Controlling for Statistical Dependencies in CSCL Using General Estimating Equations

E. Michael Nussbaum, Gwen C. Marchand, University of Nevada, Las Vegas
4505 Maryland Parkway, Box 453003, Las Vegas, NV 89154-3003
nussbaum@unlv.nevada.edu, Gwen.Marchand@unlv.edu

Abstract: Statistically analyzing small-group discourse in CSCL requires controlling for statistical dependencies among group members that arise from the fact that group members influence one another's behaviors. Although some researchers in the learning sciences have addressed this problem by using multilevel modeling, that approach requires large group sizes. This poster presents an alternative approach, known as General Estimating Equations (GEE), which is more suitable when small groups such as dyads or triads are analyzed.

The Problem

Computer-supported collaborative learning (CSCL) by definition involves group tasks and group learning. CSCL research often involves experimental and quasi-experimental research that employs statistical hypothesis testing to draw conclusions about the effectiveness of different instructional designs on students' discursive behaviors. The problem addressed in this poster is that student behaviors are not statistically independent from one another, which violates basic, underlying statistical assumptions of many analytic methods.

By way of illustration, suppose students in a study (N = 150) are randomly grouped into triads and asked to engage in a synchronous discussion about some topic. Furthermore, half the triads receive some sort of intervention to improve the quality of the arguments made during CSCL discussion (Weinberger, Stegmann, & Fischer, 2010), such as being given information on group members' prior opinions (Buder & Bodemer, 2008). The other half of the triads do not receive the intervention. The number of counterarguments and rebuttals generated in the discussions is found to be higher in the first condition than in the second, but is the difference statistically significant?

A t-test could be performed, using the formula for the standard error: Sqrt($2\sigma^2/N$). What in this case is the effective N? It cannot be all 150 students, because there are not 150 statistically independent data points; the members of each triad affect one another. If one student offers a counterargument, her partner might do so as well for purposes of refutation. Furthermore, students also often imitate one another's discourse moves (Anderson et al., 2001). As a result, the chance of obtaining relatively more extreme observations increase because one extreme observation begets another. If we treat the observations as statistically independent, we will underestimate this probability, increasing the chance of Type I errors.

On the other hand, we could treat the triads as the unit of analysis, but doing so reduces statistical power, because the effective N is now only 50 rather than 150. This problem has been typically addressed by using a multilevel model (see Cress, 2008), which fits a separate regression line for each small group. However, Hox (2010) recommends that there should be a minimum of 20 students in each group; otherwise, the regression estimates will not be reliable. Small-groups, such as triads, clearly do not meet the sample size criterion. As a result, researchers must either use multi-level model inappropriately or use groups as the unit of analysis. In this paper, we propose a solution that has rarely been used in CSCL and other learning sciences research. The solution involves a new application of an existing methodology.

The Proposed Solution: General Estimating Equations (GEE)

The GEE methodology (Hardin & Hilbe, 2003) also attempts to adjust for statistical dependence in clusters of observations but unlike multilevel modeling, does not require the clusters to be large. It therefore is a more useful methodology for analyzing behavior in small groups, as long as there are a sufficient number of clusters.

GEE models statistical dependence by estimating to what degree student behaviors are correlated. For example, to what extent are the number of counterarguments generated by one person in a triad correlated with the number of counterarguments generated by another? Fifty dyads would provide 50 data points. Unlike multilevel modeling, the GEE methodology does not attempt to estimate regression lines for each group.

Technical Details

The procedure begins by estimating a working correlation matrix. With dyads, only one correlation needs to be estimated, but if there three or more members in a group, a matrix would be estimated. In the case of three members, the working correlation matrix would be:

$$\mathbf{R} = \begin{array}{c|cccc} . & p1 & p2 & p3 \\ p1 & 1 & \rho_1 & \rho_3 \\ p2 & \rho_1 & 1 & \rho_2 \\ p3 & \rho_3 & \rho_2 & 1 \end{array},$$

where *p1* is person #1, *p2* is person #2, etc. Three correlation parameters would need to be estimated, but we could make a simplifying assumption that all three parameters are equal to one another (this option is known as using an "exchangeable" structure). This is not a mandatory assumption, and one can test which type of correlation matrix (exchangeable, unstructured, etc.) best fits one's data. One can also assume an autoregressive structure if the statistical dependence is associated with repeated measurements. If there is doubt about which correlational structure is correct, a "robust" standard error can be estimated that is less sensitive to choice of the correlation structure.

The second step is to use one's statistical model to generate predicted values for each individual (i = 1 …N) on the dependent variable. From the predicted means one can calculate associated variances based on generalized linear models (for example, in analyzing counts using Poisson regression, the variance equals the mean). If there are three members in each group, a 3x3 diagonal variance matrix is generated for each of the j groups, $(\mathbf{A}_j)^{1/2}$, and this matrix is multiplied by the working correlation matrix $\mathbf{R}$ (and $(\mathbf{A}_j)^{1/2}$ again) to produce a 3x3 variance-covariance matrix, $\mathbf{V}_j$. The third step is to calculate a p x p variance-covariance matrix, where p is the number of parameters in the overall regression equation. For example, if there is one predictor, then there would be two parameters (an intercept and slope), and so the variance-covariance parameter matrix would be 2x2. It is calculated using the following formula: $(\mathbf{Cov})_j = \left(\mathbf{D}_j' \mathbf{V}_j \mathbf{D}_j\right)^{-1}$, where $\mathbf{D}_j = \mathbf{A}_j \mathbf{X}_j$. The matrices for the different groups are then summed, and the final matrix used to derive standard errors or to update parameter estimates. Further details and a more detailed example can be found in Nussbaum (in press).

The GEE methodology is currently available in various statistical packages, such as Stata, SAS, MATLAB, and R. It is also currently available in SPSS, but only for time-series data.

Conclusions

The GEE methodology can be used to control for statistical dependencies among group members when analyzing data for individuals engaged in collaborative learning. The methodology does not require the groups to be large, but there should be a large number of groups (at least 30). This stands in contrast to multilevel modeling, which requires fewer but bigger groups. The GEE methodology is currently underutilized in CSCL and other learning sciences research (but for exception, see Lin et al., 2012). This poster will hopefully raise awareness on the merits of this approach.

References

Anderson, R. C., Nguyen-Jahiel, K., McNurlen, B., Archodidou, A., Kim, S., Reznitskaya, A., Tillmanns, M., & Gilbert, L. (2001). The snowball phenomenon: Spreading ways of talking and ways of thinking across groups of children. *Cognition and Instruction, 19*(1), 1-46.

Buder, J., & Bodemer, D. (2008). Supporting controversial CSCL discussions with augmented group awareness tools. *International Journal of Computer Supported Collaborative Learning, 3*, 123-139.

Cress, U. (2008). The need for considering multilevel analysis in CSCL research—An appeal for the use of more advanced statistical methods. *International Journal of Computer-Supported Collaborative Learning, 3*, 69-84.

Hardin, J. W., & Hilbe, J. M. (2003). *Generalized estimating equations*. Boca Raton, FL: Chapman & Hall.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Hoboken: Taylor & Francis.

Lin, T.-J., Anderson, R. C., Hummel, J. E., Jadallah, M., Miller, B. W., Nguyen-Jahiel, K., Morris, J. A., Kuo, L.-J., Kim, I.-H., Wu, X., & Dong, T. (2012). Children's use of analogy during collaborative reasoning. *Child Development, 83*(4), 1429-1443.

Nussbaum, E. M. (in press). *Categorical and nonparametric data analysis: Choosing the best statistical technique.* NY: Taylor & Francis.

Weinberger, A., Stegman, K., & Fischer, F. (2010). Learning to argue online. Scripted groups surpass individual groups (unscripted groups do not). *Computers in Human Behavior, 26,* 506-515.