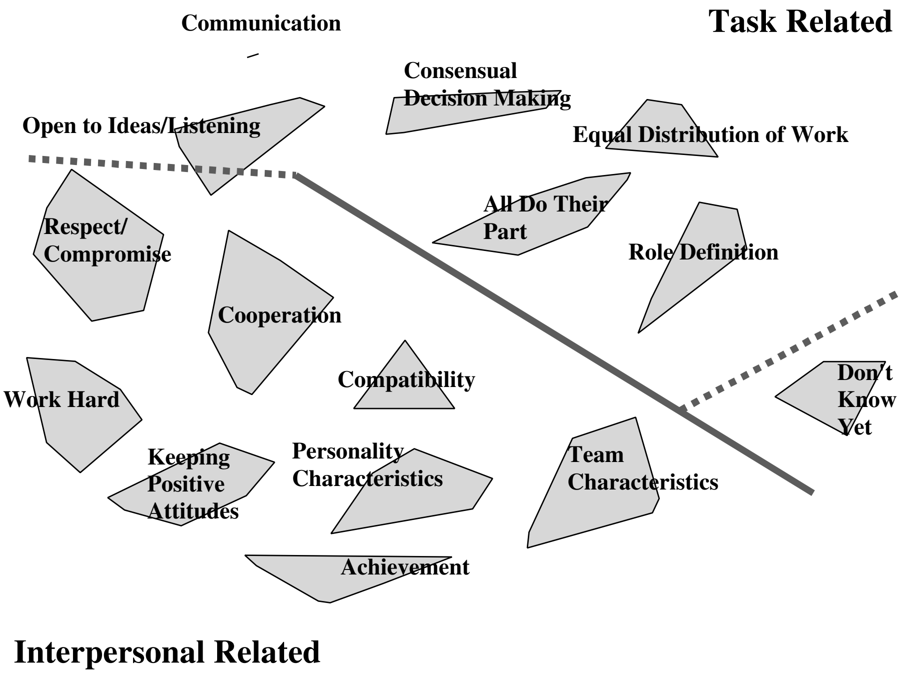

Figure 3: Final Map With Cluster Labels and Global Interpretation

Based on this relational positioning or "structured conceptualization" of the data, it is also possible to take this analysis one step further by examining the map for "regions" of meaning. A region on the map represents clusters that can be meaningfully grouped together more tightly than they are with other regional groups of clusters. This is apparent by more separation, or white space, between regions of the map. Decisions about regional distinctions can be driven by theoretical preconceptions or simply through discussion.

For example, the solid and dotted lines overlaid on Figure 3 represent one interpretation of how group norms might be conceptualized at more of a "global" level. From the literature on group norms, we know that group work is composed of task-process activities related to the division and coordination of labor but also depends on interpersonal interaction among its members (Guzzo & Shea, 1992). There is a clear division along these lines between the "east" (more interpersonal) and "west" (more task-related) sides of the map, such that if it is folded on the diagonal, the two regions fall on either side. Interestingly, there are two clusters in the "northwest" corner of the map that seem to bridge task-process- and interpersonal-process-related norms: "communication" and "openness to ideas/listening." There is also a clear separation between the "don't know" concept and the rest of the concepts on the map. As mentioned above, the interpretation of these results is constrained by the amount of involvement by the researchers. However, our involvement illustrates how this analysis can be grounded quite heavily in existing theory if the research objective allows it.

Table 3
Final Cluster Solution

Cluster Name	*Statement Bridging Value*
Cluster 1: Don't know	
(153) We have not had any opportunity to establish group norms.	.00
(155) We haven't spent that much time together.	.00
(64) We haven't actually worked together yet (on the project) but through a few interactions.	.00
(102) We haven't done that much together.	.00
(114) We have done little group work so far.	.00
(118) We do not have any established yet.	.00
(133) Did not meet enough to figure out.	.00
(20) Not sure yet.	.00
(48) We have not met enough to determine.	.00
(76) I'm not sure I can answer yet?	.00
(1) Don't know yet.	.02
(11) Don't know yet.	.02
(55) Don't know yet.	.02
(66) I don't know yet.	.02
(135) Don't know yet.	.02
(14) None yet.	.03
(41) None yet.	.04
(7) N/A	.09
(34) N/A	.09
(86) N/A	.09
(146) N/A	.09
(16) We haven't had that much time to work together.	.10
(26) There is not really a "group" goal yet.	.16
(138) Our group has limited work time; I'll get back to ya!	.22
(105) We haven't discussed anything regarding beliefs, attitudes.	.32
Average bridging	.05
Cluster 2: Open to ideas/listening	
(120) To communicate our beliefs until everyone is satisfied.	.29
(23) Open.	.30
(39) Openness.	.30
(40) Open.	.30
(43) Openness.	.30
(101) To listen to each other.	.33
(78) Everyone is not afraid to voice their opinion.	.34
(57) Discussion.	.35
(117) Norms include: group discussions.	.35
(122) Hearing one another's opinions.	.36
(35) Consideration for opinion of others.	.37
(38) Make sure everyone understands what is going on.	.40
(10) We are able to discuss our ideas, nobody holds back.	.44
Average bridging	.34
Cluster 3: Communication	
(12) No lack of communication.	.18
(13) Communicative.	.18

(continued)

Table 3 continued

Cluster Name	*Statement Bridging Value*
Cluster 3: Communication	
(21) So far, we seem to communicate well.	.18
(27) Everyone communicates everything.	.18
(60) Communicating among the group about our tasks.	.18
(71) Communication	.18
(129) Communicative.	.18
(151) The norms of my team: our strong communication skills.	.18
(50) We want to have a channel of open communication.	.21
Average bridging	.19
Cluster 4: Consensual decision making	
(32) We let everyone have a say in decisions.	.36
(70) Everyone should agree on a final decision.	.36
(18) We are pretty much a democratic group.	.37
(130) We must make key decisions as a group.	.39
(112) Everyone should contribute.	.40
(24) Majority rules.	.42
(111) We review work all together.	.44
(131) We are all contributing members of the team.	.45
Average bridging	.40
Cluster 5: Equal distribution of work	
(85) Splitting up the work fairly.	.22
(89) Dividing up the work as fairly as possible.	.22
(113) Work should be distributed evenly.	.22
(96) We divide work.	.22
(99) Norms include: divvying up the work.	.22
(109) Everyone agrees to split up work evenly.	.22
(68) Each person must contribute.	.25
(59) We are all expected to contribute equally.	.30
(143) To all contribute equally.	.33
(29) Dividing up the tasks as fairly as possible.	.39
(147) Doing your portion of the work.	.40
(149) Make sure to be a part of the group.	.47
Average bridging	.29
Cluster 6: All do their part	
(58) Complete all work that is assigned to you.	.49
(46) Do what's asked of you.	.49
(83) That everyone will help when they are needed.	.50
(62) The group norm: take responsibility.	.53
(100) Jeremy and I seem to argue a lot (compromising in the end) and take charge.	.57
(154) Responsibility.	.57
(123) Just that everyone works hard.	.59
(22) We are all expected to do our work on time.	.60
(139) The group norm: do your work.	.67
Average bridging	.56

Table 3 continued

Cluster Name	*Statement Bridging Value*
Cluster 7: Role definition	
(90) There will probably end up a certain structure—like a leader, secretary, caller, etc	.36
(88) There is one member that volunteers to collect everyone's papers and email lists.	.37
(144) Larissa—organizer/recorder. Lisa & Christopher—implementers. Me (Drew)—(idea person.	.37
(128) The girl offers to do the grunt work and contributes but isn't too forceful	.41
(73) Who the leaders are.	.43
(127) Probably will have one leader that leads w/o meaning to, someone who is an organizer.	.44
(75) Only a few take initiative to write or do stuff.	.47
(63) One person or two assign the work and the others follow.	.50
Average bridging	.42
Cluster 8: Respect/compromise	
(91)Everyone respects each other's ideas.	.35
(65) We are always respectful (well...so far).	.38
(121) Respect.	.38
(142) Mutual respect for each other.	.38
(150) Cooperative.	.44
(4) Must compromise.	.44
(84) So far, we are all willing to compromise.	.45
(97) Actually, everyone seems to work very well together.	.46
(126) We all seem willing to work together.	.47
(2) To work together	.49
(61) Compromising is sometimes hard.	.52
(51) To always help each other w/patience.	.63
(5) Norms include: lots of interactive helping.	.63
Average bridging	.46
Cluster 9: Cooperation	
(17) If everyone does not agree, then there should be no unnecessary hostility.	.44
(156) Good meshing of attitudes.	.46
(42) No one is better than another person.	.50
(54) We all seem willing to meet.	.50
(81) So far, we are all willing to work together.	.50
(148) I think a norm in our team is honesty.	.52
(56) Everyone is open-minded.	.53
(53) Showing up to meetings.	.57
(44) We are all expected to do a quality job.	.62
Average bridging	.52
Cluster 10: Work hard	
(25) We all believe in hard work.	.43
(115) The desire to achieve.	.45

(continued)

Table 3 continued

Cluster Name	*Statement Bridging Value*
Cluster 10: Work hard	
(72) Hard work.	.47
(37) Must work hard	.48
(103) Get the job done.	.49
(141) We all want what is best for the group.	.49
(125) We try to get things done as quickly as possible.	.51
(92) Attendance.	.53
(74) To be there.	.54
(19) We all believe in well thought out decisions.	.62
(104) To help get all the work done.	.64
(15) We all know we have a job to do.	.67
Average bridging	.53
Cluster 11: Keeping positive attitude	
(31) The desire to be the best.	.39
(110) Positive attitudes.	.43
(119) The norms of my team: our positive attitude.	.43
(98) Everyone keeps a positive attitude.	.45
(33) My group norm is, and I am very happy about it, professionalism.	.48
(47) The group norm: do a good job.	.50
(49) Joking.	.52
(134) Humor is a key.	.52
(106) I think a norm in our team is fun.	.54
(108) Positive outlook.	.60
Average bridging	.48
Cluster 12: Achievement	
(69) Our team wants to learn.	.38
(36) Wants to exceed.	.39
(94) We want to learn a lot from this project.	.40
(6) We have a solid belief that we all want to do well on this project.	.41
(3) We will work as hard as possible to achieve a good grade.	.44
(79) We want to run an upscale restaurant.	.47
(95) We all want to do our work well, and get it behind us.	.57
(28) Our team and project will have our own style and attitude.	.63
Average bridging	.46
Cluster 13: Compatibility	
(8) Our group seems fairly diverse with a multitude of attitudes and backgrounds	.52
(137) We are a good blend of different people.	.52
(116) Seemingly compatible.	.54
(80) We get along well.	.54
(52) Our group members have these characteristics: quiet, unsure, persuadable; leader; (intelligence; optimistic	.58
Average bridging	.54

Table 3 continued

Cluster Name	*Statement Bridging Value*
Cluster 14: Personality characteristics	
(9) Kindness.	.51
(77) Our team seems to contain dedicated individuals who want to do well in HA 136	.51
(30) Must be calm.	.53
(67) Very outgoing.	.57
(136) Must be personable.	.57
(145) Regular meetings will be held.	.83
Average bridging	.59
Cluster 15: Team characteristics	
(124) I think there are cliques starting already.	.46
(93) Male-based.	.51
(82) Everyone has their own personal objectives it seems like.	.52
(152) Intelligent.	.53
(140) For some reason we all agree we are the team with the lowest cumulative IQ.	.54
(45) We are all transfers.	.54
(107) 2 kids offer very little input and say “whatever” a lot.	.61
(132) We seem to be divided on most issues.	.71
(87) Some keep talking and do not listen to others when they talk.	1.00
(140) For some reason we all agree we are the team with the lowest cumulative IQ.	.54
Average bridging	.60

Reliability and Validity

Concept mapping presents a visually appealing classification of text data, but more important, it also offers several advantages over existing word-based and code-based methods in terms of reliability and validity. Decisions made at each stage of the analysis can either increase or decrease how representative results are of the sample versus how much they are informed by existing theory. Krippendorff (1980) has outlined a useful framework for discussion of the reliability and validity of content analysis that will be applied here.

Reliability

Reliability has been defined as obtaining data from research that represent “variations in real phenomena rather than the extraneous circumstances of measurement” (Krippendorff, 1980). Krippendorf (1980) discusses three types of reliability in content analysis: stability, reproducibility, and accuracy. *Stability* refers to the degree to which the same coder at different times codes the same data in a similar manner. *Reproducibility* refers to the extent to which similar results can be reproduced in different times and locations and with different coders. *Accuracy* refers to the amount of error (intraobserver inconsistencies, interobserver disagreements, and systematic deviations from the standard).

The reliability of concept mapping can be assessed in several ways (Trochim, 1993). The stability of the method can be addressed, for example, by having each sorter repeat his or her sort at a later time, then assess the correlation between the two sort matrices. Reproducibility refers to intercoder reliability and can be assessed by correlating each individual sorter's matrix against the entire sample of sorters (a form of item-total reliability assessment that in effect treats the aggregate as the "errorless solution"). This has been discussed in detail by Trochim (1993). Intercoder reliability is especially important to consider when making decisions about whom to choose as sorters because it has implications for the validity of results. In the analysis of text data, meaning is constructed in readers' minds through an interaction of their interpretation of the text and their own experiences or reality (Findahl & Hoijer, 1981; Lindkvist, 1981). Therefore, if sorters with a different experience or background sort the responses, they may interpret them differently than the original respondents intended. The most obvious way to minimize the potential for misunderstanding is to have the original respondents serve as sorters. We highly recommend this. However, there are times and circumstances that may make that difficult. If so, we recommend careful selection of proxies per the guidelines described above.

One major reliability benefit to the concept mapping method is that the accuracy of each coder is not a problem compared to more traditional notions of intercoder reliability. There is no preestablished category structure to which to conform. Each sorter makes his or her own judgments about how many categories to create, what each category should contain, and what each category should be called. Therefore, intersorter error or disagreement is taken into account through statistical aggregation of the similarity judgments of the individual coders. Occasionally one coder will generate a sort that is radically different from the other coders' (for example, one coder does a sort with only 2 piles, whereas the rest of the coders generated 10 to 12 piles). As discussed above in regard to stability, when individual sort matrices are correlated against the aggregate, any outliers will be identified by very low correlations. At that point, the researcher must make a judgment as to whether the sorter followed instructions and can do a reproducible sort. A radically different sort may represent a legitimate interpretation of similarity between concepts, but often it represents that the coder did not understand the purpose or instructions of the sort. These situations must be carefully considered, and decisions about including or excluding the sort must be well justified.

In addition to the reliability issues discussed above, Krippendorff (1980) identified four more common reliability concerns: (a) Some units are harder to code than others; (b) some categories are harder to understand than others; (c) subsets of categories can sometimes be confused with larger categories; and (d) individual coders may be careless, inconsistent, or interdependent. Each of these will be discussed.

Some Units Are Harder to Code Than Others and Some Categories Are Harder to Understand Than Others. A strength of the concept mapping method is that it offers a nonforced, systematic way for sorters to create categories that they understand from their unique perspectives. Sorters are given instructions to create as many or as few piles as it takes to group the statements (i.e., units) according to how similar they are to each other. Therefore, sorters can create their own categories and place "hard to categorize" statements in whichever pile they feel is appropriate. If they feel they do not understand how to categorize a particular statement, they have the option of leaving it in its own pile instead of forcing it into a preestablished category.

The software also generates a useful statistic called a "bridging value" that helps the researcher to identify the degree to which any given statement is related to ones that are similar in meaning or tend to "bridge" a more diverse set of statements. Statements that are difficult to sort will show up as having a high bridging value (Concept-Systems, 1999). Bridging values can be used at several points in the analysis. For example, while choosing the final cluster solution, the decision makers can examine bridging values of each statement as a guide to whether that statement should be included in a different cluster. Bridging values are also available for each cluster (see Table 3). Cluster bridging values are an indicator of how cohesive the statements are with the other statements around them—it is the average bridging value of all statements in a cluster (Concept-Systems, 1999).

Subsets of Categories Can Sometimes Be Confused With Larger Categories. The emergence of subcategories in the concept mapping methodology is not an issue. Each statement (or unit) is placed on a separate card and represents only one concept (in this example, one group norm). Clusters of points on the map represent patterns of shared similarity in meaning but do not necessarily represent exact agreement on the meanings. The results from all of the sorts are aggregated to give us the most accurate "model" of reality based on the sorters' perspectives in aggregated form. Therefore, instead of subcategories, statements that are understood as categorically/thematically similar but conceptually different will be sorted into separate piles as understood by the sorters and will most likely emerge as proximally located on the map through the MDS analysis of the aggregated sort results.

Individual Coders May Be Careless, Inconsistent, or Interdependent. Another strength that concept mapping brings to reliability is that sorters (coders) in this methodology are always independent of each other. There is no need for sorters to discuss how to conceptualize problematic concepts or reach a greater degree of interrater agreement. In our experience and as mentioned by others (Boster, 1994), because sorters are conceptualizing their own similarity judgments, their attention level and enthusiasm for the task tends to be high—unless they are given too many statements to sort. As mentioned above in the discussion of stability, reproducibility, and accuracy, carelessness or inconsistencies can easily be identified by low correlations between matrices.

Validity

Qualitative data pose an interesting obstacle to validity. If we know nothing about the subject, we cannot capture meaning effectively—conversely, if we know a lot about the subject, our own biases might interfere (Krippendorff, 1980; Miles & Huberman, 1994; Patton, 1990). Concept mapping helps to ease this tension somewhat by combining statistical analysis and human judgment. The degree to which theory guides the concept mapping analysis is introduced through choices about whom to include as decision makers in the analysis. The more respondents are used at each stage of the analysis, the greater the resulting map represents their collective understanding of the topic at hand. Because concepts are social constructions, there is really no way to establish a standard by which to judge the degree of error (Krippendorff, 1980). The main strength that concept mapping offers to validity is that by using multidimensional scaling and cluster analysis to represent the similarity judgments of multiple coders, it allows meaning and relationships to emerge by aggregating the "biases" or "constructions" of many. Instead of arbitrary bias and potentially forcing values of the investigator with a priori categories or semantic encoding choices, sorting concepts allows for a web of concept relationships to be represented by sorters immersed in the context of their own social reality.

An additional value of concept mapping is that by having multiple sorters create their own categories, we can help ensure that the categories are exhaustive—an especially important validity concern considering the variability of concepts produced in open-ended survey responses. Nonexhaustive categorical coding schemes pose a common threat to validity in code-based content analysis (Seidel & Kelle, 1995). A more word-based analysis of encoding the co-occurrence of concepts would be below the level of interest in this study because the context of the concepts (in relation to each other and to each cluster) is more important than just co-occurring.

Construct Validity. Krippendorff (1980) has identified two types of construct validity in content analysis. First, *semantical validity* is the "degree to which a method is sensitive to the symbolic meanings that are relevant within a given context" (p. 157). It has an internal component in terms of how the units of analysis are broken down as well as an external component in terms of how well coders understand the symbolic meaning of respondents' language.

Internal semantical validity refers to the process of unitizing—the reduction of the original text to individual phrases. Open-ended survey responses usually generate list-like phrases, which lend themselves well to unitizing. A unit should consist of a sentence or phrase containing one single concept. This is the most time-consuming step of the analysis. Although is not likely that the researcher would be able to bias decisions about how to break up responses based on the result he or she hoped to obtain from the analysis, it is our experience that involving two or three of the original respondents in this process is useful. Most of the units are created by separating complex sentences into simple sentences (see the example above). The important issue in unitizing is to absolutely retain the original language and meaning of the original statement. Discussion with collaborators and/or original survey respondents can be used to reduce any uncertainty about a decision. Although this step in the analysis can potentially introduce threats to validity, the sparseness and list-like nature of open-ended survey responses usually does not create very much uncertainty in creating units (this would not be the case in trying to unitize denser texts).

In terms of *external* semantical validity, if unitizing is done well and enough of the context of the statements was preserved, the meaning should be clear to the sorters (this is why choosing sorters based on well-justified criteria is important). Statements that are hard to interpret can easily be identified in this analysis because they will have high bridging values and/or often appear more toward the center of the map (they are pulled in several directions by the MDS analysis). The researcher can then use this information to revisit how the units of analysis were created.

The second type of construct validity, *sampling validity*, "assesses the degree to which available data are either an unbiased sample from a universe of interest or sufficiently similar to another sample from the sample universe so that data can be taken as statistically representative of that universe" (Krippendorff, 1980, p. 157). The survey data presented in this example were taken from a census rather than a sample and therefore were representative of the population of interest (that semester's class). This is likely the case in most analyses of open-ended survey data. Random sample selection, sample size guidelines, and how to handle missing data are issues that apply to this method as much as they to do other survey methods.

A final validity consideration is the external validity of concept mapping results/classifications. By using human judgment and statistical analyses in concert, the categories in concept maps are more data-driven than they are in traditional content analysis (where they are instead typically picked by the researcher). They also do not depend on researcher judgments about which concepts to encode or include in exclusion dictionaries, as do word-unit approaches. Concept mapping is a systematic way of formalizing a choice in syntax/context relationship. That being said, the final judgment about this representation is based on human interpretation. Having the actual respondents participate in determining the final cluster solution (the cluster replay analysis) serves as a check for validity of meaning.

Limitations

The method proposed here has only addressed the analysis of a relatively simple and sparse type of qualitative data. More dense or complex textual data, such as long interview transcripts or board meeting minutes, pose a different series of methodological, reliability, and validity questions, which will be the subject of future research efforts. In this example, it was relatively straightforward to reduce three- to five-sentence responses into units of analysis containing only one concept. Each sentence did not rely on the context of those preceding or following it for meaning. For example, they were not part of a complex argument or reasoning process.

The answers here were also all stated in one direction. They did not contain "if, then" statements; positive or negative qualifications; or conditional judgments. Consider the following statement: "I would hire her if she had more education and different job experience." This statement contains two different concepts (more education and different experience) that qualify a decision (to hire or not). This poses problems for unit-of-analysis reduction. This also causes problems for sorters who might be faced with statements in the same set such as, "I would hire her if she had a college degree and a better GPA"; "I won't hire her"; and "She doesn't have enough experience, but I think job training will help her." In the case of this type of data, semantical or semantic network text analysis would probably be more effective.

Another limitation of this methodology is that of resource restriction and/or sorter burden. This data set contained 156 statements, which can be considered as being on the upper end of what can be reasonably be processed by sorters. More than 200 statements tend to overwhelm sorters and greatly reduce their willingness to remain engaged and finish the task. For the statement set presented in this example, it took each sorter about 30 to 40 minutes to finish the sort and give each of their piles a label. This is somewhat quick compared to traditional content analysis coding, but the concept mapping method requires at least 10 to 12 carefully selected sorters to produce a reliable map. It is always possible to reduce a large list of statements by combining or eliminating redundant or near-redundant ones, and it is always possible to achieve a lower number of discrete statements by broadening the criteria of what constitutes redundancy. It should be noted that in this example, repeat units of analysis were allowed. For example, if 15 people said, “I don’t know,” then 15 “I don’t know” statements were sorted. Ordinarily, in a larger data set, 15 repetitions of “I don’t know” are not necessary. If the decision to eliminate redundant statements is made, caution must be used in drawing inference from the results (e.g., if redundant statements are eliminated, no inference about importance or frequency can be made).

Future Directions

There are several directions in which this methodology can be developed and extended in future studies to build theory. The identification of regions on the maps can lead to theorizing about scale subdimensions or uncover theoretical areas that need more investigation. Concept mapping can also be used to generate items and identify dimensions in the process of scale development (e.g., see Jackson, Mannix, Peterson, & Trochim, 2002) and can also be used to develop coding schemes and content for interview and/or follow-up interview questions (Penney & Trochim, 2000). Another interesting application of this methodology is to compare how different groups of people (e.g., from different disciplines, industries, or levels of organizational hierarchy) might generate different concept mapping solutions depending on their experiences and understanding of the same phenomena. This can potentially guide researchers in identifying boundary conditions for theory.

An extension of the core concept mapping results may also be of use in organizational research. For example, once the final map has been produced, comparisons among different stakeholders groups can be made by gathering Likert-type scale ratings of each statement on any dimension (e.g., importance or relevance to job or group) and according to any demographic characteristic of interest (e.g., management vs. line workers, engineers vs. marketing, or new employee vs. employee with long tenure). In this way, there is a map that represents the entire population of interest that also allows differences among participants to be identified within the clusters. Intergroup agreement or differences can be statistically assessed. This can be dummy coded and used in a regression to predict performance.

Conclusions

Concept mapping is one alternative for the analysis of open-ended survey responses. It has several notable strengths or advantages over alternative approaches. Unlike word-analysis approaches, it does not rely on precoded, computer-recognized semantic relationships or frequency counts, therefore retaining the context of the original concept. Unlike code-analysis approaches, it does not use forced category classifications that are laden with researcher bias. Instead, it enables estimation of the similarity between concepts and clusters of concept categories that are representative of a combination of human judgment/respondent experience and statistical analysis. Concept mapping cuts analysis time down significantly (the above analysis was completed in 3 hours, including sort-material preparation and data entry) while at the same time offering improvements to some of the reliability and validity challenges of word-based and code-based analysis methods.

Concept mapping of open-ended survey questions appears to be especially well suited for the following types of organizational research questions: (a) when the researcher does not want to impose bias or suggest relationships by forcing the data into a preconceived coding scheme, (b) when existing coding schemes or theoretical frameworks do not already exist or when the purpose of the research is to explore possibilities for conceptual categories, and (c) when there are competing theoretical explanations or frameworks. The degree to which these three objectives can be accomplished, of course, depends on decisions made at each stage of the analysis. Concept mapping is a promising alternative for the analysis of open-ended survey questions in organizational research and for building stronger theory from their results. By involving human judgment at various steps of a statistical mapping analysis, it combines the best of interpretive and representative techniques. It appears to be a rather promising addition to the analysis techniques currently available.

Notes

1. This process can also be done on the computer or over the Web.

2. One approach in studies using multidimensional scaling (MDS) analysis is to run analyses for multidimensional solutions and then pick the dimension that accounts for the greatest amount of variance as a goodness-of-fit test (e.g., see Pinkley, 1990). The concept mapping method does not do this for two reasons. First, two dimensions are easier to interpret and understand in the final maps. Second, as Kruskal and Wish (1978) point out, “when an MDS configuration is desired primarily as the foundation on which to display clustering results, then a two-dimensional configuration is far more useful than one involving three or more dimensions” (p. 58).

3. The software we used generates these decision tools. We would like to point out that the concept mapping analysis can be conducted using most commercial statistical packages. However, some of the output that is generated by the software we used would require more postprocessing.

4. The bridging value, ranging from 0 to 1, tells how often a statement was sorted with others that are close to it on the map or whether it was sorted with items that are farther away on the map (Concept-Systems, 1999). Lower bridging values indicate a “tighter” relationship with other statements in the cluster. This information can be used as a “backup” to human judgment about the appropriateness of a cluster solution.

5. Depending on the level of detail desired, this range may increase or decrease. A range of 8 to 20 is recommended for most data sets of this size.

6. To ensure maximum validity of how the structure is represented by thematic clusters, it is recommended that a group of original respondents make the final cluster solution decisions. In this example, the researchers made the final cluster solution decisions because the purpose of the analysis was merely to explore what kinds of norms the students would mention. There was no intention of making predictions; generalizing the results; or drawing conclusions about the agreement, usefulness, similarities, or most frequent type of norm. We felt that imposing our theoretical understanding of group norms, drawing from a vast literature on group norms, was acceptable in this case. However, in additional projects in which we have used this methodology (e.g., Jackson, Mannix, Peterson, & Trochim, 2002), when the intention of inference was high and the respondents’ reality was totally unknown to us, we used the original respondents in every step of the analysis (sorting, cluster replay, and labeling). If proxy sorters are chosen to make the final cluster solution, the same guidelines for choosing them should apply and be justified.

References

Afifi, A., & Clark, V. (1996). *Computer-aided multivariate analysis* (3rd ed.). Boca Raton, LA: Chapman & Hall/CRC.

Axelrod, R. (Ed.). (1976). *Structure of decision: The cognitive maps of political elites*. Princeton, NJ: Princeton University Press.

Boster, J. (1994, June). The successive pile sort. *Cultural Anthropology Methods*, pp. 11-12.

Carley, K. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. In P. Marsden (Ed.), *Sociological methodology* (Vol. 23, pp. 75-126). Washington, DC: Blackwell for the American Sociological Association.

Carley, K. (1997). Network text analysis: The network position of concepts. In C. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inference from texts and transcripts* (pp. 79-100). Mahway, NJ: Lawrence Erlbaum.

Carley, K., & Kaufer, D. (1993). Semantic connectivity: An approach for analyzing symbols in semantic networks. *Communication Theory*, *3*, 183-213.

Carley, K., & Palmquist, M. (1992). Extracting, representing, and analyzing mental models. *Social Forces*, *70*, 601-636.

Concept-Systems. (1999). *The Concept System facilitator training manual*. Ithaca, NY: Concept Systems Inc. Available from http://www.conceptsystems.com

Denzin, N., & Lincoln, Y. (Eds.). (2000). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Doerfel, M., & Barnett, G. (1999). A semantic network analysis of the international communication association. *Human Communication Research*, *25*, 589-603.

Erickson, P. I., & Kaplan, C. P. (2000). Maximizing qualitative responses about smoking in structured interviews. *Qualitative Health Research*, *10*, 829-840.

Findahl, O., & Hoijer, B. (1981). Media content and human comprehension. In K. Rosengren (Ed.), *Advances in content analysis* (pp. 111-132). Beverly Hills, CA: Sage.

Fine, G., & Elsbach, K. (2000). Ethnography and experiment in social psychological theory building: Tactics for integrating qualitative field data with quantitative lab data. *Journal of Experimental Social Psychology*, *36*, 51-76.

Geer, J. G. (1991). Do open-ended questions measure “salient” issues? *Public Opinion Quarterly*, *55*, 360-370.

Gerbner, G., Holsti, O., Krippendorff, K., Paisley, W., & Stone, P. (Eds.). (1969). *The analysis of communication content: Development in scientific theories and computer techniques*. New York: John Wiley.

Guzzo, R., & Shea, G. (1992). Group performance and intergroup relations in organizations. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 3, pp. 269-313). Palo Alto, CA: Consulting Psychologists Press.

Jackson, K., Mannix, E., Peterson, R., & Trochim, W. (2002, June 15-18). *A multi-faceted approach to process conflict*. Paper presented at the International Association for Conflict Management, Salt Lake City, UT.

Jehn, K. (1995). A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative Science Quarterly*, *40*, 256-282.

Kelle, U., & Laurie, H. (1998). Computer use in qualitative research and issues of validity. In U. Kelle (Ed.), *Computer-aided qualitative data analysis: Theory, methods, and practice* (pp. 19-28). Thousand Oaks, CA: Sage.

Kraut, A. (Ed.). (1996). *Organizational surveys: Tools for assessment and change*. San Francisco: Jossey-Bass.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology* (Vol. 5). Newbury Park, CA: Sage.

Kruskal, J., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Lindkvist, K. (1981). Approaches to textual analysis. In K. Rosengren (Ed.), *Advances in content analysis* (pp. 23-41). Beverly Hills, CA: Sage.

Miles, M., & Huberman, M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Mohammed, S., Klimoski, R., & Rentsch, J. (2000). The measurement of team mental models: We have no shared schema. *Organizational Research Methods*, *3*, 123-165.

Novak, J. (1998). *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Mahwah, NJ: Lawrence Erlbaum.

Novak, J., & Gowin, D. B. (1997). *Learning how to learn*. New York: Cambridge University Press.

Patton, M. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Penney, N., & Trochim, W. (2000, November 1-5). *Concept mapping data as a guide for developing qualitative interview questions*. Paper presented at the American Evaluation Association: Increasing Evaluation Capacity, Honolulu, HI.

Pinkley, R. (1990). Dimensions of conflict frame: Disputant interpretations of conflict. *Journal of Applied Psychology*, *75*, 117-126.

Pool, I. d. S. (Ed.). (1959). *Trends in content analysis*. Urbana-Champagne: University of Illinois Press.

Pothas, A.-M., Andries, D., & DeWet, J. (2001). Customer satisfaction: Keeping tabs on the issues that matter. *Total Quality Management*, *12*, 83-94.

Rea, L., & Parker, R. (1997). *Designing and conducting survey research: A comprehensive guide*. San Francisco: Jossey-Bass.

Roberts, C. (1997). A theoretical map for selecting among text analysis methods. In C. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 275-283). Mahwah, NJ: Lawrence Erlbaum.

Ryan, G., & Bernard, R. (2000). Data management and analysis methods. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 769-802). Thousand Oaks, CA: Sage.

Seidel, J., & Kelle, U. (1995). Different functions of coding in the analysis of textual data. In U. Kelle (Ed.), *Computer-aided qualitative data analysis: Theory, methods, and practice* (pp. 52-61). Thousand Oaks, CA: Sage.

Shapiro, G. (1997). The future of coders: Human judgments in a world of sophisticated software. In C. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 225-238). Mahwah, NJ: Lawrence Erlbaum.

Sproull, N. (1988). *Handbook of research methods: A guide for practitioners and students in the social sciences* (2nd ed.). Lanham, MD: Scarecrow Press.

Stone, P. (1997). Thematic text analysis: New agendas for analyzing text content. In C. Robert (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 35-54). Mahwah, NJ: Lawrence Erlbaum.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Trochim, W. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning*, *12*, 1-16.

Trochim, W. (1993, November 6-11). *Reliability of concept mapping*. Paper presented at the American Evaluation Association, Dallas, TX.

Weber, R. (1990). *Basic content analysis* (2nd ed.). Newbury Park, CA: Sage.

Weller, S., & Romney, A. K. (1988). *Systematic data collection* (Vol. 10). Newbury Park, CA: Sage.

Young, M. (1996). Cognitive mapping meets semantic networks. *Journal of Conflict Resolution*, *40*, 395-414.

Kristin M. Jackson is a Ph.D. candidate at the Johnson Graduate School of Management at Cornell University. Her research interests include social research methods, leadership in groups and teams, conflict and work strategies in autonomous groups, and group decision making.

William M. K. Trochim is a professor of policy analysis and management in human ecology at Cornell University. His research interests include social research methods, group decision support systems, concept mapping, pattern matching, and decision analysis.