Evaluation and Program Planning 60 (2017) 277–283

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Use of multiple cluster analysis methods to explore the validity of a community outcomes concept map

Rebecca Orsi[a,b,*]

[a] School of Education, Colorado State University, United States
[b] Social Work Research Center, School of Social Work, Colorado State University, United States

ARTICLE INFO

Article history:
Received 26 July 2016
Accepted 22 August 2016
Available online 28 August 2016

Keywords:
Concept mapping
Multidimensional scaling
Cluster analysis
R statistical software
Validity

ABSTRACT

Concept mapping is now a commonly-used technique for articulating and evaluating programmatic outcomes. However, research regarding validity of knowledge and outcomes produced with concept mapping is sparse. The current study describes quantitative validity analyses using a concept mapping dataset. We sought to increase the validity of concept mapping evaluation results by running multiple cluster analysis methods and then using several metrics to choose from among solutions. We present four different clustering methods based on analyses using the R statistical software package: partitioning around medoids (PAM), fuzzy analysis (FANNY), agglomerative nesting (AGNES) and divisive analysis (DIANA). We then used the Dunn and Davies-Bouldin indices to assist in choosing a valid cluster solution for a concept mapping outcomes evaluation. We conclude that the validity of the outcomes map is high, based on the analyses described. Finally, we discuss areas for further concept mapping methods research.

1. Introduction

Literature in the field of theory-based evaluation (Chen, 1990) has suggested that statistical tools may be used in the development of program theory, particularly in the area of outcomes evaluation. Also, such literature suggests (Leeuw, 2003) that three primary methods for reconstructing program theories can be used: a policy-scientific approach, a strategic assessment approach and an elicitation approach. The elicitation approach draws on ideas from cognitive and organizational psychology, and Leeuw notes that Trochim's concept mapping method (1989) is an example of the elicitation approach. Programmatic outcomes can be understood as one domain in the context of Chen's (1990) six-domain framework for program theory. Alternatively, outcomes can be regarded as part of a simpler program theory framework based on a linear logic model to represent inputs, outputs and outcomes (University of Wisconsin-Extension, 2002; W.K. Kellog Foundation, 2001). The purpose of this study is to explore the utility of concept mapping (Trochim, 1989) as a tool for articulating outcomes from a complex social intervention. Specifically, we examine variations of the basic concept mapping process and how such variations can assist evaluators in validly articulating programmatic outcomes.

Concept mapping was first presented as a cohesive evaluation tool more than 25 years ago (Trochim, 1989). Briefly, the six steps for the process are: (1) *preparation*, which includes selection of participants and a brainstorming focus statement, (b) *generation* of focus response statements via brainstorming, (c) *structuring* of statements via sorting and rating, (d) *concept mapping analysis* (also known as *representation* of statements on a map), (e) *interpretation* of maps and (f) *utilization* of maps (Kane & Trochim, 2007). The current study constitutes an in-depth look at step four: concept mapping analysis. It will explore several alternative cluster analyses in an effort to produce the most valid representation of participant responses as possible. The discussion is primarily methodological with a focus on how valid results were obtained. Readers interested in substantive concept mapping outcomes as they relate to the social program in question are referred to Orsi (2014).

Shadish, Cook, and Campbell (2002) note that validity refers to whether or not an inference or knowledge claim or proposition is approximately true. They further note that validity is properly understood as a property of inferences, not a property of methods. Thus, when considering validity and concept mapping, the question is not whether concept mapping is a valid method, but whether concept mapping produces valid knowledge propositions for a specified situation or context. As noted briefly above, the fourth step of concept mapping is the *representation of statements* on a map. This involves conducting one or more cluster analyses.

* Correspondence to: School of Social Work, Campus Delivery 1586, Colorado State University, Fort Collins, CO 80523-1586, United States.
E-mail address: becky.orsi@colostate.edu (R. Orsi).

http://dx.doi.org/10.1016/j.evalprogplan.2016.08.017
0149-7189/

Ward's hierarchical method of clustering is commonly used, although other types of cluster analysis are available. Furthermore, cluster analysis typically involves judgement, as more than one set of clusters may represent concept mapping data in a meaningful way. Decisions about what type of cluster analysis to use and how to select a final set of clusters have implications for the validity of the evaluation results produced. Therefore, in the current study, we ask whether or not using several clustering methods (e.g. agglomerative, divisive or non-hierarchical) and calculating clustering fit metrics serves the purpose of better enabling valid articulation of programmatic outcomes via concept mapping.

A few past studies have examined the validity of concept mapping results. Dumont (1989) considered whether maps formed by multidimensional scaling (MDS) are a valid representation of a participants' individual conceptualizations for a construct of interest. This study concluded that the maps were not entirely valid representations of participants' experiences. Dumont did not, however, entirely use Trochim's methodology. Cacy (1996) produced three concept maps relating to the nature of doctors' practice-based research networks. Participants were asked to choose a concept map that "makes the most sense" to them, based on professional experience (1996, p. 95). Results showed that faculty doctors consistently chose a community practitioners' concept map, rather their own map. Community practitioners did not consistently pick any of three possible maps. Cacy concluded that the study did not provide any certain evidence for the validity of concept maps (1996). More recently, Rosas and Kane (2012) examined possible measures of representational validity from a wide variety of concept mapping studies, focusing primarily on internal representational validity. Measures discussed include acceptable values for the stress statistic and also configural similarity. These measures focus on validity understood as "determining the overall match between the participant-structured input and the mathematically generated output" (Rosas & Kane, 2012, p. 237). Internal representational validity was found to be good across the studies examined.

Although the focus of the current study is on validity from a quantitative and statistical point of view, concept mapping also includes qualitative data. Therefore, alternative considerations of validity are also appropriate. For example, Jackson and Trochim (2002) take a perspective from content analysis and suggest that collective conceptualizations from concept mapping are potentially more valid than are results from other methods which rely substantially on the researcher's role or interpretations. However, with Krippendorf, they note that because concept mapping deals with social constructions, " . . . there is really no way to establish a standard by which to judge the degree of error" in the expression of participants' perceptions (2002, p. 330; Krippendorf, 1980). Also, Orsi (2014) discusses the *credibility* of outcomes using data from the current study and finds that outcomes are indeed credible to program participants who review concept maps. It remains the case, however, that no single and universally accepted measure of validity for concept mapping yet exists (Bedi & Alexander, 2009; Trochim, Cook, & Setze, 1994).

2. Methods

The context for applying concept mapping in the current study was a grassroots community organization in a western United States city. At the time, the organization was working to address several community issues including access to healthcare, local and state education reforms, college access, citizenship, and neighborhood safety. To create an outcomes map for the community organizing program, twenty-one grassroots community leaders from schools and local religious congregations participated. The twenty-one participants generated responses to the following focus statement: "Think about yourself, your family, your child's school, your church and your neighborhood. When [our organization] does community organizing, this is what happens: __________." This resulted in 125 statements for sorting and rating. However, Kane and Trochim (2007) suggested limiting the number of statements for sorting to about 100. Experience from the author's pilot study suggested there should be even fewer statements to reduce the time for sorting, a task which pilot participants stated was burdensome. Therefore, statements reflecting a similar theme or topic were combined to remove redundancy and to reduce the number of sort statements. The final list of brainstormed statements numbered 89. This is consistent with the average number of statements per study (about 96) found by a recent overview of concept mapping studies (Rosas & Kane, 2012).

The next step of the concept mapping process was the sorting of statements. In total, twenty-one sorted solutions for the 89 statements were provided by participants. These formed the data set for multidimensional scaling and cluster analyses. Table 1 displays information which characterizes the entire group of participants in terms of experience with community organizing, personal education level, age, and childcare responsibilities. Detailed results concerning community organizing outcomes are reported elsewhere (Orsi, 2014). In the current study, we focus on methods to ensure validity and on providing details from statistical analysis using the R statistical software environment (R Development Core Team, 2011).

2.1. Data preparation

As noted by McLinden (2013), in order to perform multidimensional scaling (Bartholomew, Steele, Moustaki, & Galbraith, 2002; Kruskal & Wish, 1978) and cluster analysis (Johnson & Wichern, 2007; Kaufman & Rousseeuw, 1990), it is necessary to reformat data from the sort solutions into a matrix for analysis in the R statistical package. Each response statement was numbered from 1 to 89. Each participant's sort results were transcribed into a Word document, then cut, pasted and edited in the R package as a vector and finally, transformed from a vector into a symmetric binary

Table 1
Participant Characteristics (n = 21).

Characteristic	
Experience with organizing	
1 year or less	5%
2–3 years	33%
4 or more years	62%
Committee affiliation	
Congregation	72%
School	33%
Education level	
No HS diploma	14%
HS diploma	33%
2-year degree	5%
4-year degree +	43%
Missing	5%
Age	
Under 35 years	19%
35–50 years	33%
Over 50 years	48%
Caring for children?	
Yes	38%
No	62%

Note: Percentages of committee affiliation do not add to 100% because some participants are affiliated with both a congregation and school organizing committee.

similarity matrix. There was one binary similarity matrix for each of the 21 participants. Each binary similarity matrix was square and symmetric (dimension 89×89). The R code building the similarity matrices contained two checks for accuracy. The first check counted the number of statements transcribed into each participant's groupings and checked to see it was equal to 89. The second check counted all possible pairs in each sort group and compared it to the sum of entries in the binary similarity matrix. Each participant's similarity matrix was reviewed for accuracy using these checks.

After a binary similarity matrix was created for each participant, these were summed to form a group similarity matrix (Bedi, 2006). Values of the cells in the group similarity matrix represent how many times each pair of statements was sorted together by participants. The values in the group similarity matrix can range from 0 (a pair of statements never sorted together) to 21 (a pair of statements sorted together by every participant). As Bedi notes, "the value in . . . [the group similarity matrix] for any pair of statements indicates how many participants placed that pair of statements together in a pile regardless of what other statements were included or excluded from that pile" (2006, p. 28). Finally, the group similarity matrix was transformed into a *dissimilarity* matrix by subtracting similarity values from a constant. Values in the group dissimilarity matrix now ranged between 0 and 21, but now a high value of 21 indicated that a pair of statements was *not* often sorted together.

2.2. Multidimensional scaling and cluster analysis

The dissimilarity matrix forms the basis for non-metric multidimensional scaling (MDS) analysis (Bedi, 2006; Carter, Enyedy, Goodyear, Arcinue, & Puri, 2009; Trochim, 1989). Details on this analysis are contained in Kruskal and Wish (1978). Essentially, however, MDS takes the information in the dissimilarity matrix and expresses it as a distance between each pair of points. A high dissimilarity value corresponds to a large distance between points. These distances are used to build a map showing the relationship of each of the 89 statements to every other statement. An MDS analysis can provide a map solution in any number of dimensions specified by the researcher. Usually, a two-dimensional solution is used with concept mapping due to its ease of representation (Trochim, 1989). It is often difficult to visualize a solution in three dimensions and nearly impossible in four. This study followed the convention of using a two-dimensional solution.

After the 89 response statements were put on a two-dimensional map, a second statistical technique – cluster analysis – was applied to mathematically group the mapped response statements into non-overlapping clusters of conceptually similar statements (Trochim, 1989). In this case, a matrix of the pairwise distances between these points was used as the input for cluster analysis. The R statistical package allows points which represent sorted statements to be specified either by a set of coordinates or by a set of pairwise distances among points; either of these two sources of data can serve as input for the cluster analysis and will produce geometrically equivalent results.

Johnson and Wichern (2007) suggest trying multiple clustering methods; if outcomes from several methods are consistent, a researcher may be able to support a case for natural groupings. Kaufman and Rousseeuw (1990) also suggest using multiple clustering methods. For a concept mapping application it is critical to choose from among clustering methods which partition the MDS space, rather than methods which would allow any pair of points to be clustered together, regardless of location. For this study, four types of clustering discussed by Kaufman and Rousseeuw (1990) were used in order to develop a valid conceptual representation of outcome measures. These clustering methods include: partitioning around medoids (PAM), fuzzy analysis (FANNY), agglomerative nesting (AGNES) and divisive analysis (DIANA). PAM is a non-hierarchical method. PAM chooses k representative statements called medoids and builds clusters by assigning each statement to the nearest medoid. FANNY is also a non-hierarchical method. It does not definitively assign each statement to one cluster, but rather assigns each statement to every cluster with a percentage likelihood. So, a statement might be described as belonging with x% probability to cluster 1, y% to cluster 2, and z% to cluster 3. If x% is the largest magnitude percentage, the statement should probably be assigned to cluster 1, but there is some possibility that it belongs to one of the other clusters (Kaufman & Rousseeuw, 1990). A "hard" clustering can be obtained using FANNY by assigning each statement to the cluster with the largest likelihood. AGNES is an agglomerative hierarchical clustering method which begins with every statement in its own group and joins statements together. Once a pair is joined by AGNES, it cannot be split again. Clusters of any size can be formed, and several methods are available in AGNES. In the current study, the Ward's method was chosen to be consistent with traditional concept mapping methodology. Finally, DIANA is a divisive hierarchical method. However, DIANA works in the opposite direction from AGNES. It starts with all statements in one group and then breaks the large group into any number of smaller clusters. All of the cluster analyses for this study have been conducted using the 'cluster' package (Maechler, 2011) in the R statistical software environment (R Development Core Team, 2011). The 'cluster' package has fully implemented the four types of clustering just described, using the algorithms of Kaufman and Rousseeuw (1990).

2.3. Application of four cluster analysis methods

Use of the four clustering methods described above constitutes an expansion to the *representation of statements* (Trochim, 1989) step of the concept mapping process. (Traditionally, only one clustering method has been used.) In the current study, *representation of statements* first yielded an MDS point map, with one point for each response statement. The statements on the point map were then partitioned using all four methods. Each method was used to create four solutions of 4, 5, 6 and 7 groupings for a total of sixteen cluster maps (four for each method). Then, by process of elimination, we sought to select the most valid cluster map from the set of sixteen. We sequentially removed cluster solutions from consideration to arrive the cluster solution which was the most valid representation of the participant-sorted statements. As part of this process, we used several measures of clustering structure (provided in the R package) including the agglomerative and divisive coefficients and a normalized version Dunn's index (Dunn, 1974, 1976). We also evaluated additional considerations regarding the quality of clustering solutions (as described below) to make final selections.

3. Results

3.1. Multidimensional scaling

The goodness-of-fit measure for an MDS solution is called the *f-stress* statistic, or simply *stress* (Kruskal & Wish, 1978; Ripley, 2011). The stress value of the two-dimensional solution for these program data was 0.24. This is higher than the 0.10 guideline suggested by Kruskal and Wish (1978), but it is consistent with average values of the stress metric for several recent studies utilizing concept mapping (Rosas & Kane, 2012).

3.2. Cluster analysis

The AGNES clustering results were evaluated by an agglomerative coefficient which ranges between 0 and 1; higher values indicate relatively more clustering structure is present in the data, compared to lower values (Kaufman & Rousseeuw, 1990). For the AGNES solutions, the agglomerative coefficient was very high, displaying a value of 0.98. The DIANA algorithm produces a similarly-purposed divisive coefficient with a value of 0.93. Kaufmann and Rousseeuw have noted, however, that both the agglomerative and divisive coefficients can be influenced by even a single outlier. For the community organizing data, an outlier is a statement not sorted consistently with any other group of statements (e.g. statement 89, see Fig. 1, right side). Therefore, the high values of these coefficients for the community organizing data may be more attributable to outliers rather than meaningful clusters in the dataset.

The FANNY algorithm calculates a normalized version of Dunn's index (Dunn, 1974, 1976) to assess the clarity of cluster structures produced with this algorithm (Kaufman & Rousseeuw, 1990). The normalized measure ranges from 0 to 1 with 1 indicating a completely well-partitioned (e.g. non-fuzzy) cluster solution. The normalized partition coefficients for the FANNY cluster solutions ranged from 0.133 to 0.214, indicating a set of rather poor cluster solutions. Poor partitioning measures were also validated by a visual inspection of the FANNY cluster maps (not shown). With the exception of one well-defined cluster at the left of the map, the rest of the clusters are not well-differentiated. The PAM algorithm does not offer a numeric measure of the goodness of the clustering solution.

We considered additional aspects the fit of clustering solutions to the MDS map as we sought the most valid representation of outcomes. One of the four clustering methods (FANNY) did not produce a 7-group solution; its "7-group" solution contained only six clusters which were virtually identical to its 6-cluster solution. Also, two of the 7-group solutions (DIANA and AGNES) had one cluster which was very small, with either two or four statements, respectively. In the context of community organizing outcomes, it seemed unlikely that groups of two or four statements would be a valid representation of an entire group of community outcomes. Based on these considerations, all of the 7-group solutions were eliminated. Next, the remaining DIANA solutions (4–6 clusters) were eliminated. We made this decision because the 4-group solution appeared to poorly differentiate a visually obvious cluster on the left side of the map and because the 5- and 6-group DIANA solutions also displayed a cluster with only two members. These decisions narrowed the set of possible maps from sixteen possibilities to nine.

The remaining nine cluster solutions consisted of the 4, 5 and 6-cluster solutions for AGNES, FANNY and PAM. These displayed a great deal of similarity. Each one presented cluster solutions arranged in a more-or-less oval-shaped pattern around a relatively empty area slightly to the upper left of the center of the map. To compare the nine remaining maps and in order to choose the best map from a statistical perspective, we calculated two cluster validation indices: the Dunn index (using a non-normalized version) and the Davies-Bouldin index (Davies & Bouldin, 1979; Dunn, 1974; Halkidi, Batistakis, & Vazirgiannis, n.d.). Both assess the separation of clusters, the former based on resemblance and the latter based on difference. These indices can be calculated in the R statistical environment using a number of different intracluster diameter and intercluster distance measures. The point map contains both outlier and overlapping points; therefore, average intracluster diameter and intercluster distance measures were used to calculate the indices in order to limit the effect of outliers and overlaps. Average intracluster diameter is defined as the average of all distances between point pairs in a cluster (Nieweglowski, 2009). Other distance metric options are complete linkage (for both diameter and distance) and single linkage (for intercluster distance). These measurements, however, are based on

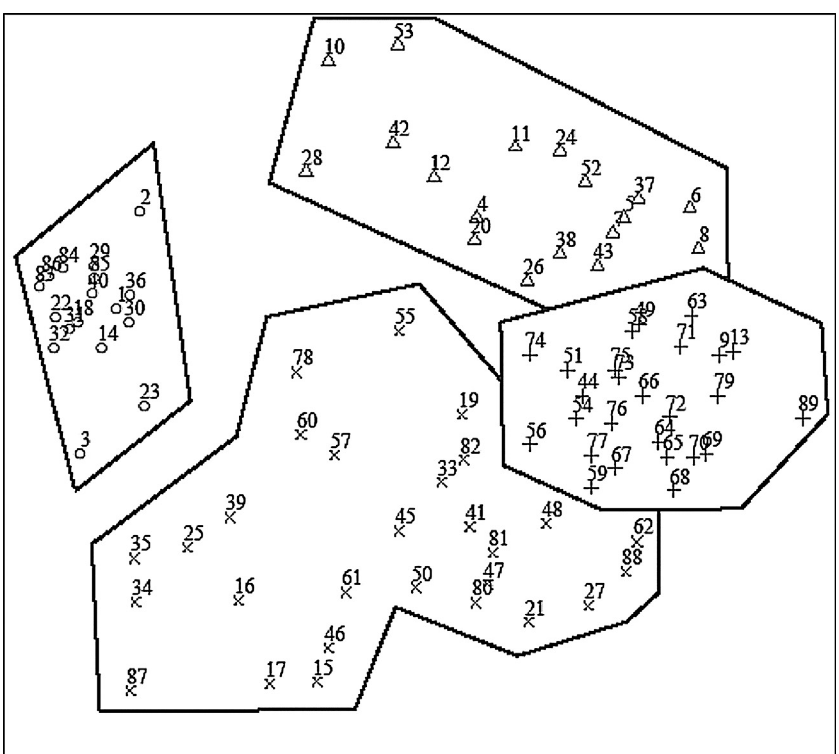

Fig. 1. AGNES four cluster map for community organizing outcomes.

the furthest apart and closest together point pairs, respectively and would be unduly influenced by outlier or overlapping points. Average linkage distances are more representative of the entire cluster.

Dunn's index is a measure of difference between clusters. Thus, to identify a statistically accurate solution, we sought to find the cluster solution with the highest value of Dunn's index. For the nine solutions, the values of Dunn's index ranged from 1.54 to 3.04. The two solutions with the highest indices are the PAM 4-cluster solution (3.04) and the AGNES 5-cluster solution (3.00). The Davies-Bouldin index is a measure of resemblance between clusters. Thus, we sought to find the cluster solution with the lowest value for the Davies-Bouldin index. For the nine solutions, the values of Davies-Bouldin ranged from 0.47 to 0.56. The solution with the lowest index was the AGNES 4-cluster solution. Based on these metrics, we chose to present two maps to community leaders for interpretation: the AGNES 4- and 5-cluster solutions, one with a high measure for Dunn's index and one with a low measure of the Davies-Bouldin index. In addition to statistical evidence supporting this choice, there were two other reasons to identify these solutions as most valid for the interpretation step. First, three of the non-overlapping clusters were identical between the AGNES 4- and 5-group solutions. The 5-cluster solution broke the least cohesive cluster from the 4-cluster solution in two. Presenting both solutions allowed participants to name the groups either individually or as one. The second reason for our choice was that the AGNES algorithm appeared to do a better job than the PAM algorithm of differentiating the most visually obvious cluster at the left side of the map. Both the AGNES 4 and 5-group solutions are displayed as Figs. 1 and 2.

Following the selection of maps, an interpretation session was held with members of the community organization. Participants were able to characterize the three clusters that were the same on both maps as: *victories* (circular plot characters), *personal development* (triangular plot characters) and *public leadership skills* (cross-shaped plot characters). These represented, respectively, public organizing outcomes, newly-acquired personal skills and feelings of empowerment, and political and democratic process skills. For the 4-cluster AGNES map, the fourth cluster (Fig. 1; x-shaped plot characters) was difficult for participants to name. The AGNES 5-cluster solution (which split the fourth cluster in two) resulted in names for both new clusters. One was named *relationships with power people* (Fig. 2; x-shaped plot characters). The other was named *culture of civic engagement* (Fig. 2; diamond plot characters). For a complete discussion of the interpretations session and its results see Orsi (2014).

4. Discussion

4.1. Validity evidence from multiple clustering methods

As described above, there were originally sixteen different cluster maps created. Both DIANA and 7-group cluster solutions were eliminated, leaving a group of nine potential solutions. All nine viable solutions appeared visually quite similar. Similar solutions from different clustering methods lend support to the notion of natural clusters in the data (Johnson & Wichern, 2007). To further support conclusions based on visual inspection, we ran similarity measures for pairs of cluster solutions. We used a similarity index based on a measure called partition-distance (Almudevar & Field, 1999; Gusfield, 2002). A partition for a set of statements is simply the division of that set into smaller, non-overlapping groups (Ross, 1988). For any two different partitions, the partition distance is the minimum number of statements which must be deleted so that the two partitions (when restricted to the remaining statements) become equal (Giurcaneanu, Tabus, Shmulevich, & Zhang, 2003). The similarity index ranges from 0 to 1, and a value of 1 indicates identical partitions (Nieweglowski, 2009). The original context for use of a similarity index by Giurcaneanu et al. (2003) was a comparison of multiple cluster

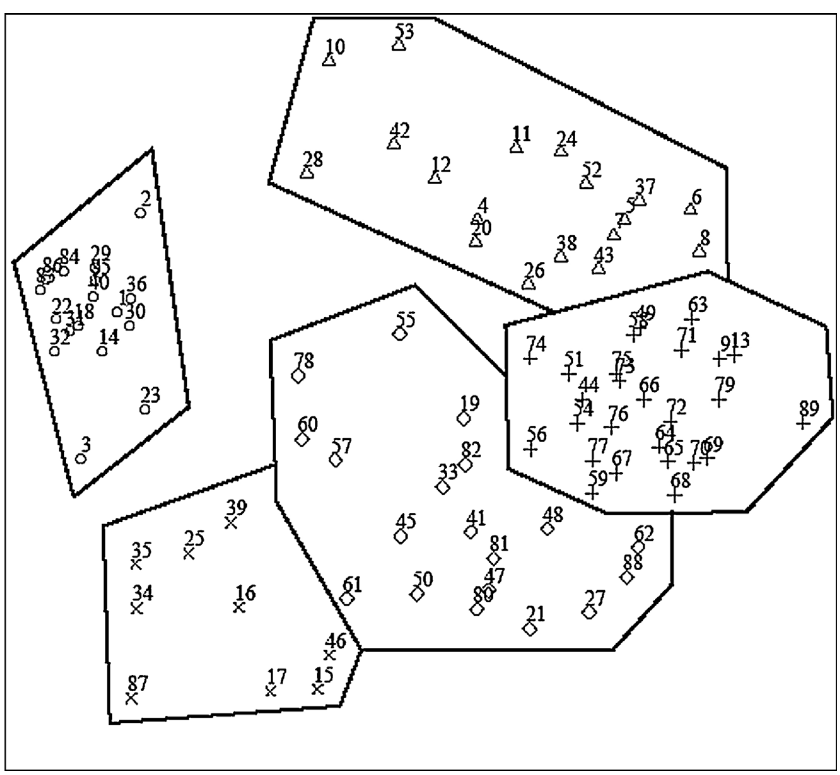

Fig. 2. AGNES five cluster map for community organizing outcomes.

solutions (with different numbers of clusters) to a “true” partition for a microarray dataset. In the current study, we use the index to compare eighteen pairs chosen from the nine cluster maps and to assess how similar are the pairs. As can be seen in Table 2, all of the 18 comparisons show a great deal of similarity between cluster solutions; none of the indices were lower than 0.73 on a scale of 0–1. The use of multiple clustering solutions has yielded nine cluster maps which are both visually similar and quantitatively similar. The similarity of the cluster solutions lends clear support for the validity of the final cluster map. We can be confident that results are not an artifact of the clustering solution used. It is reasonable to expect that if some of the other nine cluster maps (i.e. not AGNES 4 and AGNES 5) had been presented at the interpretation session, similar group names would have resulted. Finally, it is worthwhile noting that the chosen solutions result from a hierarchical clustering solution. This is consistent with Kane’s and Trochim’s conclusion that a hierarchical method is a better fit for concept mapping than a centroid-based method (Kane & Trochim, 2007) and lends further support for the validity of the final map.

4.2. Limitations

This study is limited in several ways. First, the study would have benefited from more participants sorting statements. Prior research points to more sorters being positively correlated with reliability of sort solutions for concept mapping studies (Rosas & Kane, 2012). In this study, more sort solutions might also have helped to better characterize the statements in the last cluster (*civic engagement*). Its placement suggests that the statements in this cluster have not been consistently sorted together with any other group of statements. With a relatively small number of participants, we do not know whether this cluster is genuinely ambiguous or whether it is simply an artifact of small sample size. Another study limitation was “participant fatigue” in relation to the concept mapping process. Three participants commented that sorting took a great deal of time. Sort time could be reduced by limiting the number of statements. It would be worth considering how a large list of statements could be reduced so that it would still be representative of programmatic outcomes but would also be less burdensome to sort. Kane and Trochim (2007) suggest borrowing methods from content analysis to assist in limiting the number of statements. Using content analysis to reduce the set size might have alleviated participant fatigue and generated tighter clusters. Finally, if there had been a larger number of participants at the interpretation session (and if participant fatigue had not been a concern) it would have been valuable to provide participants with one cluster map from the seven which were initially eliminated (i.e. a seven group solution or a DIANA solution). If the participants could have analyzed the clusters on the eliminated map and determined that they were not as reflective of the organization’s outcomes, this would have provided disconfirmatory evidence for the validity of the final map.

Table 2
Similarity Indices Comparing Different Cluster Solutions.

Comparison Pair	Similarity Index
Comparisons with 4 clusters	
AGNES 4 vs. PAM 4	0.90
AGNES 4 vs. FANNY 4	0.92
PAM 4 vs. FANNY 4	0.90
Comparisons with 5 clusters	
AGNES 5 vs. PAM 5	0.91
AGNES 5 vs. FANNY 5	0.88
PAM 5 vs. FANNY 5	0.91
Comparisons with 6 clusters	
AGNES 6 vs. PAM 6	0.90
AGNES 6 vs. FANNY 6	0.75
PAM 6 vs. FANNY 6	0.74
AGNES Comparisons	
4 vs. 5 Clusters	0.90
4 vs. 6 Clusters	0.84
5 vs. 6 Clusters	0.94
FANNY Comparisons	
4 vs. 5 Clusters	0.81
4.vs. 6 Clusters	0.73
5 vs. 6 Clusters	0.80
PAM Comparisons	
4 vs. 5 Clusters	0.78
4.vs. 6 Clusters	0.73
5 vs. 6 Clusters	0.89

4.3. Additional research

As with the exploration of multiple clustering methods in this study, there are also aspects of the MDS analysis that would benefit from methodological research. For example, it would be worthwhile to investigate the question of whether evaluations should always present MDS results in two dimensions as a basis for cluster analyses. Trochim (1989) notes that two dimensions are best used to provide a map that participants can visualize. Additional dimensions would improve fit, though with a decreasing benefit from each additional dimension because MDS selects initial dimensions to account for as much variability as possible. Nevertheless, computing capacity has increased substantially since the development of concept mapping. Additional graphical displays could be explored which would allow a three-dimensional solution to be presented to participants in a meaningful way. This could include developing a program to print a concept map using a 3D printer (C. Petrucci, personal communication, March 5, 2015) or using additional dimensions to display rating data (W. Trochim, personal communication, October 20, 2015). The community organizing data generated some evidence for a benefit from three dimensions. The stress value for a three-dimensional solution was 0.17, which is a 29% reduction from the two-dimensional solution stress value of 0.24. A comparison of the AGNES 5-cluster solution based on a two-dimensional MDS layout (used in the current study) and an additional AGNES 5-cluster solution based on a three-dimensional layout (both with non-overlapping clusters) yielded a similarity index (Nieweglowski, 2009) of 0.69 (Results not shown in Table 2). The fact that this value is less than all of the two-dimensional solutions shown in Table 2 suggests that clusters based on the three-dimensional distances differ and could have yielded different interpretation results.

Another question for research is whether MDS is suitable to portray a large number of paired statements with overlapping dissimilarity metrics. There are over 3900 possible pairs of 89 statements. For an MDS analysis, each pair requires a dissimilarity measure. Ideally, each pair would have a *distinct* dissimilarity measure, different from all other pairs in the data. However, with only 21 participants and a univariate dissimilarity measure (i.e. whether a pair was sorted in the same group or not), the dissimilarity measures for each pair are not distinct. In fact, the group dissimilarity matrix contains many zeros (never sorted together) and many ties (sorted together the same number of times). Texts which provide detail on the statistics behind multidimensional scaling do not offer any obvious guidance on the question of using MDS in a situation with many pairs and a relatively undifferentiated dissimilarity measure (Cox & Cox, 2001; Everitt & Rabe-Hesketh, 1997). Note, however, that two non-metric scaling examples given in Everitt and Rabe-Hesketh (1997, pp. 34,

43) use many fewer statement pairs (60 and 84, respectively), and they present many fewer ties in the dissimilarity measures. Further research regarding an appropriate number of pairs and appropriate dissimilarity measures could start with a literature review examining studies which used MDS and document the lowest and highest number of data points used.

Finally, the question of standardizing dissimilarity measures prior to analysis could merit further research. Some users of concept mapping suggest that the standardization of dissimilarity measures (in this case, rescaling dissimilarity measures ranging from 0 through 21 so that they range from 0.0 to 1.0) may facilitate the MDS being less sensitive to the specific computational algorithm used (H. Bar, personal communication, February 12, 2015). This issue has been raised in other fields of research where MDS is used (Austin, 2013; Kenkel & Orlóci, 1986). Evaluation methodologies could benefit from the examination of standardization for MDS in the context of the social science data.

5. Conclusion

In conclusion, this study builds additional evidence for the validity of program outcomes articulated via concept mapping. Conducting analyses in the R statistical software package (R Development Core Team, 2011) offers a means of using several methods of cluster analysis. Results reconfirm Kane and Trochim's conclusion that hierarchical methods prove most useful for concept mapping (Kane & Trochim, 2007). However, R offers the ability to try multiple methods as part of any concept mapping analysis with relative ease. Also, new validation measures are suggested by this study, including the Dunn and Davies-Bouldin indices (Davies & Bouldin, 1979; Dunn, 1974). These assist in choosing among a set of cluster solutions. Similarity indices are also provided (Giurcaneanu et al., 2003) for the purpose of comparing pairs of cluster solutions. All of these indices are provided in the 'cluster' (Maechler, 2011) package in the R statistical software environment. Further research into whether and how concept maps could be meaningfully displayed in more than two dimensions and into the issue of scaling raw data would be welcome. In short, even after more than 25 years of using concept mapping, there remain many options and refinements to be explored.

Acknowledgement

The author acknowledges R. Brian Cobb, Ph.D. for his mentorship in conducting the study.

References

Almudevar, A., & Field, C. (1999). Estimation of single-generation sibling relationships based on DNA markers. *Journal of Agricultural Biological and Environmental Statistics*, *4*(2), 136–165.

Austin, M. P. (2013). Inconsistencies between theory and methodology: A recurrent problem in ordination studies. *Journal of Vegetation Science*, *24*(2), 251–268. http://dx.doi.org/10.1111/j.1654-1103.2012.01467.x.

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2002). *The analysis and interpretation of multivariate data for social scientists*. Boca Raton, FL: Chapman & Hall/CRC.

Bedi, R. P., & Alexander, D. A. (2009). Using multivariate concept-mapping for assessing client understandings of counselling. *Canadian Journal of Counselling*, *43*(2), 76–91.

Bedi, R. P. (2006). Concept mapping the client's perspective on counseling alliance formation. *Journal of Counseling Psychology*, *53*(1), 26–35. http://dx.doi.org/10.1037/0022-0167.53.1.26.

Cacy, J. R. (1996). The reality of stakeholder groups: A study of the validity and reliability of concept maps. (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (Publication No. AAT 9608493).

Carter, J. W., Enyedy, K. C., Goodyear, R. K., Arcinue, F., & Puri, N. N. (2009). Concept mapping of the events supervisees find helpful in group supervision. *Training and Education in Professional Psychology*, *3*(1), 1–9. http://dx.doi.org/10.1037/a0013656.

Chen, H.-t (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage.

Cox, T. F., & Cox, M. A. A. (2001). *Multidimensional scaling*, 2nd ed. Chapman and Hall.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*(2), 224–227.

Dumont, J. (1989). Validity of multidimensional scaling in the context of structured conceptualization. *Evaluation and Program Planning*, *12*(1), 81–86. http://dx.doi.org/10.1016/0149-7189(89)90026-8.

Dunn, J. C. (1974). Well separated clusters and optimal fuzzy partitions. *Cybernetics and Systems: An International Journal*, *4*(1), 95–104. http://dx.doi.org/10.1080/01969727408546059.

Dunn, J. C. (1976). Indicies of partition fuzziness and the detection of clusters in large data sets. In M. Gupta (Ed.), *Fuzzy automata and decision processes*New York: Elsevier.

Everitt, B. S., & Rabe-Hesketh, S. (1997). *The analysis of proximity data*. London: Arnold.

Giurcaneanu, C. D., Tabus, I., Shmulevich, I., & Zhang, W. (2003). Stability-based cluster analysis applied to microarray data. *Paper presented at the seventh international symposium on signal processing and its applications* [Retrieved from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1224814&queryText=Giurcaneanu&newsearch=true].

Gusfield, D. (2002). Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, *82*(3), 159–164. http://dx.doi.org/10.1016/s0020-0190(01)00263-0.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (n.d.). Clustering validity checking methods: Part II. Retrieved from: http://www.db-net.aueb.gr/files/HBV_SIGREC02%28II%29.pdf.

Jackson, K. M., & Trochim, W. M. K. (2002). Concept mapping as an alternative approach for the analysis of open-ended survey responses. *Organizational Research Methods*, *5*(4), 307–336. http://dx.doi.org/10.1177/109442802237114.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis*, 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall.

Kane, M., & Trochim, W. M. K. (2007). *Concept mapping for planning and evaluation*. Thousand Oaks, CA: Sage.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley & Sons.

Kenkel, N. C., & Orlóci, L. (1986). Applying metric and nonmetric multidimensional scaling to ecological studies: Some new results. *Ecology*, *67*(4), 919–928. http://dx.doi.org/10.2307/1939814.

Krippendorf, K. (1980). *Content analysis: An introduction to its methodology*, 1st ed. Beverly Hills, CA: Sage.

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

Leeuw, F. L. (2003). Reconstructing program theories: Methods available and problems to be solved. *The American Journal of Evaluation*, *24*(1), 5–20. http://dx.doi.org/10.1016/S1098-2140(02)00271-0.

Maechler, M. (2011). Package 'cluster'. Retrieved March 23, 2011, from http://cran.r-project.org/.

McLinden, D. (2013). Concept maps as network data: Analysis of a concept map using the methods of social network analysis. *Evaluation and Program Planning*, *36*(1), 40–48. http://dx.doi.org/10.1016/j.evalprogplan.2012.05.001.

Nieweglowski, L. (2009). Package 'clv.'. Retrieved March 25, 2011, from http://cran.r-project.org/.

Orsi, R. (2014). Concept mapping: A method for articulating grassroots community organizing outcomes. *Journal of Community Psychology*, *42*(4), 398–413.

R Development Core Team (2011). *R: A language and environment for statistical computing (Version 2.12.2)*. Vienna, Austria: R Foundation for Statistical Computing Retrieved from www.r-project.org.

Ripley, B. (2011). Package 'MASS'. Retrieved March 23, 2011, from http://cran.r-project.org/ .

Rosas, S. R., & Kane, M. (2012). Quality and rigor of the concept mapping methodology: A pooled study analysis. *Evaluation and Program Planning*, *35*(2), 236–245. http://dx.doi.org/10.1016/j.evalprogplan.2011.10.003.

Ross, S. (1988). *A first course in probability*, 3rd ed. New York: Macmillan.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton-Mifflin.

Trochim, W. M. K., Cook, J. A., & Setze, R. J. (1994). Using concept mapping to develop a conceptual framework of staff's views of a supported employment program for individuals with severe mental illness. *Journal of Consulting and Clinical Psychology*, *62*(4), 766–775. http://dx.doi.org/10.1037/0022-006X.62.4.766.

Trochim, W. M. K. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning*, *12*(1), 1–16. http://dx.doi.org/10.1016/0149-7189(89)90016-5.

University of Wisconsin-Extension. (2002). Enhancing program performance with logic models. Retrieved June 7, 2011, from http://www.uwex.edu/ces/pdande/evaluation/pdf/lmcourseall.pdf.

W.K. Kellogg Foundation. (2001). Logic model development guide. Retrieved June 7, 2011, from http://www.wkkf.org/knowledge-center/resources/2006/02/WK-Kellogg-Foundation-Logic-Model-Development-Guide.aspx.