Applying Group Communication Analysis to Educational Discourse Interactions at Scale

Nia M. Dowell, University of Michigan, ndowell@umich.edu
Christopher Brooks, University of Michigan, brooksch@umich.edu
Oleksandra Poquet, National University of Singapore, sasha.poquet@nus.edu.sg

Abstract: The learning sciences field is in need of new automated methodological approaches that offer deeper insights into the dynamics of learner interaction and discourse across scaled learning platforms. In this paper, we explore MOOC learners' discourse by employing *Group Communication Analysis* (GCA), a methodology for quantifying and characterizing the discourse between learners in online interactions. Commonly used approaches in MOOCs derive insight into the learning processes from aggregated text or structural data. In contrast, GCA makes use of linguistic cohesion analysis across sequences of learners' interactions in multi-party communication. GCA calculates six inter- and intra-personal sociocognitive measures of such interactions and from these identify distinct interaction profiles through a cluster analysis. With this method, we were able to diagnostically reveal four robust profiles amongst MOOC learners. This study presents a unique analysis of the sociocognitive processes that comprise the interaction between learners. The scalability of the methodology opens the door for future research efforts directed towards understanding and improving scaled peer-interactions.

Introduction

The importance of peer interactions for the learning process has been emphasized in learning sciences and educational research (Bransford, Brown, & Cocking, 2000; Stahl, Koschmann, & Suthers, 2006). Studies in the distance courses, online and blended courses, and more recently in Massive Open Online Courses (MOOCs) environments have all stressed the need for developing peer to peer interactions to promote student learning and achievement of course goals (Bernard et al., 2009; Borokhovski, Tamim, Bernard, Abrami, & Sokolovskaya, 2012; Joksimović, Gašević, Loughin, Kovanović, & Hatala, 2015). Technology-based interactions, such as blogs and forums, play a key role in facilitating discussions among peers, instructors, and teaching assistants. While MOOC forums hold the potential for scalable peer-based learning, they are typically characterized by low overall participation and responsiveness (Yang, Sinha, Adamson, & Rose, 2013). Thus, a question facing the Learning at Scale community is how the potential of large scaled peer-learning environments can be achieved.

A burgeoning research body of literature focuses learning processes in groups across computer-supported collaborative work and broader learning sciences (Stahl et al., 2006; Stahl & Rosé, 2013; Suthers, Dwyer, Medina, & Vatrapu, 2010), often with the emphasis on sociocognitive group processes, such as coordination, common ground, elaboration and integration of ideas. Applying these well-developed theoretical lenses at scale, such as in MOOCs, is challenged by the vast amount of data. In order to move forward, the field is in need of new automated approaches that offer deeper insights into the dynamics of learner interaction and discourse across scaled learning platforms.

Scaled online educational platforms afford the collection of learning data at high granularity. However, thus far, efforts towards characterizing the dynamics of MOOC learners' interactions in forums have been limited to aggregate and surface level features. In cases where the analysis of MOOC forums reaches beyond quantifying the number of forum contributions, the efforts to derive meaning around peer interactions have been driven by SNA and automated content analysis of forum posts. Studies employing social network analysis (SNA) (Boroujeni, Hecking, Hoppe, & Dillenbourg, 2017; Dowell et al., 2015; Hecking, Chounta, & Ulrich Hoppe, 2017; Poquet, Dawson, & Dowell, 2017) characterize learners' social interactions in terms of reply-to behaviors and network attributes. Insights derived from SNA are limited because these strictly structural measures do not capture the deeper level interpersonal sociocognitive and semantic information found in the discourse interaction. Alternatively, automated natural language processing techniques could provide a productive path towards characterizing scaled peer-learning interactions. Indeed, language and communication has proven quite useful in explorations of group interaction phenomena, and has been applied to characterize the quality of learner-generated text at scale (e.g., Joksimović et al., 2018). Automated content analysis of forum posts at scale has been applied at both the student level (i.e. individual posts or totality of them per person) and group level (ie. text of the overall thread transcript). Aggregating the text of the individual or a thread offers a summative account of learner or group discourse characteristics, but it provides only coarse-level granularity, and disregards the sociocognitive processes that reside in the interaction between learners' discourse contributions. Such approach is inadequate for obtaining insights around many core peer interaction issues because the analyses focus on individual cognition, rather than sequential meaning-making characteristics. In particular, these practices tend to obscure the sequential structure, semantic references within group discussion, and situated methods of interaction through which learning emerges (Çakır, Zemel, & Stahl, 2009; Reimann, 2009; Stahl, 2017; Suthers et al., 2010). As a result, current studies of peer discourse at scale cannot offer insight on many aspects of peer interaction such as coordination, and regulation, among others, and more nuanced techniques are needed.

To start unpacking sociocognitive processes at scale, learning and computational sciences require new automated methodological approaches that will provide deeper understanding of learners' communication patterns and interaction dynamics across MOOC platforms. Drawing on this, we explore MOOC learners' discourse by employing *Group Communication Analysis* (GCA), a methodology for quantifying and characterizing the discourse dynamics between learners in online multi-party interactions (Dowell, Nixon, & Graesser, 2018 under-review). GCA applies automated computational linguistic analysis to the *sequential* interactions of participants in online group communication. GCA both captures the structure of the group discussion, and quantifies the complex semantic cohesion relationships between learners' contributions overtime, revealing interpersonal processes in group communication. In doing so, this methodology goes beyond previous models for automated group communication, which often rely on counting the number of utterances exchanged between learners.

This study is focused on identifying learner interaction profiles based on the quality of each person's contributions to the forum discourse. Cluster analysis was used to identify prototypical MOOC learner interaction profiles. Each profile comprises six sociocognitive measures resulting from the GCA methodology of responsiveness, social impact, internal cohesion, newness, communication density, and participation. Measures used for cluster analysis are derived by tracking the sequential intra- and interpersonal patterns of cohesion between participants' discourse contributions. As such, the GCA provides a novel approach to describing learner behavior in MOOC settings. Its significance is in opening for further investigations of group level learning processes at scale, such as monitoring and social regulation, integration of ideas, creation of the common ground, sharing of information, and deepening of group ideas.

Theoretical foundations and group communication analysis

The GCA framework incorporates definitions and theoretical constructs that are based on research and best practices from several areas where sociocognitive processes, group interaction, and collaborative skills have been assessed. These include areas such as CSCL, CSCW, teams, organizational psychology, assessment in work contexts and PISA (Oecd, 2013). Despite differences in orientation between the disciplines where these frameworks have originated, the conversational behaviors that have been identified as valuable are quite similar, and the GCA provides six learning-relevant interaction measures (summarized in Table 1), which are briefly reviewed below (Dowell et al., 2018 under-review).

Posting a message on the forums is often operationalized by researchers and instructors as participation (Hrastinski, 2008) and considered a requirement for any online learning group interaction. It signifies a willingness and readiness of learners to externalize and share information and thoughts (Hesse, Care, Buder, Sassenberg, & Griffin, 2015). Participation, has been shown to have a beneficial influence on various learning outcomes, including retention rates, learner satisfaction, and social capital (Hrastinski, 2008). GCA approaches ***participation*** as a necessary, but not sufficient component for characterizing the interactions between MOOC learners.

Internal cohesion is a sociocognitive measure that can serve as a proxy for individual self-monitoring and reflection processes during peer interactions. That is, successful collaboration requires that each individual monitor and reflect on their own knowledge and contributions to the group (Barron, 2000; Oecd, 2013); a behavior explained within self-regulation theory (Chan, 2012; Malmberg, Järvelä, & Järvenoja, 2017; Zimmerman, 2001). Consequently, during peer-learning individuals need to appropriately build on and integrate their own views with those of the group (Kreijns, Kirschner, & Jochems, 2003). Given that a participant's current and previous contributions should be, to some extent, semantically related to each other, a measure of internal cohesion can indicate the extent to which they have monitored and reflected on their previous discourse (i.e. self-regulation). Overly high levels of internal cohesion might suggest that a participant is not evolving their thoughts, but rather reiterating the same static view. Conversely, low levels of internal cohesion might indicate that a participant has no consistent perspective to offer the conversation, and is echoing the views of others, or is only engaging at a surface level within discussion thread topics.

Learners must also monitor and build on the perspectives of their collaborative partners to achieve and maintain a shared understanding of the task and its solutions (Dillenbourg & Traum, 2006; Graesser, Dowell, & Clewley, 2017; Hmelo-Silver & Barrows, 2008; Stahl & Rosé, 2013). In the CSCL literature this shared understanding has been referred to as knowledge convergence, or common ground (Clark & Brennan, 1991; Roschelle & Teasley, 1995). It is achieved through communication and interaction, such as building a shared representation of the meaning of the goal, coordinating efforts, and viewpoints of group members, and mutual monitoring of progress towards the solution. ***Responsivity*** is a sociocognitive GCA measure, which captures monitoring and regulatory processes externalized during communication with peers. This measure reflects the extent to which an individual monitors and incorporates the information provided by the peers in their new contributions. The measure is implemented by examining the semantic relatedness between the individual's current contribution and the previous contributions of their collaborative partners. For example, if an individual's contributions are, on average, only minimally related to those of their peers, it would the individual exhibits low responsivity.

The GCA's ***social impact*** measure captures the extent to which a learners' contributions are seen as meaningful, or worthy of further discussion (i.e. uptake), by their peers. Social impact is measured through the analysis of the semantic relatedness between the learner's current contribution and those that follow from their collaborative partners. Individual messages that are more semantically related to the subsequent contributions indicate a high social impact of their authors on the unfolding group discourse.

Peer interactions provide the opportunity to expand the pool of available information, thereby enabling groups to reach higher quality solutions than could be reached by any one individual. However, despite the intuitive importance of (new) information sharing, a consistent finding from research is that groups predominantly discuss information that has been already shared (known to all participants) at the expense of information that has not been shared (known to a single member) (see Mesmer-Magnus & Dechurch, 2009 for a review). The distinction between given (old) information versus new information in discourse is a foundational distinction in theories of discourse processing (Price, 1981). Given information includes words, concepts, and ideas that have already been disclosed in the discourse; new information involves words, concepts and ideas that have not yet been mentioned, and builds on the given information or launches a new thread of ideas. The GCA captures the extent to which learners provide new information, compared to referring to previously shared information, with a measure called ***newness***.

The team performance literature also advocates for concise communication between group members (Gorman, Cooke, & Kiekel, 2004). An example of this can be seen in formal teams, like military units, which typically adopt conventionalized terminology and standardized patterns of communication. It is suggested that this concise communication is possible when there is more common ground within the team and the presence of shared mental models of the task and team interaction (Klein, Feltovich, Bradshaw, & Woods, 2005). The GCA's ***communication density*** measure was first introduced by Gorman et al. (2003) in team communication analysis to measure the extent to which a team conveys information in a concise manner. Specifically, the rate of meaningful discourse is defined by the ratio of semantic content to number of words used to convey that content.

Semantic-based GCA measures

Five of the GCA measures are semantic-based metrics (i.e., all but participation). The GCA relies on Latent Semantic Analysis (LSA) to infer the semantic relationship among the individual contributions. LSA, an automated high-dimensional associative analysis of semantic structure in discourse, can be used to model and quantify the quality of coherence by measuring the semantic similarity of one section of text to the next. LSA represents the semantic and conceptual meanings of individual words, utterances, texts, and larger stretches of discourse based on the statistical regularities between words in a large corpus of natural language (Landauer, McNamara, Dennis, & Kintsch, 2007). When used to model discourse cohesion, LSA tracks the overlap and transitions of meaning of text segments throughout the discourse.

Conversations, including MOOC forum discussions, commonly follow a statement-response structure, in which new statements are made in response to previous statements, and subsequently trigger further statements in response. Learners may, in a single contribution, refer to concepts and content presented in multiple previous contributions, made throughout the conversation either by themselves or other learners. Thus, a single contribution may be in response, to varying degrees, to many previous contributions, and it may in turn trigger, to varying degrees, multiple subsequent responses.

The analytical approach of the GCA was inspired by analogy to the cross- and autocorrelation measures from time-series analysis. Cross-correlation similarly measures the relatedness between two variables, but with a given interval of time (or lag) between them. That is, for variables x and y, and a lag of τ, the cross-correlation would be the correlation of $x(t)$ with $y(t+\tau)$, across all applicable times, t, in the time-series. Such cross-correlation plots are a commonly used in the qualitative exploration of time series data. While we might apply standard auto- and cross-correlation to examine temporal patterns in *when* participants contribute, we are primarily interested in understanding the temporal dynamics of *what* they contribute, and what the evolution of the conversation's semantics can teach us about the scaled peer-interaction. With this in mind, the GCA provides a fine-grained measure of the similarity of participants' contributions to capture the multi-responsive and social impact dynamics that may be present in online interactions. That is, the semantic cohesion of contributions at fixed lags in conversations can be computed much in the same way that cross-correlation evaluates correlation between lagged variables. Various measures of this auto- and cross-cohesion form the basis of the GCA's semantic-based measures.

Methods

MOOC and participants

We analyzed forum discussions from an eight-week long iteration of the course [removed for review] offered on the Coursera platform in 2013. The subject area of the analyzed MOOC fell under the domain of a combination of learning and life sciences. The course objective was to introduce the educational theory as it relates to health professionals, familiarize students with a variety of teaching techniques, as well as with the practical approaches for matching instructional methods with desired educational outcomes. The dataset for the analysis in this study included 644 participants, i.e. all those who posted messages on the course forum. Forum data was collected from the Coursera platform and included all the information specified within the Coursera discussion forums data documentation.

Modeling conversational structure and preprocessing

MOOC forums make use of a tree hierarchical data structure of enchained messages: threads, which initiate a new discussion; posts, which are messages on a thread; and comments, which are messages used to reply to a post. During the 8-week course period for the MOOC used in this research, there were 180 threads, 2,335 posts, and 1,437 comments. The number of contributions (i.e., comments or posts) varied across threads (*M*=137.80, *SD*=81.43, *Q1*=85, *Q3*=209, *Min*=1, *Max*=244). The GCA was applied to all threads individually.

Conversations within a MOOC thread can be organized in two ways, visual or temporal. Under the visual approach, all posts and associated comments on thread would be organized as one would view them if they browsed at a historical (i.e., completed course) thread. That is, the order of the conversation respects the labeled dependency between posts and comments on posts, but ignores the temporal order of the contributions. Consequently, the visual organization hides the true semantic relations between learner contributions. As the name suggests, the temporal approach orders learners' contributions in temporal order. That is, this organization reflects the way a learner would encounter posts and comments at a given moment in time, thereby retaining the temporal semantic relations between learner contributions. The temporal ordering more accurately represents the evolving development of learners' ideas over time and thus was used to analyze the data in this study. Prior to the GCA analysis, all comments and posts were preprocessed and information that was not a part of the actual discourse (e.g., HTML) was removed in a cleaning process. The openNLP R library (Hornik, 2016) was used for word tokenization, sentence segmentation, and parsing. Table 1 provides overview of GCA measures.

Table 1: GCA Measures

Measure	Description
Participation	Mean participation of any participant above or below what you would expect from equal participation in a group of the size of theirs
Overall Responsivity	Measure of how responsive a participant's contributions are to all other group members' recent contributions
Internal Cohesion	How semantically similar a participant's contributions are with their own recent contributions
Social Impact	Measure of how contributions initiated by the corresponding participant have triggered follow-up responses
Newness	The amount of new information a participant provides, on average
Communication Density	The amount of semantically meaningful information

Analysis

A k-means cluster analysis approach was adopted to discover communication patterns associated with specific learner profiles during MOOC forum interactions. Cluster analysis is a common data mining technique that involves identifying subgroups of data within the larger population who share similar patterns across a set of variables. Cluster analysis has been previously applied in the analysis of learner behavior in MOOCs (e.g., Kizilcec, Piech, & Schneider, 2013) and has proven useful in building an understanding of individuals' behaviors in many digital environments more broadly (e.g., Wise, Speer, Marbouti, & Hsiao, 2012).

In our analysis, we first compute the mean for each learner across the six GCA measures, which provides a global account of individuals interaction dynamics across the conversations in the MOOC forums. The data were then normalized and centered to prepare them for analysis. Prior to clustering, multicollinearity, collinearity and cluster tendency were assessed through variance inflation factor (VIF), Pearson correlations, and Hopkins statistic, respectively. The results support the view that multicollinearity and collinearity were not an issue with VIF > 7, and at $|r| \geq 0.7$. The assessment of cluster tendency is a particularly important in the context of unsupervised machine learning because clustering methods will return clusters even if the data does not contain any inherent clusters. The Hopkins statistic (factoextra R package) did show evidence of clustering, H = .08, which is well below the threshold of H > .5 (Han, Pei, & Kamber, 2011, p. 486). Given that cluster analysis can return any number of specified clusters, we used the NbClust R package which provides 26 indices for determining the relevant number of clusters (Charrad, Ghazzali, Boiteau, & Niknafs, 2014). Detailed specification of each index can be found in Charrad et al. (2014), and majority vote between indices indicated k=4 was appropriate.

Findings

Investigation of the cluster centroids helps identify if the clusters are conceptually distinguishable. Centroids represent the prototypical entity (learner) in each cluster. With K-means, the centroids are in the means of the points in the cluster. In the context of GCA profiles, we may interpret the centroids as measures of behavior typical of a distinct style of interaction. The centroids for the four-cluster k-means solution are presented in Figure 1. The clustering was performed on normalized data; zero in this figure represents the population average (all 644 learners) for each measure, while positive and negative values represent values above or below that average.

We see some interesting patterns across the four-cluster solution. Cluster 1 (*N* = 135) were above average participators; they also exhibit above average social impact, responsiveness, newness and communication density, coupled with high internal cohesion. Learners in these clusters are investing a high degree of effort in the discussions and display self-regulatory and social-regulatory skills. Not only are these individuals responsive to other learners, but they are engaging in effective information sharing and their contributions seem to warrant further discussion from the group members or provide new information (i.e., social impact and newness). This suggests these learners are invested in the prevailing social climate. Cluster 2 (*N* = 363), makes up more than half of the MOOC learners. Here, we see learners with only average participation, but when they do contribute, they attend to other learners' contributions and provide meaningful information that furthers the discussion (i.e., overall responsiveness, and social impact). It is interesting to note that these students are not among the highest participators, but their discourse signals a social positioning that is conducive to a productive exchange within the MOOC interactions. This pattern is suggestive of a student that is engaged in the interaction, but perhaps takes a more thoughtful and deliberative stance, than learners in cluster 1. Cluster 3 (*N* = 27), is the smallest cluster and was characterized by learners who have the highest participation, newness, communication density and second highest internal cohesion, however very low scores across all other measures. This suggests that, when they contribute, their discourse is more in response to themselves than other MOOC learners since they exhibit relatively higher internal cohesion than responsiveness or social impact. Furthermore, their contributions do not seem to warrant further discussion from learners. This pattern might be reflective of individuals driving the discourse agenda forward by offering a larger volume of contributions as compared the Cluster 1 or 2. In contrast to the cluster 1 and cluster 2 these learners have a higher degree of internal cohesion compared to social impact or responsiveness, which signals the they may be more concerned with a personal narrative than with productive peer interactions. It is plausible that these learners are highly motivated, exhibiting strong individual learning goals through their leadership and at times dominance in the discourse. Cluster 4 (*N* = 119), in contrast, is characterized by a combination of low responsivity, internal cohesion, social impact, and participation, coupled with high newness and average communication density. This pattern is similar to lurking and off-topic behaviors. Overall, the four-cluster model appears, at least upon an initial visual inspection, to produce meaningful MOOC learner profiles. We then proceeded to evaluate the quality and validity of this model.

In the current research, we evaluated the internal, cluster coherence, and stability validation criteria. A commonly reported internal validity measure, Silhouette, measures how well an observation is clustered by estimating the average distance between clusters and ranges from -1, indicating that observations are likely placed in the wrong cluster to 1, indicating that the clusters perfectly separate. The average silhouette (AS) for our model was positive (AS = .38), indicating the MOOC learners in a cluster had higher similarity to other learners in their own cluster than to students in any other cluster. A MANOVA, ANOVAs, followed by Tukey's *post hoc* were used to evaluate cluster coherence or the extent to which learners in the cluster groups significantly differed from each other on the six GCA variables. This coherence evaluation showed that the four-cluster model exhibited nice separation across the GCA measures. The stability and validity of the cluster model was assessed using a non-parametric bootstrapping procedure (B=100 runs), which resampled from the original data with replacement to construct bootstrap matrices and clusters, and iteratively used the *Jaccard coefficient* to compute the structural similarity of the resampled clusters with the cluster derived from the original data. The Jaccard's similarity values showed very strong prediction for all four clusters with .96, .97, .93, and .88 for clusters 1-4, respectively. Given the extent of these evaluations, we feel that the identified learner profiles can be considered as robust and stable constructs in the space of scaled peer interactions, and that the GCA measures capture the critical socio-cognitive processes necessary for identifying such profiles.

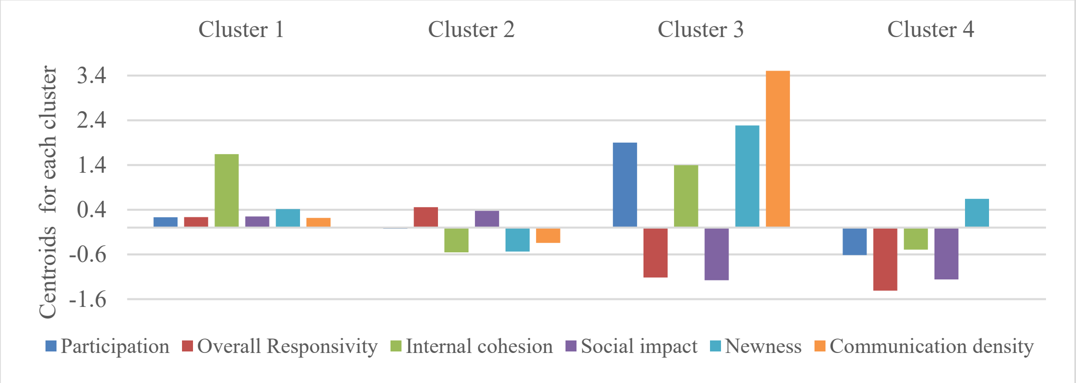

Figure 1. Centroids for the four-cluster solution across the GCA variables.

Discussion

A primary objective of this research was to evaluate the application of a new methodological approach, *Group Communication Analysis* (GCA), in the context of scaled peer-interactions. There are multiple lenses by which one might view discourse and learner profiles in scaled learning environments (Chua, Tagg, Sharples, & Rienties, 2017; e.g., Hecking, Chounta, & Hoppe, 2016; Kizilcec et al., 2013). The GCA lets us view discourse as a dynamic and evolving sociocognitive process that resides in the interaction between learner's communicative contributions. Our results suggest that these sociocognitive discourse patterns, as captured by the GCA, diagnostically reveal MOOC learner interaction profiles, and that the observed patterns are robust. The study presents initial interpretation of the clusters and the methodological validation of such profiling. However, further validation of these profiles is needed including triangulating qualitative analysis of individual learner discourses, as well as replicating the results in other contexts, and understanding the behaviors and learning outcomes associated with different profiles.

The findings present some methodological, and practical implications scaled peer-interaction research in the L@S, Artificial Intelligence in Education (AIED), and Learning Science communities. The application of the GCA to scaled peer-interactions represents a novel methodological contribution, capable of detecting distinct patterns of interaction representative of the learner profiles in MOOC interactions. The natural language metrics that make up the GCA provide a mechanism to operationalize such profiles, and insights into how they are constructed and maintained through the sociocognitive processes within scaled learning interactions. Moreover, as the methodology is readily automated, substantially larger corpora can be analyzed with the GCA than is practical when human judgements are required to annotate the data. Ideally, the GCA provides the greater learning sciences community with a toolkit to obtain objective, domain independent, and deeper explorations of the micro-level inter- and intra-personal patterns associated learners' interaction profiles.

The individual GCA measures and the detection of MOOC learners' interaction profiles could be of practical value for instructional designers of MOOCs. For instance, the internal cohesion of learners in cluster 1 was exceptionally high in comparison to the more socially oriented measures, like responsivity or social impact. Instructional designers could use this information to provide real-time scaffolding to encourage learners to reflect more on their peer's viewpoints. Such interventions could lead to the design of improved online learning environments and more collaborative, scaled peer-interactions.

The previous applications of GCA focused on detecting roles in small group interactions and validating the methodology in the context of practical learning outcomes (Dowell et al., 2018 under-review). A particularly notable discovery out of that research suggested the difference in outcome measures across the social roles was not a product of learners being more prolific during interactions. Instead it appeared that simply participating a lot was far less important than the nature of that participation (as captured by the internal cohesion, responsivity, and social impact measures). That is, the quality of conversation, more than the quantity, appears to be the key element in the success for both groups and individuals in small group learning interactions. Our future research efforts will focus on further validating the clusters with human judgements and exploring the practical implications of the identified profiles with learning relevant processes, such as retention and performance in scaled learning environments.

References

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *Journal of the Learning Sciences*, *9*(4), 403–436.

Bernard, R. M., Abrami, P. C., Borokhovski, E., Wade, C. A., Tamim, R. M., Surkes, M. A., & Bethel, E. C. (2009). A Meta-Analysis of Three Types of Interaction Treatments in Distance Education. *Review of Educational Research*, *79*(3), 1243–1289.

Borokhovski, E., Tamim, R., Bernard, R. M., Abrami, P. C., & Sokolovskaya, A. (2012). Are contextual and designed student–student interaction treatments equally effective in distance education? *Distance Education*, *33*(3), 311–329.

Boroujeni, M. S., Hecking, T., Hoppe, H. U., & Dillenbourg, P. (2017). Dynamics of MOOC Discussion Forums. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 128–137). New York, NY, USA: ACM.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: brain, mind, experience, and school* (Exp Sub edition). Washington, D.C: National Academies Press.

Çakır, M. P., Zemel, A., & Stahl, G. (2009). The joint organization of interaction within a multimodal CSCL medium. *International Journal of Computer-Supported Collaborative Learning*, *4*(2), 115–149.

Chan, C. K. K. (2012). Co-regulation of learning in computer-supported collaborative learning environments: a discussion. *Metacognition and Learning*, *7*(1), 63–73.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6).

Chua, S. M., Tagg, C., Sharples, M., & Rienties, B. (2017). Discussion analytics: Identifying conversations and social learners in FutureLearn MOOCs. In L. Vigentini, W. Y. P. L, & M. L. Urrutia (Eds.), *7th International Learning Analytics and Knowledge Conference* (pp. 36–62). ACM.

Clark, H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.

Dillenbourg, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *Journal of the Learning Sciences*, *15*(1), 121–151.

Dowell, N. M., Nixon, T., & Graesser, A. C. (2018). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multi-party interactions. *[revise and Resubmit] Behavior Research Methods*. Retrieved from https://arxiv.org/abs/1801.03563

Dowell, N. M., Skrypnyk, O., Joksimović, S., Graesser, A. C., Dawson, S., Gašević, S., … Kovanović, V. (2015). Modeling learners' social centrality and performance through language and discourse. In C. Romero & M. Pechenizkiy (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 250–257). International Educational Data Mining Society.

Gorman, J. C., Cooke, N. J., & Kiekel, P. A. (2004). Dynamical perspectives on team cognition. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting Human Factors and Ergonomics Society. Meeting*, *48*(3), 673–677.

Gorman, J. C., Foltz, P. W., Kiekel, P. A., Martin, M. J., & Cooke, N. J. (2003). Evaluation of latent semantic analysis-based measures of team communications content. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting Human Factors and Ergonomics Society. Meeting*, *47*, 424–428.

Graesser, A. C., Dowell, N. M., & Clewley, D. (2017). Assessing Collaborative Problem Solving Through Conversational Agents. In A. A. Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative Assessment of Collaboration* (pp. 65–80). Springer International Publishing.

Han, J., Pei, J., & Kamber, M. (Eds.). (2011). *Data mining: Concepts and techniques*. Boston, MA: Elsevier.

Hecking, T., Chounta, I.-A., & Hoppe, H. U. (2016). Investigating Social and Semantic User Roles in MOOC Discussion Forums. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 198–207). New York, NY, USA: ACM.

Hecking, T., Chounta, I. A., & Ulrich Hoppe, H. (2017). Role Modelling in MOOC Discussion Forums. *Journal of Learning Analytics*, *4*(1), 85–116.

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 37–56). Springer Netherlands.

Hmelo-Silver, C. E., & Barrows, H. S. (2008). Facilitating collaborative knowledge building. *Cognition and Instruction*, *26*(1), 48–94.

Hrastinski, S. (2008). What is online learner participation? A literature review. *Computers & Education*, *51*(4), 1755–1765.

Joksimović, S. a., Gašević, D. a., Loughin, T. M. c., Kovanović, V. b., & Hatala, M. d. (2015). Learning at distance: Effects of interaction traces on academic achievement. *Computers and Education*, *87*, 204–217.

Joksimović, S., Dowell, N., Poquet, O., Kovanović, V., Gašević, D., Dawson, S., & Graesser, A. C. (2018). Exploring development of social capital in a CMOOC through language and discourse. *The Internet and Higher Education*, *36*(Supplement C), 54–64.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 170–179). New York, NY, USA: ACM.

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground coordination in joint activity. In *Organizational Simulation* (pp. 139–184). John Wiley & Sons, Inc.

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: A review of the research. *Computers in Human Behavior*, *19*(3), 335–353.

Malmberg, J., Järvelä, S., & Järvenoja, H. (2017). Capturing temporal and sequential patterns of self-, co-, and socially shared regulation in the context of collaborative learning. *Contemporary Educational Psychology*, *49*(Supplement C), 160–174.

Mesmer-Magnus, J. R., & Dechurch, L. A. (2009). Information sharing and team performance: a meta-analysis. *The Journal of Applied Psychology*, *94*(2), 535–546.

Oecd. (2013). *PISA 2015 collaborative problem solving framework*. Oxford, U.K: OECD Publishing.

Poquet, O., Dawson, S., & Dowell, N. (2017). How Effective is Your Facilitation?: Group-level Analytics of MOOC Forums. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 208–217). New York, NY, USA: ACM.

Price, E. F. (1981). Toward a taxonomy of given/new information. In P. Cole (Ed.), *Radical Pragmatics*. New York, NY: Academic Press.

Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, *4*(3), 239–257.

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem- solving. In C. E. O'Malley (Ed.), *Computer-supported collaborative learning* (67–97). Berlin: Springer-Verlag.

Stahl, G. (2017). Group practices: a new way of viewing CSCL. *International Journal of Computer-Supported Collaborative Learning*, *12*(1), 113–126.

Stahl, G., Koschmann, T., & Suthers, D. D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409–426). Cambridge: Cambridge University Press.

Stahl, G., & Rosé, C. P. (2013). Theories of Team Cognition: Cross-Disciplinary Perspectives. In E. Salas, S. M. Fiore, & M. P. Letsky (Eds.), *Theories of team cognition: Cross-disciplinary perspectives* (pp. 111–134). New York, NY: Routledge.

Suthers, D. D., Dwyer, N., Medina, R., & Vatrapu, R. (2010). A framework for conceptualizing, representing, and analyzing distributed interaction. *International Journal of Computer-Supported Collaborative Learning*, *5*(1), 5–42.

Wise, A. F., Speer, J., Marbouti, F., & Hsiao, Y.-T. (2012). Broadening the notion of participation in online discussions: examining patterns in learners' online listening behaviors. *Instructnl. Science*, *41*, 323–343.

Yang, D., Sinha, T., Adamson, D., & Rose, C. P. (2013). Turn on, Tune in, Drop out": Anticipating student dropouts in Massive Open Online Courses. In *Proceedings of the 2013 NIPS data-driven education workshop* (pp. 1–8). NIPS Foundation.

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: theoretical perspectives*. Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.

Acknowledgements

This research was supported in part by the Michigan Institute for Data Science (MIDAS).