Local versus global connection making in discourse

Wesley Collier, The University of Wisconsin-Madison, collier.wesley@gmail.com
Andrew Ruis, The University of Wisconsin-Madison, andrewruis@gmail.com
David Williamson Shaffer, The University of Wisconsin-Madison, dws@education.wisc.edu

Abstract: This paper examines techniques for modeling relationships among domain concepts and practices in discourse to assess learning in a CSCL environment. We compare two approaches: a traditional psychometric approach, which models the global correlation structure of student discourse markers across the learning intervention, and a model that accounts for the local correlation structure of discourse markers within activities. We investigate whether: (a) analysis of local correlation structure can identify significant differences between novices and relative experts; (b) these differences reflect meaningful differences the discourse; and (c) analysis of global correlation structure can identify significant differences between novices and relative experts. We assess whether an approach that models local relationships among concepts in a domain provides useful information beyond what might be extracted from a more traditional modeling approach. Our results indicate that techniques that account for local correlation structure can identify patterns in discourse not reflected in global correlation structure.

Keywords: assessment, epistemological cognition, learning analytics, simulations

Introduction

Digital learning environments capture student discourse as they interact with mentors and peers to solve complex problems (Boulos, 2006). This poses a challenge for assessment of student work. While logged discourse contains concepts from the domain in which the problem is situated, what makes discourse better is typically not just that it contains more domain concepts, or even that it contains different concepts. Chi et. al.'s classic study of expertise in physics problems (1981), for example, found that experts organize their understanding differently than novices do. Similarly Bransford et. al. (1999) showed that experts have acquired a great deal of content knowledge that is organized in ways that reflect a deep understanding of their subject matter.

In other words, the ability to solve complex problems depends not only on having access to domain concepts, but also in understanding the appropriate relationships among them — and being able to mobilize those relationships in the context of real-world problem solving.

To assess complex thinking therefore may require constructing models of the way students use their understanding of the relationships among domain concepts during the problem solving process. In this paper, we look at two classes of mathematical models that can be used to measure this kind of connectivity in discourse. One is a class of models commonly used in educational research that looks at connectivity within the discourse as a whole — that is, the global relationship among concepts within students' discourse. The second approach is a more novel modeling tool that is sensitive to how concepts are related to one another within individual topics or activities within the discourse — that is the local relationship among concepts.

In what follows, we examine these two modeling approaches. In particular we look to see (1) whether the local relationship model can identify statistically significant differences between groups of novices and experts; (2) whether the statistically significant differences are meaningful on a closer qualitative analysis of the data; and (3) whether this same difference is identified in a global relationship model.

That is, we assess whether an approach that models local relationships among concepts in a domain provides useful information beyond what might be extracted from a more traditional modeling approach.

Theory

For several decades, work in the Learning Sciences has examined how complex thinking in a domain involves not only mastering basic skills and concepts, learning how these skills and concepts are systematically linked to one another. Summarizing a broad range of studies, Bransford et. al. (1999) describe the difference between experts and novices as being less about the amount of information that experts have then it is about the way that experts organize the knowledge that they have. diSessa (1988), for example, suggests that that while solving physics problems does require understanding basic concepts from the discipline, deep and systematic understanding comes from linking those concepts to one another within a theoretical framework. Novices have

what diSessa describes as “knowledge in pieces,” whereas experts understand the connections among different elements of the domain. Shaffer (2004) similarly characterizes learning as the development of an epistemic frame: a pattern of associations among knowledge, skills, habits of mind, and other cognitive elements that characterizes communities of practice, or groups of people who share similar ways of framing, investigating, and solving complex problems.

In other words, a good model of expertise needs to characterize way in which an individual (or group of individuals) understand the relationships among elements in the domain. To do so, of course, requires some context in which these relationships would be expressed: some record of how individuals approach problems in the domain. This might involve think-aloud protocols (Chi, 1997), problems that require students to “show their work” (McNeil, 2009), or records of group discussions during problem solving (Hmelo-Silver, 2004). That is, characterizing an individual’s understanding of the relationships among concepts and practices in a domain requires some record of work that can be analyzed. This record of work — which is often in the form of a logfile (Peled, 1999) or transcript — then needs to be annotated or coded for evidence of the key concepts and practices of interest. And finally, a model needs to be created that accounts for the relationships among the concepts and practices as reflected in these codes.

Traditional psychometric techniques provide several possible approaches to characterizing relationships of this kind. In theory, one could use correlation matrices to show the correlation structure of the codes based on their frequency in the logfile. However, it is difficult to compare multiple correlation matrices simultaneously due to the large amount of information contained within even a relatively small matrix (Alper, 2013) — which perhaps explains why the review we conducted did not find any examples analyzing concept-concept relations by comparing correlation matrices in the literature. Statistical techniques for analyzing the structure of correlation matrices can reduce the amount of information, making comparison possible, but summary statistics obtained from these techniques are geared towards analysis of single matrices (Cudeck, 1989). Thus, simultaneous comparison of many correlation matrices remains an active topic of research in quantitative methods and data visualization (Alper, 2013; Elmqvist, 2008).

Because it is difficult to compare correlation matrices, education researchers often use dimensional reduction techniques to analyze relationships as linear combinations of observed variables (Hall, 1977). For example, Beishuizen et. al. (2001) analyzed relationships among concepts in an essay writing activity. They identified the concepts of interest and computed the frequencies with which the concepts occurred in each essay. To identify which concept-concept relations were most common, they used a dimensional reduction to group concepts based on the structure of correlations in concept frequencies.

While there are many dimensional reduction techniques in the literature — including principal components analysis, factor analysis, item response theory, multidimensional scaling, and diagnostic classification models — most dimensional reduction techniques are similar in that they model correlation structure across all the data for a given unit of analysis. That is, they characterize the *global correlation structure* (GCS) of the data. In the analysis conducted by Beishuizen and colleagues (2001), for example, all correlations within each essay were considered equally meaningful.

In some cases, however, this approach may not accurately operationalize the relationships learners form among concepts. For example, research on discourse processing suggests that connections between concepts may be made primarily on a topic-by-topic basis rather than across discourse as a whole. Gernsbacher (1991; see also Graesser, Gernsbacher, and Goldman, 1997) argues that meaning is constructed through the hierarchical organization of ideas. A key element of this theory is that coherent discourse is structured by topic, with utterances having clear relationships to other utterances within topics, and few or no relationships across topics. Put another way, meaning is often localized within topics, and thus a model of how learners connect concepts to one another needs to account for this topic-based or *localized correlation structure* (LCS) of discourse.

One example of an approach that measures LCS’s is *Epistemic Network Analysis* (ENA), a suite of tools that can be used for identifying and quantifying connections among elements in coded data and representing them in dynamic network models. A key feature of ENA is that it enables researchers compare networks, both visually and through summary statistics that reflect the weighted structure of connections. ENA can be used to address a wide range of qualitative and quantitative research questions.

In this study, we look at a specific discourse context to explore whether GCS and LCS models suggest different interpretations of the discourse — and if so, which provides a more useful representation of the salient features of the data. That is, we examine the difference between a dimensional reduction technique that is insensitive to topical structure and a technique that is sensitive to topical structure. To do this, we analyze the chat discourse of high school and college students in *Land Science*, a virtual internship in which students work at a fictitious urban planning firm to solve an authentic urban redevelopment problem using two approaches: (1)

We use epistemic network analysis (ENA), which models the structure of relationships among domain concepts and practices using correlation structures that account for the topical structure of the discourse; (2) we then analyze the global correlation structures in the frequencies of these domain concepts and practices in the same data using principal components analysis (PCA). We compare the results of these analyses to determine whether and to what extent the two approaches find different structures of connections in student discourse.

Specifically, we ask:

RQ1: Are there statistically significant differences between novices' and relative experts' local correlation structures that can be detected using ENA?

RQ2: Are there meaningful differences between novices' and relative experts' local correlation structures?

RQ3: Does PCA detect these same differences in local correlation structure by measuring global correlation structures?

Methods

Land Science: A virtual internship in urban planning

In the virtual internship *Land Science* (Shaffer, 2008; Bagley, 2010; Bagley, 2015), students play the role of interns at Regional Design Associates, a fictional urban and regional planning firm. Their task is to develop a rezoning plan for the city of Lowell, Massachusetts that addresses the requests of various stakeholder groups. Students assess stakeholder preferences to understand what community members desire in terms of socio-economic and ecological issues. Not all of the stakeholders' competing concerns can be met, so students must make decisions about which demands to meet and how to meet them. To make these decisions, students discuss options with their project teams via online chat, and they use professional tools, such as a geographic information system model of Lowell and preference surveys, to model the effects of land-use changes and obtain stakeholder feedback. At the end of the internship, students write a proposal in which they present and justify their rezoning plans.

Coding of student chats in the *Land Science* logfile

All actions and interactions that occur during implementations of *Land Science* are recorded a log file. In this analysis, we focus on the chat conversations that students had while solving the rezoning problem. The logfile contains team chat conversations (41,332 lines of chat in total) from 265 students who used *Land Science*, including high school students (novices) (N = 110) and college students (relative experts) enrolled in an introductory urban science course (N = 155). The chat utterances were coded for 17 concepts and practices from the epistemic frame of urban planning, including:

Knowledge of stakeholder representation – knowledge of stakeholders, whose requests pertain to social, economic, and environmental issues

Skills and practices urban planning using tools of the domain – discussion or actions involving the tools – broadly defined – of the urban planning domain, such as a virtual site visit to key regions in the city, a stakeholder preference survey, and *iPlan*, a geographic information system-enabled zoning model

Data-based justifications – justifications using data such as graphs, results tables, numerical values, or research papers

We used an automated coding process based on conjunctive keyword and expression matches that was previously developed and validated (Bagley & Shaffer, 2015; Nash & Shaffer, 2011). Next, we analyzed the coded chats of relative experts and novices using ENA to measure the development of connections among elements of the urban planning epistemic frame. Then, we analyzed correlation structures in the relative frequencies with which students used these elements by PCA.

Epistemic network analysis (ENA)

Epistemic Network Analysis (ENA) is a method of identifying and quantifying connections among elements in coded data and representing them in dynamic network models. ENA enables researchers to compare networks, both visually and through summary statistics that reflect the weighted structure of connections.

To create network models of individual students' discourse, ENA creates a series of adjacency matrices for each student. Each adjacency matrix represents the co-occurrence of codes in one students' discourse during a single activity. The adjacency matrices are binary, meaning that co-occurrences of codes are indicated simply as present or absent: If two codes co-occur in an activity, a 1 is placed in the cell in the adjacency matrix for that activity that corresponds to the intersection of the two codes; cells for codes that do not co-occur in the activity receive a 0. Binary accumulation of co-occurrences is appropriate in this study because a student who says something twice as much does not necessarily understand it twice as well.

Each adjacency matrix thus represents the relations among the different urban planning concepts or practices made by one student in one activity, which means that each student is represented by a series of adjacency matrices. To identify the structure of connections made by each student, the adjacency matrices are summed into a cumulative adjacency matrix, where each cell represents the number of activities in which the unique pair of codes was present. The data set used for this study contained 17 activities coded for 18 urban planning epistemic frame elements. Thus, the cumulative adjacency matrix for a given student is the summation of 17 adjacency matrices, each with 153 (18 choose 2) possible unique co-occurrences of codes.

Once ENA creates the set of cumulative adjacency matrices for all the students in the data set, each matrix is converted into an adjacency vector by copying the cells from the upper diagonal of the matrix row by row into a single vector. These vectors exist in a high-dimensional space such that each dimension represents a unique pairing of two codes.

ENA spherically normalizes the adjacency vectors to calculate the relative frequencies of co-occurrence. In this high-dimensional ENA space, each adjacency vector represents the pattern of associations of a single unit, and the length of a vector is potentially affected by the number of activities that are contained in the student's discourse. More activities are likely to produce more co-occurrences, which results in longer vectors. This is problematic because two vectors may represent the same patterns of association, and thus point in the same direction, but represent different numbers of activities, and thus have different lengths. ENA solves this problem by spherically normalizing the vectors to unit Euclidean length. The resulting normalized vectors thus quantify for a student the relative frequencies of co-occurrence of codes independent of the number of activities in the model for any given student.

ENA then performs a dimensional reduction via singular value decomposition. (A singular value decomposition is similar to a principal components analysis, but it does not rescale the data.) This provides a rotation of the original high-dimensional ENA space, such that the reduced number of dimensions in the rotated space capture the maximum variance in the data. For every student in the data, ENA creates a point that is the location of the normalized vector under the singular value decomposition.

Finally, ENA positions the network nodes. Normally with dimensional reduction techniques, the basis vectors, or the *loadings*, would tell us how to interpret the positions of points in the space. However these basis vectors represent connections between codes in the original data. That is, each point represents one of the cells in the cumulative adjacency matrix. That makes it hard to interpret, because if we have 18 codes, as in this study, then we can have up to 153 basis vectors, each of which corresponds to a unique co-occurrence of codes. To interpret the dimensions of this rotated space, ENA uses an optimization routine to position nodes in ENA space such that for any student, the centroid of the student's network model, corresponding to the student's cumulative adjacency matrix, is as close as possible to the location of the projected point. The utility of this correspondence is that we can compare the structures of networks by comparing the locations of their projected points. Projected points are positioned such that their associated networks will have their strongest connections distributed as weightings relative to the positions their projected points. Thus, the position of a point in ENA space summarizes the structure of connections in the networks being modeled.

The result is that: the ENA dimensional reduction models the variance among the different networks being analyzed; the corresponding network graphs allow us to interpret the significance of the locations of points in the ENA model; and we can interpret what aspects of the network structure explain the differences between units in the model.

Principal components analysis (PCA)

We computed the frequencies of each student's codes in the logfile and performed a principal components analysis using the R language's prcomp function. The code frequencies were standardized such that the variance in the frequencies of each code was equal to one, and the dimensionality of the code frequency space was reduced by singular value decomposition. Student code frequencies were then projected into the reduced principal component space.

Results

RQ1: Are there statistically significant differences between novices' and relative experts' local correlation structures that can be detected using ENA?

We computed the structure of connections made by each student using ENA and projected them into the space created by the dimensional reduction. Figure 1 shows the projected points of each student's network in ENA space. The projected points showed statistically significant differences between novices (blue) and relative experts (red) on both the first dimension ($mean_{expert} = –0.123$, $mean_{novice} = 0.088$; $t = 7.446$, $p < 0.001$, Cohen's $d = 0.969$) and the second dimension ($mean_{expert} = –0.023$, $mean_{novice} = 0.032$; $t = –2.031$, $p = 0.043$, Cohen's $d = –0.258$).

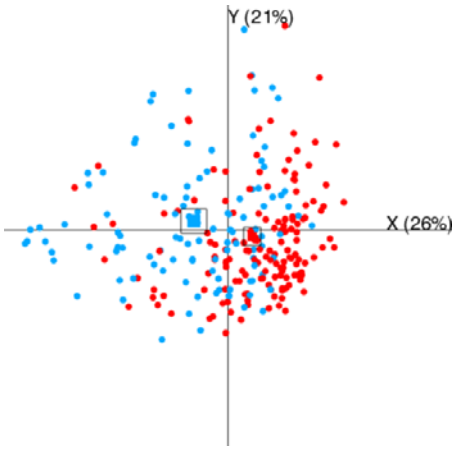

Figure 1. ENA scatter plot showing novice (blue) and relative expert (red). Each point is a single student; the squares are group means; the boxes are 95% confidence intervals (t-distribution) on each dimension; the numbers in parentheses indicate the percentage of variance in the data accounted for by that dimension.

To investigate which connections accounted for the differences between the two groups, we compared their mean epistemic networks. One of the most salient differences between the mean networks was in connections to data-based justifications (lower right in Figure 2; complementary colors indicate connections with data-based justifications). The novice network (blue) showed that data-based justifications were connected only with knowledge elements, while the relative expert network showed that data-based justifications were connected with knowledge elements, skills and actions, and other justification codes. In other words, relative experts made more and more diverse connections to data-based justifications than novices. To understand the meaning of these differences in local correlation structure, we analyzed instances of data-based justification in the student chats.

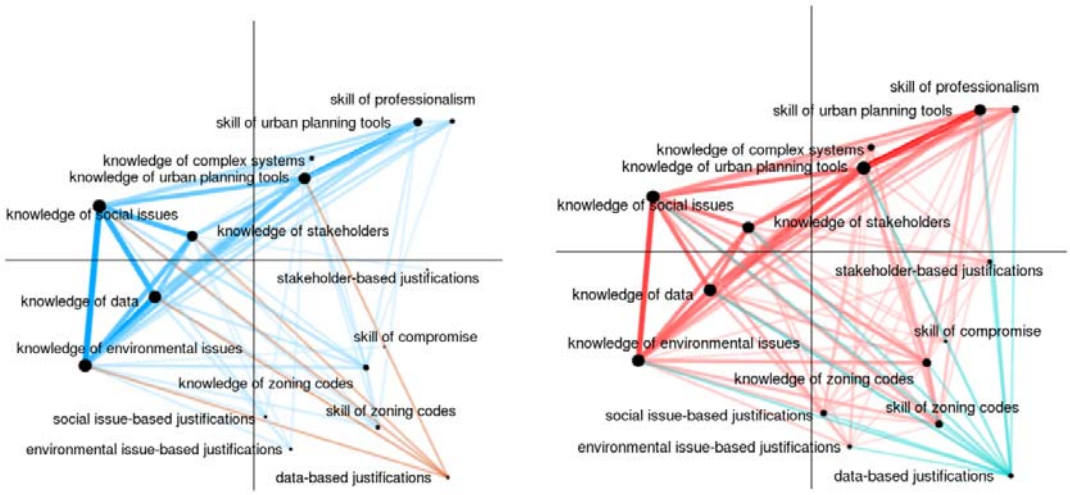

Figure 2. Mean ENA network diagrams showing the connections made by the two groups of students described in Figure 1. Novices (left) connected data-based justifications with knowledge elements; relative experts (right) connected data-based justifications with knowledge elements as well as with skills and other justifications.

RQ2: Are these differences meaningful on closer qualitative analysis of the data?

In one activity in *Land Science*, students are introduced to *iPlan*, the geographic information system tool with which they model how zoning changes affect the socioeconomic and environmental indicators that the stakeholders care about. Students made changes to their plans and observed the effects of changes on the indicator graphs. After finalizing their changes, students used the chat tool to participate in a reflective discussion with their assigned mentors.

In the chat discussions, novices made connections primarily between (1) knowledge of urban planning concepts and practices and (2) data-based justifications. One novice (below) explained the stakeholder Hao's request (red, knowledge of stakeholder representation) and explained the warrant for this request (blue italics: data, blue underline: the warrant).

> Hao says that *the number of Baltimore Orioles is decreasing* as a result of the development of the town. She claims that an environment that is not healthy for birds is not healthy for us.

This connection between stakeholder representation and warranting a stakeholder request on the basis of data obtained in *iPlan*'s graph indicator for animal life population is an example of novices justifying domain-relevant knowledge using data.

While novices connected data-based justifications with domain-relevant knowledge, relative experts connected data-based justifications with a wider range of the domain's concepts and practices, indicating a more sophisticated grasp of the domain. For example (below), in the same activity, a relative expert asserted that, whatever her team's next changes in their plan might be, they must increase the amount of housing (red, skills with urban planning tools). The expert justifies her assertion (blue italics: data-based warrant) by appeal to socioeconomic issues (population), environmental issues (runoff), which she can know only by making adjustments with iPlan and observing changes in graph indicators.

> Going back to Colby's original question, I think the plan may be forced to increase housing anyways, *due to the increasing population, and the runoff is inevitable*.

Thus, in this example, the relative expert built a data-based argument for a design decision in the domain. In contrast, the novice made a data-based argument to justify a specific piece of knowledge he had acquired. The same distinction is seen as well for other types of justification than those based on data. Figure 3 depicts the same mean networks as Figure 2, but the only connections shown are those to justification codes. As was the

case with data-based justifications, relative novices connect other types of justifications primarily with knowledge elements, whereas relative experts connect justifications with knowledge, skills and actions, and other justification codes.

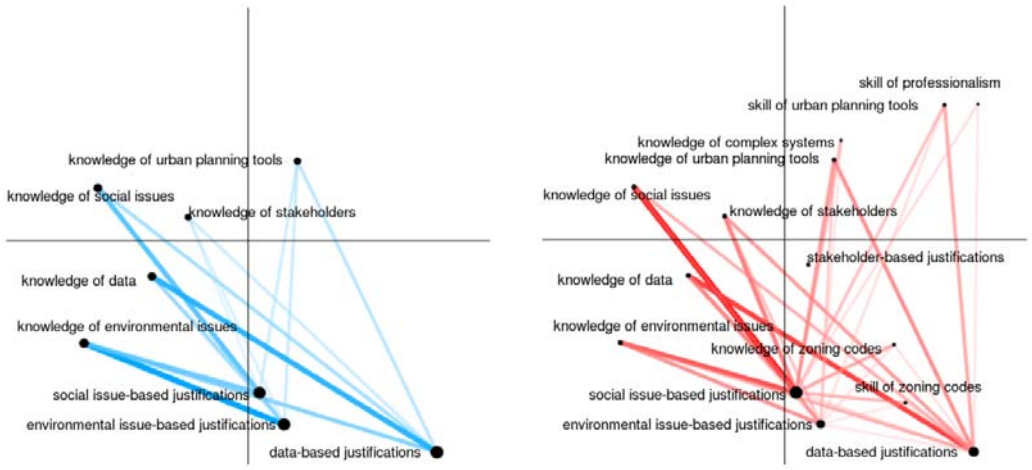

Figure 3. Mean ENA network diagrams of novices (left) and relative experts (right) showing only connections to justification codes. All other connections are hidden.

RQ3: Does PCA detect these same differences in local correlation structure by measuring global correlation structures?

We investigated whether there were differences in the correlation structures of the code frequencies of novices and relative experts using a common dimensional reduction technique, PCA. We computed students' code frequencies, created a reduced space using PCA, and projected the code frequencies into this space. Significance tests did not show significant differences between the novices' and relative experts projections on either PCA component one ($mean_{novice}$ = -0.0814, $mean_{expert}$ = -0.2252; t = 0.6625, p = 0.5083) or PCA component two ($mean_{novice}$ = -0.08983, $mean_{expert}$ = 0.2529; t = -1.5774, p = 0.1159). In other words, the result obtained above using ENA that models using the local nature of students' cognitive connections in discourse is not show by PCA, which relies on correlation structures in code frequencies overall in each students' data.

Discussion

Our results showed that ENA was able to detect meaningful differences in the logged discourse of *Land Science* where PCA was not. Our PCA approach, which measured the global correlation structures in the frequencies of concepts and practices from the urban planning domain as they appeared in the student discourse, was unable to distinguish the novices from relative experts. ENA, which measured concepts' local correlation structures within activities, was able to make this distinction. Moreover, these distinctions were meaningful in a qualitative analysis of the data. In sum, our results support the claims of diSessa, Shaffer, and others, that student discourse is appropriately analyzed using correlation structures, and more appropriately analyzed using correlation structures that are sensitive to local contexts in which concepts are connected to one another.

This study had several limitations. First, we compared one example of a an approach which measures local correlation structures and one example of an approach that measures global correlation structures, although PCA is a very common tool used for analysis of global correlation structure and there are few examples besides ENA of techniques that systematically model local correlation structure in discourse. There is also, of course, the problem that this study is based on the analysis of only one data source. Despite these limitations, however, the work here suggests local correlation structure-based approaches may be more appropriate than traditional global correlation structure-based methods for assessing that assessment student discourse.

References

Alper, B., Bach, B., Riche, N.H., Isenberg, T., & Fekete, J. (2013). Weighted graph comparison techniques for brain connectivity analysis. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13. (pp. 484-492). New York, NY, USA. ACM.

Bagley, E. (2010). The epistemography of an urban and regional planning practicum: Appropriation in the face of resistance. WCER Working Paper 2010-8. Wisconsin Center for Education Research, University of Wisconsin-Madison.

Bagley, E. & Shaffer, D. W. (2015). Learning in an urban and regional planning practicum: The view from educational ethnography. Journal of Interactive Learning Research, 26(4), 369-393.

Beishuizen, J. J., Hof, E., van Putten, C. M., Bouwmeester, S. & Asscher, J. J. (2001). Students' and teachers' cognitions about good teachers. British Journal of Educational Psychology, 71, 185-201.

Boulos, M. N. K., Maramba, I. & Wheeler, S. (2006). Wikis, blogs and podcasts: a new generation of web-based tools for virtual collaborative clinical practice and education. BMC Medical Education, 41(6).

Bransford, J. D., Brown, A.L., & Cocking, R.R. (1999). How people learn: Brain, mind, experience, and school. National Academy Press, Washington, D.C.

Chi, M. T. H., Feltovich, P. & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5(2), 121-152.

Chi, M.H.T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. Journal of the Learning Sciences, 6(3), 271-315.

Cudek, R. (1989) Analysis of correlation matrices using covariance structure models. Psychological Bulletin, 105(2), 317-327.

diSessa, A. A. (1988). Knowledge in Pieces. In Forman, G. & Pufall (Eds). Constructivism in the Computer Age, New Jersey: Lawrence Erlbaum Publishers.

Elmqvist, N., Do, T. N., Goodell, H., Henry, N., & Fekete, J. D. (2008). ZAME: Interactive large-scale graph visualization. In Proceedings of IEEE Pacific Visualization Symposium (pp. 215-222). IEEE.

Gernsbacher, M.A. (1991). Cognitive processes and mechanisms in language comprehension: The structure building framework. In G. H. Bower (Ed)., The psychology of learning and motivation (pp. 217-263). New York: Academic Press.

Graesser, A. C., Gernsbacher, M. A., & Goldman, S.R. (1997). Cognition. In T. A. van Dijk (Ed)., Discourse: A multidisciplinary introduction (pp. 292-319). London: Sage.

Hall, C. (1977). Dimensional reduction analysis. Journal of Experimental Education, 45(4), 4-8.

Hmelo-Silver, C.E. (2004). Problem-based learning: What and how do students learn? Educational Psychology Review, 16(3), 235-266.

McNeil, N.M., Uttal, D.H., & Sternberg, R.J. (2009). Should you show me the money? Concrete object both hurt and help performance on mathematics problems. Learning and Instruction, 19(2), 171-187.

Nash, P. & Shaffer, D.W. (2011). Mentor modeling: The internalization of modeled professional thinking in an epistemic game. Journal of Computer Assisted Learning, 27(2), 173-189.

Peled, A., & Rashty, D. (1999). Logging for Success: advancing the use of WWW logs to improve computer mediated distance learning. Journal of Educational Computing Research, 21(4), 413-431.

Shaffer, D. W. (2004). Epistemic frames and islands of expertise: Learning from infusion experiences. In Proceedings of the 6th international conference on Learning Sciences (pp. 473-480). Santa Monica, CA: International Society of the Learning Sciences.

Shaffer, D. W. (2006). Epistemic frames for epistemic games. Computers and Education, 46(3), 223-234.

Shaffer, D. W. (2008). How computer games help children learn. New York, NY: Palgrave Macmillan.

Shaffer, D. W., Hatfield, D. L., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E. A., & Frank, K. (2009). Epistemic Network Analysis: A prototype for 21st century assessment of learning. International Journal of Learning and Media, 1(1), 1-21.

Shaffer, D. W. (2014). User guide for Epistemic Network Analysis web version 3.3. Madison, WI: Games and Professional Simulations Technical Report 2014-1.

Acknowledgments

This work was funded in part by the National Science Foundation (DRL-0918409, DRL-0946372, DRL-1247262, DRL-1418288, DUE-0919347, DUE-1225885, EEC-1232656, EEC-1340402, REC-0347000), the MacArthur Foundation, the Spencer Foundation, the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.