The Binary Replicate Test: Determining the Sensitivity of CSCL Models to Coding Error

Brendan R. Eagan, University of Wisconsin–Madison, beagan@wisc.edu
Zachari Swiecki, University of Wisconsin–Madison, swiecki@wisc.edu
Cayley Farrell, University of Wisconsin–Madison, cefarrell@wisc.edu
David Williamson Shaffer, University of Wisconsin–Madison and Aalborg University, Copenhagen, dws@education.wisc.edu

Abstract: The process of labeling, categorizing, or otherwise annotating data—or *coding* in the computer-supported collaborative learning (CSCL) literature—is a fundamental process in CSCL research. It is the process by which researchers identify salient properties about segments of CSCL data: what they are, what they contain, or what they mean. Coding, like all processes in research, is subject to error. To reduce the potential impact of coding error, CSCL researchers typically measure inter-rater reliability (IRR). However, there is no extant method to determine what level of IRR would invalidate a CSCL result or model. One way of assessing the potential impact of such inaccuracies is by conducting *sensitivity analyses,* which measure the *level of error* that would need to be present in the data to invalidate a given inference. This paper introduces a new method for conducting sensitivity analyses in CSCL: the Binary Replicate Test.

Introduction

Inaccuracies in measuring collaborative learning in *computer-supported collaborative learning* (CSCL) environments can result in poor learning outcomes, inappropriate or ineffective pedagogical responses, misguided design decisions, or misallocation of scarce resources. In order to effectively *control* for the impact of measurement error, researchers need to be able to *measure* it. One general approach to assessing the impact of inaccuracies in data is through *sensitivity analyses,* which measure the *level of error* that would need to be present in the data to invalidate a given inference: that is, the extent to which the data could be altered until the original result becomes invalid (Frank & Min, 2007) [(1)]. Here, we develop and test a method to apply this general strategy to the specific case of *coding error*.

In CSCL research, one of the largest sources of error involves *coding*. At its most basic level, coding is the process used to classify data: to assert that it should be labeled with a given construct or thought of as containing a given property (Chi, 1997; Shaffer, 2017). There are, of course, many sources of coding error. For example, hand coding performed by qualitative researchers is subject to individual bias, misinterpretation, or even just human fatigue—ask any researcher rating their 250th excerpt of data! Similarly, automated coding processes introduce some level of error, no matter how good the algorithm or approach. To reduce the potential impact of coding error, CSCL researchers typically measure inter-rater reliability (IRR) (Eagan, et. al. 2017; Hammer & Berland, 2014). While the common method of establishing IRR in CSCL research requires some methodological refinement (Eagan, et. al. 2017; Hammer & Berland, 2014; Shaffer, 2017), even when applied correctly, there is no extant method to determine what level of IRR (and thus what level of coding error) would invalidate a result or model.

In this study, we propose and provide an example of a method for conducting a Monte Carlo rejective test, a form of null hypothesis significance test, that empirically measures the impact of coding error on the results of any CSCL analysis. We consider the specific case of *binary coding*—the assertion of the presence or absence of a given construct for some observation—because this is a commonly-used approach to coding in CSCL research, and we refer to the proposed technique as a Binary Replicate Test (BRT). And in this initial paper on the BRT method, we apply the BRT approach using one specific modeling tool and one specific set of CSCL data. Our results suggest, however, that the BRT method has the potential to empirically determine thresholds for IRR measurement in CSCL analyses more generally.

Theory

Coding error

Binary coding is the process of assigning one of two possible values (typically 1 or 0) to pieces of data, reflecting the presence or absence of some property of interest in the data: for example, whether a student is asking a question in a particular turn or talk (1) or not (0). It follows that there are two types of error in binary coding. *Type I errors*

(false positives) are when an observation *does not* contain a given construct but is *coded* for the construct—for example, a turn of talk coded with a 1, indicating that the student is asking a question, is a Type I error if the student is, in fact, not asking a question. *Type II errors* (false negatives) are when an observation *does* contain a construct but is *not coded* for the construct—for example, if a turn of talk is coded with a 0, indicating the student was not asking a question, when she actually was.

In most CSCL settings, however, determining Type I and Type II errors is complex, because there is not necessarily a clear ground truth: two observers may see the same event in the data and code it differently. Researchers typically use measures of IRR, such as Cohen's kappa, to determine the overall rate of agreement—that is, some combined level of Type I and Type II errors between two raters, taking one rater as the standard to which the second rater is compared. The assumption—often implicit—in using IRR in this way is that the ground truth lies somewhere between the two sequences of 1s and 0s that each coder assigns to the data [(2)].

Put another way, the IRR between two raters establishes an *upper bound* on the distance that either rater can be from ground truth—assuming that both raters are conscientious and understand the concept being coded, of course. Therefore, to assess the sensitivity of a given model to coding error, we need to determine the minimum level of an IRR statistic for which the model remains valid: that is, some level at which we can change 1s to 0s and 0s to 1s without altering the result of a given analysis conducted on the coded data.

In the case of IRR, guidance for establishing such thresholds is scarce. Many statistics have no commonly accepted threshold for what constitutes an acceptable level of agreement. For example, researches refer to levels of Precision, Recall, or F statistics as being "high" without ever defining what constitutes a "high" level for that statistic (Schwarm, & Ostendorf, 2005). For other metrics, there are commonly used levels of acceptable agreement that have no statistical or inferential foundation. For example, Viera & Garrett (2005) suggested that "substantial agreement" for Cohen's kappa is above 0.61, and that has been used as a standard in many fields (Andrews, Leonard, Colgrove, & Kalinowski, 2011; Goh et. al., 2008; Lasorsa, Lewis, & Holton, 2012), despite the fact that neither they, nor Cohen, provided a justification for that choice.

Thresholds of significance

This use of IRR, is, of course, a form of sensitivity analysis. Sensitivity analyses in general quantify the conditions under which the validity of a result could be called into question by establishing a *threshold*: the point at which a particular level of error would invalidate an inference based on that result. Such analyses are often conducted using *Monte Carlo* (MC) studies (Eagan et. al. 2017; Harwell, 1992). MC studies assess the performance and reliability of statistical tests in educational and psychological research by creating an *empirical sampling distribution*. That is, they create a large number of simulated datasets and calculate a test statistic for each one. The performance of the test can then be evaluated under different assumptions about the properties of the population from which the samples are drawn, or based on variations on the sampling procedure itself.

For example, Eagan and colleagues (2017) used MC studies to demonstrate the unacceptably high Type I error rates associated with using commonly applied methods for IRR measurement in CSCL, learning analytics, and related fields. Similarly, Harwell, Rubinstein, Hayes & Olds (1992) used MC studies to show that the Welch test, a variant of ANOVA, has inflated Type I error rates and lower power when used with non-normal distributions than with normal distributions. In the case of coding error, researchers can create simulated code sets by introducing different levels of error to an *actual* dataset to examine the performance of a model under different error conditions.

In this paper we conduct a sensitivity analysis on the results of one analytic technique used in CSCL and learning analytics, epistemic network analysis, but the BRT method is agnostic to the specific analytic approach.

Model significance

Analytic models can be used to detect differences between groups, such as contrasting treatment and control, comparing different teams, highlighting instructor differences, and so on. In addition, analytic models can be used to assess learners at an individual level, which is useful for, among other things, providing feedback, grading, and selecting interventions. As a result, researchers need methods for establishing the accuracy of a learning analytic model for assessing both group differences and individual learners.

MC rejective methods

Researchers can establish the robustness of analytic results using *rejective methods*. Using a criterion for rejecting a null hypothesis, such as employing p-value thresholds in hypothesis testing, is an example of a rejective method. Classically, when researchers observe a p-value under a critical value—often $\alpha < 0.05$—they reject the null hypothesis under the assumption that they would be doing so incorrectly 5% of the time if the null hypothesis were indeed true. While many rejective methods rely on a theoretical model, such as parametric tests using p-

values, others rely on empirical simulations to establish a criterion for rejecting a null hypothesis. For example, the MC rejective method Shaffer's rho generates an empirical null hypothesis distribution used to determine the likelihood of seeing a specific IRR measurement given the null hypothesis—that the actual IRR was below a threshold of interest (Shaffer, 2017). Similarly, an MC rejective method can be used to test a given analytical model's robustness to coding error when measuring group or individual differences.

ENA

One technique that has been frequently used to analyze coded data produced in CSCL environments is *epistemic network analysis* (ENA) (see Shaffer, 2017, and references). To do so, ENA takes interaction data coded for elements of complex thinking, collaboration, learning, or any other phenomena of interest and creates weighted network models of the connections between those elements for individuals or groups.

As with any learning analytic technique applied to coded data, errors in coding can significantly affect statistical results and interpretations. In what follows, we present a study on the impact of coding error on one particular ENA result. However, we emphasize that our goal is less to show a method for conducting sensitivity analyses on one particular modeling technique than it is to present this as an example of a method that can be used to conduct sensitivity analyses on any CSCL or learning analytic result generated from coded data.

Specifically, we ask: ***How much coding error can be introduced to a dataset before we are no longer confident in the statistical significance (at the group level) and accuracy (at the individual level) of a given CSCL result?*** We address this research question by using a MC rejective method to examine the impact of coding error on one validated ENA result.

Methods

Gold-Standard ENA model

General approach. Our general approach to developing and testing a methodology for conducting sensitivity analyses was to create a "Gold-Standard" result: that is, an original, statistically significant result using coded data whose sensitivity we wanted to test. To do so we created an ENA model using a well-studied dataset (Arastoopour et. al. 2016; Chesler et al., 2015). We then evaluated the effects of making perturbations to the coded dataset upon which the Gold-Standard result was based. Specifically, we measured the sensitivity of the ENA model to coding error by randomly introducing coding error into the dataset used to create the Gold-Standard model, and then looked to see the level of error at which the original result was no longer statistically significant or accurate at the individual level.

Data source. To create our Gold-Standard result, we analyzed the discourse of students in the engineering virtual internship *RescuShell* (Chesler et al., 2015). In *RescuShell*, students work in project teams to design robotic exoskeletons for use by rescue workers in disaster situations. The Gold-Standard model has two conditions: (a) students in the first condition (relative novices, hereafter referred to as Novices) were participating in an engineering virtual internship for the first time; (b) students in the second condition (relative experts, hereafter referred to as Experts) had previously participated in a different engineering virtual internship.

Model. Previous research (Arastoopour et. al. 2016) found differences in the discourse patterns of Novices and Experts that were both statistically and interpretively significant (that is, meaningful). Chesler and colleagues (2015) found that Novices made connections mostly among basic skills and knowledge while Experts made additional connections with epistemological elements of engineering which are indicative of complex problem solving in engineering. To create the Gold-Standard model we replicated this previous research by applying Epistemic Network Analysis (ENA; Shaffer, 2017; Shaffer, Collier, & Ruis, 2016; Shaffer & Ruis, 2017) to data from *RescuShell* using the ENA Web Tool (version 1.5.2) (Marquart, Hinojosa, Swiecki, Eagan, & Shaffer, 2018).

The ENA technique has been described in detail elsewhere (Shaffer, 2017; Shaffer, Collier, & Ruis, 2016), including in previous papers presented at CSCL and ICLS (Collier, Ruis, & Shaffer, 2016; Csanadi et. al., 2017; Siebert-Evenstone et al. 2016; Swiecki, & Shaffer 2018). Briefly, ENA models connections in discourse as a network graph where nodes correspond to the codes, and edges reflect the relative frequency of co-occurrence, or connection, between two codes. These networks can be further compared both quantitatively and qualitatively to analyze and compare different discourse patterns in CSCL environments.

We are not providing the details of the data and coding scheme here, both because they are available in Chesler et al. (2015) and because the specific codes and details of the model are not the primary issue in testing the sensitivity to of the model to coding error. However, briefly: for the Gold-Standard result, we defined the units of analysis as all lines of data associated each individual student, and used a moving window to construct a network model for each line in the data, showing how codes in give line are connected to codes that occur within

the recent temporal context (Siebert-Evenstone et al., 2017), defined as 7 lines (each line plus the 6 previous lines) within a given conversation. The resulting networks were aggregated for all lines for each unit of analysis in the model. In this model, we aggregated networks using a binary summation in which the networks for a given line reflect the presence or absence of the co-occurrence of each pair of codes within the recent temporal context of the line. Our model included the following codes: CLIENT AND CONSUMER REQUESTS, COLLABORATION, DATA, DESIGN REASONING, PERFORMANCE PARAMETERS, and TECHNICAL CONSTRAINTS. We defined conversations as all lines of data associated with a single group in a single activity in each condition, Experts or Novices. We projected the resulting networks using a *means rotation*: a dimensional reduction technique that places the greatest differences between conditions on the x-axis in the ENA space.

In the resulting network graphs, the nodes were placed in ENA space so as to minimize the sum of the distances for each student between the point representing the student's network in the projected space and the centroid of the student's network graph. The result is two coordinated representations for each student: (1) a plotted point, which represents the location of that student's network in the low-dimensional projected space, and (2) a weighted network graph, which shows the pattern of connections the student made that led to his or her location in the ENA space. Because of this co-registration, the positions of the network graph nodes—and the connections they define—can be used to interpret the dimensions of the projected space and explain the positions of the plotted points in that space. It is also possible to interpret the differences between two groups by constructing their mean networks and subtracting them. The resulting *difference network* shows the connections that are stronger for each group compared to the other.

Our model had co-registration correlations of 0.99 (Pearson) and 0.99 (Spearman) for the first dimension and co-registration correlations of 0.99 (Pearson) and 0.99 (Spearman) for the second dimension. These measures indicate that there is a strong goodness of fit between the visualization and the original model, and thus indicates that the model has interpretive validity: we can use the network graphs to interpret the statistical results.

Test result. To test for discourse pattern differences between Novices and Experts, we applied a two-sample t-test assuming unequal variance to the mean location of points in the projected ENA space for each condition. Figure 1 shows the plotted points (dots), means (squares), and confidence intervals (dashed boxes) for the Novices (in red) and the Experts (in blue).

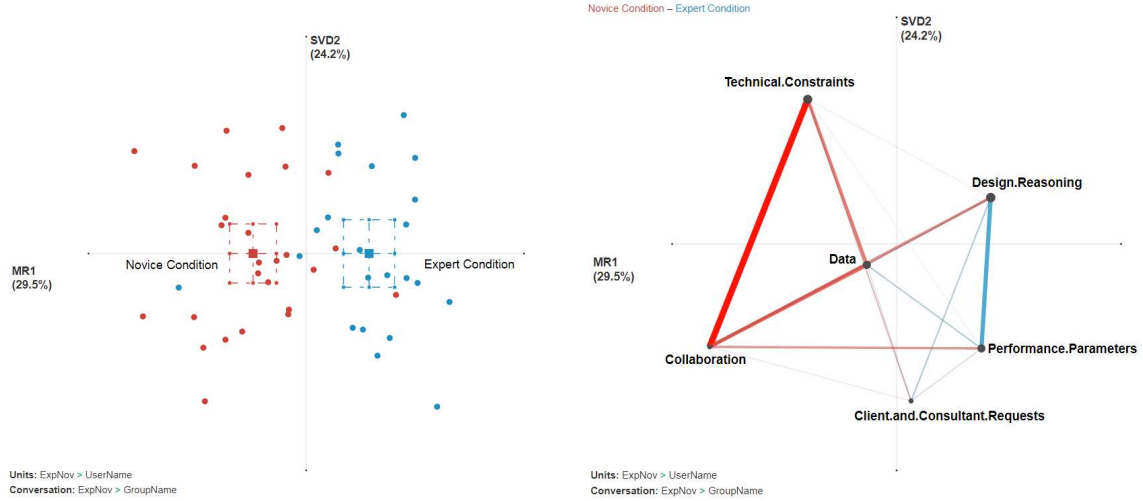

Figure 1. Plotted points (left) and difference network (right) for the Gold-Standard ENA result.

Figure 1 also shows the difference network that can be used to interpret statistical differences between conditions. Where the difference network is red, the Novices had stronger connections, and where the network is blue, Experts made stronger connections.

Along the *x* axis, a two sample t-test assuming unequal variance showed the discourse pattern of Novices (mean = –0.37, SD = 0.44, N = 27) was statistically significantly different at the alpha=0.05 level from those of Experts (mean = 0.36, SD = 0.49, N = 28; $t(52.78) = -5.76$, $p < 0.01$, Cohen's $d = 1.55$) (see Figure 1).

This statistically significant result (Figure 1, left) can be interpreted using the subtracted network graph on the right side of Figure 1. The Novices made more connections between TECHNICAL CONSTRAINTS and COLLABORATION, TECHNICAL CONSTRAINTS and DATA, COLLABORATION and DATA and COLLABORATION and PERFORMANCE PARAMETERS when compared with the Experts. The Experts made more connections between

PERFORMANCE PARAMETERS, and DESIGN REASONING than Novices. While Experts did make connections to COLLABORATION, more of their discourse focused on the relationships between PERFORMANCE PARAMETERS, DATA, and DESIGN REASONING; in other words, they were more focused on the core elements of engineering design. This result aligns with previous expert-novice comparisons from engineering virtual internships (Arastoopour et. al. 2016; Chesler et al., 2015).

Our Gold-Standard model thus includes two components that we tested in this analysis: (a) the statistically significant difference between Experts and Novices in that model, and (b) the specific locations of individual students relative to the model.

Monte Carlo approach

For our BRT analysis, we created 5,000 simulated datasets at each of the different levels of coding error we examined (5%–10%, inclusive, in 1% increments) [(3)]. To create each individual simulated dataset, at each level of coding error, we randomly introduced error at that level to the Gold-Standard dataset. All other aspects of the subsequent ENA model creation process where the same [(4)].

In the original data, codes are represented as a 1 or 0, indicating the presence or absence of a particular construct, for each turn of talk as students collaborated in groups of four [(5)]. We introduced coding error as follows: (a) for each code, we randomly selected a percentage of lines equal to the percent error rate; (b) we changed the coding of each line chosen, changing its value either from 1 to 0 or 0 to 1; and (c) we repeated this process for each code in the dataset.

Sensitivity analyses

RQ1: How much coding error can be introduced to a dataset before more that 5% of the 5,000 ENA simulated models do not find a statistically significant difference between the Novice and Expert conditions?

To answer research question 1, at each level of error: (1) we produced an ENA model for each of the 5,000 simulated datasets; (2) for each of these models we conducted the same t-test used to analyze our Gold-Standard model; (3) we calculated the number of models where the t-test *did not* show a statistically significant difference between Novices and Experts; (4) we created a 95% confidence interval on the estimate of the number of models that were not statistically significant. This tested the null hypothesis at each level of error: for example, at 6% introduced error, the null hypothesis was H_0: The difference between Novices and Experts is not significant at 6% coding error. If the confidence interval for the number of non-significant tests was below 5%, we rejected the null hypothesis and concluded that, with $\alpha = 0.05$, the group difference in the Gold-Standard result was robust to that level of introduced error.

RQ2: How much coding error can be introduced to a dataset before the average correlation between the plotted points from the 5,000 simulated ENA models and the plotted points from the Gold-Standard model falls below 0.80?

To assess the extent to which coding error affects the accuracy of assessing individual students, at each level of error: (1) we calculated the correlation between plotted points in the Gold-Standard model—which represent the results of the ENA analysis for each student—and the plotted points in each simulated ENA model; (2) we computed the average correlation between plotted points in the Gold Standard model and each simulated dataset; (3) we created 95% confidence intervals for the average correlations. This tested the null hypothesis at each level of error: for example, at 6% introduced error, the null hypothesis was H_0: The positions of students in is not different from Gold-Standard model with 6% coding error. There is no universally agreed upon threshold for significant correlations; some researchers propose 0.70–0.90 as a highly positive correlation (Hinkle, Wiersma, & Jurs, 2003), while others consider correlations of 0.80–1.00 very strong (Evans, 1996). We choose a conservative threshold of correlations greater than or equal to 0.80. If the lower bound of the 95% confidence interval was greater than 0.80, we rejected the null hypothesis and concluded that, with $\alpha = 0.05$, the Gold-Standard result preserved the positions of students in the model at that level of introduced error.

Results

The left side of Figure 2 shows the percentage of tests in the sensitivity analysis that were not statistically significant at each level of introduced coding error we tested. The upper bound of the confidence intervals is below 0.05 at error rates up to 8%; however the upper bound of the confidence interval is above 0.05 at a 9% error rate. Thus the group difference in the Gold-Standard result is robust up to an 8% rate of coding error.

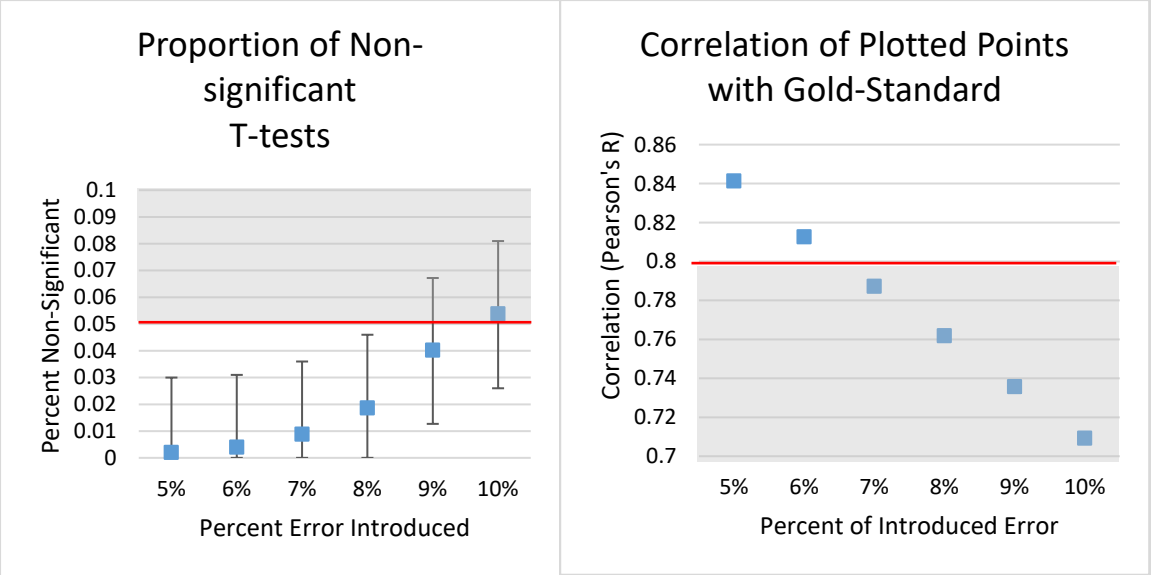

Figure 2. BRT results group difference sensitivity (left), individual level accuracy (right). For each graph, if the 95% confidence interval is at all contained by the shaded area, we fail to reject the corresponding null hypothesis.

The right side of Figure 2 shows the average correlation between the plotted points of the models with introduced error and the plotted points from the Gold-Standard model at the different levels of coding error. Because of the large number of simulated data sets (5000) at each level of error, the 95% confidence intervals are too small to be seen on the graph. The upper and lower bound of the confidence interval for error rates up to 6% are above 0.80; however; at a 7% error rate 7%, the confidence interval falls below 0.80. Thus, accuracy of individual measurements of the Gold-Standard model is robust to 6% error.

Discussion

Our results thus show that difference between the means of two groups of students in the result we tested was robust to 8% error, but the accuracy of the model for individual students was only robust to 6% error. We thus conclude that the BRT suggests the result overall was robust up to 6% error.

While this analysis was conducted with only one set of CSCL data and using only one CSCL modeling approach, we believe that our analysis provides an example of how the BRT method for conducting sensitivity analyses can be used in CSCL research. In particular, we argue that the BRT approach can be used to conduct a sensitivity analysis on any model that depends on binary-coded data. That is, BRT provides a method that could be used to measure how robust CSCL results are to coding error.

Measuring the sensitivity of CSCL results to coding error in this way would provide a useful tool for establishing confidence in CSCL models. For example, if there are two conflicting CSCL results and the first was robust up to 1% coding error and the other was robust up to 8% coding error, we might have more confidence in the more robust result. More important, however, the BRT approach potentially provides a method for determining what level of coding reliability would be required to draw conclusions from a model—that is, a way to determine empirically the appropriate IRR threshold for a given result.

There is, of course, more work that will need to be done to make it possible to use BRT in this way. In particular, this study has several clear limitations. First, while this study reports on the robustness of (a) discriminating between two groups and (b) the accuracy of individual assessment, our analysis did not investigate how coding error affected the *interpretation* of the model. While it is beyond the scope of this paper, we have conducted BRT analyses for the interpretation of this model. Briefly, ENA models are interpreted using the positions of the nodes in the network graphs. In a future paper we will show that the average correlation between ENA node positions from the Gold-Standard model and the node positions from each simulated ENA model was above our threshold of 0.80 up to a 10% error rate. This suggests that that model interpretation was robust to 10% introduced error, which aligns with previous methodological analysis of ENA (Ruis et. al. 2018).

A second limitation of this work is that we did not systematically account for properties of the codes, such as code frequency, which may impact sensitivity. Similarly, when researchers code data, errors may be systematic, rather than random. In future work we plan to account for these issues in the BRT method by

independently specifying rates of Type I and Type II coding errors. Third, we used only one model to explore the BRT method, and we deliberately chose a result with a large effect size. In future work we plan to test the BRT method using other results with a range of effect sizes, as well as other modeling techniques used in CSCL research. Finally, we used percent agreement as our IRR metric, which is a problematic measure (Cohen, 1960; Shaffer, 2017). In future work, we plan to base our error rates on more precise measures of IRR, including Cohen's kappa, precision, recall, and F statistics.

Despite these clear limitations, the results presented here suggest that the BRT method is a promising method for empirically deriving IRR thresholds.

Endnotes

(1) In statistics, sensitivity analyses often focus on bias, or systematic error, but for the purposes of this study we are focusing on the effects of random error. If a method is shown to be negatively impacted by random error, it will most likely be even more negatively impacted by systematic bias. It follows that focusing on random error is a more conservative approach for this type of sensitivity analysis.

(2) Note that this approach does not require that an actual ground truth exists; only that we can say something about likely rates of disagreement that exist about a coding process that tries to locate it. This is an important conceptual issue, but beyond the scope of the current paper.

(3) We also created simulated datasets at coding error levels from 1% - 5% and above 10% but in this study we are only reporting where the sensitivity analysis demonstrated a change statistically significant impact of coding error.

(4) We projected each ENA model into our Gold-Standard ENA space by using the same Eigen vectors for the ENA rotation matrix in the dimensional reduction step of the ENA process. This resulted in the node positions for all of the ENA models in the sensitivity analysis to be identical. The only thing that varied was the coding and resulting connection strengths and plotted point locations for each unit in each ENA space. In other words, the creation of each of the ENA sets based on datasets containing introduced coding error was identical.

(5) ENA can handle non-binary or weighted data, but this study focused on binary ENA models.

References

Arastoopour, G., Shaffer, D.W., Swiecki, Z., Ruis, A.R., & Chesler, N.C. (2016). Teaching and assessing engineering design thinking with virtual internships and epistemic network analysis. *International Journal of Engineering Education*, *32*(3B), 1492–1501.

Andrews, T. M., Leonard, M. J., Colgrove, C. A., & Kalinowski, S. T. (2011). Active learning not associated with student learning in a random sample of college biology courses. *CBE—Life Sciences Education*, *10*(4), 394-405.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The journal of the learning sciences*, *6*(3), 271-315.

Chesler, N. C., Ruis, A. R., Collier, W., Swiecki, Z., Arastoopour, G., & Shaffer, D. W. (2015). A novel paradigm for engineering education: Virtual internships with individualized mentoring and assessment of engineering thinking. *Journal of biomechanical engineering*, *137*(2), 024701.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1): 37–46.

Collier, W., Ruis, A.R., & Shaffer, D.W. (2016). Local versus global connection making in discourse. *Paper presented at the 12th International Conference of the Learning Sciences*. Singapore.

Csanadi, A., Eagan, B., Shaffer, D. W., Kollar, I., & Fischer, F. (2017). Collaborative and individual scientific reasoning of pre-service teachers: New insights through epistemic network analysis (ENA). In B. K. Smith, M. Borge, E. Mercier, & K. Y. Lim (Eds.), *Making a difference: Prioritizing equity and access in CSCL: 12th International Conference on Computer-Supported Collaborative Learning* (Vol. I, pp. 215–222).

Eagan, B., Rogers, B., Serlin, R., Ruis, A. R., Arastoopour Irgens, G., Shaffer, D. W. (2017). "Can We Rely on Reliability? Testing the Assumptions of Inter-Rater Reliability," *Making a Difference: Prioritizing Equity and Access in CSCL: 12th International Conference on Computer-Supported Collaborative Learning*, eds. B. K. Smith, M. Borge, E. Mercier, & K. Y. Lim (2017), II:529–532.

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Brooks/Cole.

Frank, K., & Min, K. S. (2007). 10. Indices of Robustness for Sample Representation. *Sociological Methodology*, *37*(1), 349-392.

Goh, N. S., Desai, S. R., Veeraraghavan, S., Hansell, D. M., Copley, S. J., Maher, T. M., ... & Bozovic, G. (2008). Interstitial lung disease in systemic sclerosis: a simple staging system. American journal of respiratory and critical care medicine, 177(11), 1248-1254.

Harwell, M. R. (1992). Summarizing Monte Carlo Results in Methodological Research. *Journal of Educational Statistics*, *17*(4), 297–313.

Hinkle DE, Wiersma W, Jurs SG. Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin; 2003.

Marquart, C. L., Hinojosa, C., Swiecki, Z., Eagan, B., & Shaffer, D. W. (2018). Epistemic Network Analysis (Version 1.5.2) [Software]. Available from http://app.epistemicnetwork.org

Lasorsa, D. L., Lewis, S. C., & Holton, A. E. (2012). Normalizing Twitter: Journalism practice in an emerging communication space. *Journalism studies*, *13*(1), 19-36.

Ruis, A.R., Siebert-Evenstone, A.L., Pozen, R., Eagan, B., & Shaffer, D.W. (2018). A method for determining the extent of recent temporal context in analyses of complex, collaborative thinking. In Kay, J. & Luckin, R (Eds.) *Rethinking Learning in the Digital Age: Making the Learning Sciences Count, 13th International Conference of the Learning Sciences (ICLS)*, *III*, (pp. 1625–1626).

Schwarm, S. E., & Ostendorf, M. (2005, June). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 523-530). Association for Computational Linguistics.

Shaffer, D. W. (2017). Quantitative ethnography. Madison, WI: Cathcart Press.

Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics, 3*(3), 9–45.

Siebert-Evenstone, A.L., Arastoopour Irgens, G., Collier, W., Swiecki, Z., Ruis, A.R., & Shaffer, D.W. (2017). In search of conversational grain size: Modelling semantic structure using moving stanza windows. *Journal of Learning Analytics, 4*(3), 123–139.

Swiecki, Z. & Shaffer, D.W. (2018). Toward a taxonomy of team performance visualization tools. In Kay J. & Luckin, R (Eds.) *Rethinking Learning in the Digital Age: Making the Learning Sciences Count, 13th International Conference of the Learning Sciences (ICLS), III*, 144–151.

Viera, A.J., & Garrett, J.M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360-363.

Acknowledgements

This work was funded in part by the National Science Foundation (DRL-1661036, DRL-1713110), 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.