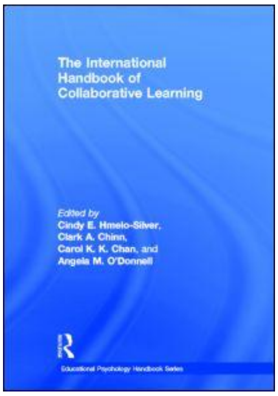

The International Handbook of Collaborative Learning

Cindy E. Hmelo-Silver, Clark A. Chinn, Carol K. K. Chan, Angela M. O'Donnell

Multilevel Analysis for the Analysis of Collaborative Learning

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203837290.ch6
Jeroen Janssen, Ulrike Cress, Gijsbert Erkens, Paul A. Kirschner
Published online on: 04 Feb 2013

6

MULTILEVEL ANALYSIS FOR THE ANALYSIS OF COLLABORATIVE LEARNING

JEROEN JANSSEN
Utrecht University, The Netherlands

ULRIKE CRESS
Knowledge Media Research Center, Tübingen, Germany

GIJSBERT ERKENS
Utrecht University, The Netherlands

PAUL A. KIRSCHNER
Open University, The Netherlands

INTRODUCTION

In a recent article "the widespread and increasing use" of collaborative learning (CL) has been called a "success story" (Cress & Hesse, chapter 5 this volume; Johnson & Johnson, 2009, p. 365). The study of CL has a long tradition, which has led to the publication of many studies which have examined the effects of CL on a range of dependent variables, such as student achievement (e.g., Nichols, 1996), time on task (e.g., Klein & Pridemore, 1992), motivation (e.g., Saleh, Lazonder, & De Jong, 2005), and use of metacognitive strategies (e.g., Mevarech & Kramarski, 2003). Johnson and Johnson (2009) identified over 1,200 studies comparing the relative effects of CL on, for example, individual learning. This line of research has become known as *effect-oriented research* because of its focus on the effects of CL (Dillenbourg, Baker, Blaye, & O'Malley, 1996; Van der Linden, Erkens, Schmidt, & Renshaw, 2000).

Traditionally, CL research has used well-known methods such as ANOVAs or (multiple) linear regression analysis to investigate how instructional interventions, student characteristics, and group characteristics affect the collaborative process and student learning. As we will show in this chapter, the datasets CL researchers collect often contain information about individual students (e.g., their motivation or their prior knowledge), about features of the groups (e.g., the type of task they work on or their composition), and sometimes even about features of the classroom (e.g., the teacher's CL teaching experience) or the school (e.g., school size or assignment to a treatment or control condition). These complex datasets create interdependencies between the different levels of analysis (i.e., student, group, classroom, or school). One way to deal with these interdependencies is to use multilevel analysis (MLA), also referred to as hierarchical linear modeling (HLM). The aim of this chapter is to demonstrate why and how MLA is an important technique for researchers who wish to investigate CL.

COMMON PROBLEMS ENCOUNTERED DURING COLLABORATIVE LEARNING RESEARCH

Consider a researcher who wants to conduct the following experiment. She wants to investigate how group member familiarity (i.e., how well do group members know each other before the start of the collaboration?) and group size (i.e., how many students constitute a group?) affect the communication and coordination process between group members and individual performance on a knowledge posttest. She therefore asks each group member to rate his or her familiarity with the other group members. Furthermore, she assigns group members to three-, four-, and five-person groups. She expects that as group member familiarity increases, the need for extensive coordination decreases. On the other hand, she also anticipates that when group size increases, the need for coordination increases. She suspects that in collaborative situations during which the need for coordination becomes very high, students' learning process may be hindered (cf., Kirschner, Paas, & Kirschner, 2009). She therefore hypothesizes that in familiar three-person groups, students will perform best on the posttest, while in unfamiliar five-person groups, students will perform worst on the posttest.

This example contains three features which require the researcher to use MLA if she wants to analyze her data appropriately: (a) Students work in groups and this creates a hierarchically nested dataset; (b) due to the mutual influence between group members, observations of the dependent variables are nonindependent; and (c) the research design contains variables measured at different levels (e.g., the dependent measures and group member familiarity are measured at the level of the student, while one of the independent variables, group size, is measured at the level of the group). Each of these problems will be discussed below.

Problem 1: Hierarchically Nested Datasets

During CL, students work in groups on an assigned group task. Often researchers are interested in the question how a certain instructional intervention (e.g., a training developed to promote helping behavior) affects the collaborative process and students' learning process. Other researchers might be interested in how contextual factors, such as the composition of the group, prior knowledge of group members, or features of the group task, affect the collaborative process. These situations lead to hierarchically nested datasets. Because groups consist of two or more students, individuals are nested within groups.

In many cases, researchers will encounter at least two levels. The lowest level is the individual, and the highest level is the group. The group is then the macro- or level-2 unit and the individual the micro- or level-1 unit (Hox, 2003; Snijders & Bosker, 1999). Researchers may encounter even more levels of analysis. A researcher might for example be interested in the effects of the teacher's experience with CL on the way his or her students cooperate. This researcher will have a dataset with three levels: students are nested within groups, while groups are nested within teachers' classrooms. Another researcher might be interested in the development of students' collaboration over time. This researcher would therefore collect data about students' collaboration on different occasions. This would also lead to a dataset with three levels: measurement occasions are nested within students and students are nested within groups (Kenny, Kashy, & Cook, 2006; Snijders & Bosker, 1999). In the example described at the beginning of this section, the researcher encounters two levels, because in her study students collaborate in small groups (students are nested in groups).

Whenever researchers encounter datasets with hierarchically nested data, MLA is needed to appropriately model this data structure because it can disentangle the effects of the different levels on the dependent variable(s) of interest (Snijders & Bosker, 1999; Strijbos & Fischer, 2007).

Problem 2: Nonindependence of Dependent Variables

Because students work in groups, the dependent variables CL researchers study are often nonindependent (Cress, 2008). This means that students within a group are likely more similar to each other than are persons from different groups (Kenny, Mannetti, Pierro, Livi, & Kashy, 2002). This nonindependence is caused by the mutual influence group members have on each other while they are interacting (Bonito, 2002; Kenny et al., 2006).

In the example described above, nonindependence may also occur. Group members have to discuss their strategies, formulate explanations, and coordinate their activities. In some groups this process will be more smooth and effective than in other groups. In the former case, it is likely that all group members will perform reasonably well on the posttest because they experienced a fruitful and effective collaborative experience. In the latter case however, it is likely that all group members will not perform well on the posttest because they experienced collaboration and coordination problems. In this example too, the observations of the dependent variable, posttest performance, may be nonindependent.

It should be noted that the mutual influence group members have on each other can not only cause students to behave more similarly, but may also cause students to behave differently from their group members. This is called the *boomerang effect* (Kenny et al., 2006). This may, for example, occur when roles are assigned to group members (cf., Schellens, Van Keer, & Valcke, 2005; Strijbos, Martens, Jochems, & Broers, 2004, 2007). If one group member, for example, is given the task of asking critical questions, while the other group member has to monitor task progress, this may lead to different behavior (e.g., the first student will ask many questions, but will display less metacognitive behavior, while the second student may display high levels of metacognitive behavior but may ask fewer questions).

Kenny et al. (2006) therefore make a distinction between *positive nonindependence* where group members influence each other in such a way that they behave more similarly and *negative nonindependence* where group members influence each other to behave differently. Thus, since group members influence each other, this will likely lead to either positive or negative nonindependence of the dependent variables that are being investigated. This nonindependence has to be dealt with during data analysis.

The degree of nonindependence can be estimated using the *intraclass correlation coefficient* (ICC, cf., Kashy & Kenny, 2000; Kenny et al., 2002). The ICC is computed by dividing the difference between the mean square between groups and the mean square within groups by the sum of the two mentioned mean squares. Values of the ICC can range from –1 to +1, and give an indication of how strongly members in a group resemble each other. A large positive value of the ICC indicates that within a group, group members tend to score similarly on the variable of interest. Conversely, a large negative value of the ICC indicates large differences between group members on the variable of interest. The ICC can also be thought of as the average correlation among all possible pairs within the groups (Kenny et al., 2006).

An alternative interpretation of the ICC is in terms of the *amount of variance that is accounted for by the group* (Kenny et al., 2006). When the ICC for satisfaction with the collaborative process is found to be .40, for example, this means that 40% of the variance in this measure is accounted for by the group and thus that 60% is accounted for by other (e.g., individual) factors.

High values of the ICC indicate a strong influence of the group on the individual. It is not uncommon for CL researchers to find ICC values of .30, .40, or higher (e.g., Janssen, Erkens, & Kanselaar, 2007; Janssen, Erkens, Kanselaar, & Jaspers, 2007; Paletz, Peng, Erez, & Maslach, 2004; Schepers, de Jong, Wetzels, & de Ruyter, 2008; Strijbos et al., 2004). Such a strong nonindependence needs to be addressed when conducting statistical analyses, because it distorts estimates of error variances, thus making standard errors, *p*-values, and confidence intervals invalid (Kenny, 1995; Kenny et al., 2006). MLA, unlike traditional statistical techniques such as *t*-tests and analysis of variance, is able to handle this distortion (Hox, 2003).

Problem 3: Different Units of Analysis

Some variables that are used during CL research are measured at the level of the student (e.g., gender, familiarity with other group members, prior knowledge), whereas other variables are measured at the level of the group (e.g., group size, group composition with respect to ability or gender). Strijbos et al. (2004), for example, studied the effects of roles on students' perceptions of the collaborative process. Their design contained two conditions. In the first condition groups were given different roles to fulfill during the collaboration, while in the second condition no roles were assigned to group members. In this case, Strijbos et al. used a dependent variable—perception of the collaborative process—that was observed at the individual level, and an independent variable—role or no role—that was observed at the group level. Such a dataset therefore contains variables with different units of analysis. Because traditional statistical techniques cannot properly take these different units of analysis into account, MLA is needed (Hox, 2003; Snijders & Bosker, 1999).

The researcher described at the beginning of this section also uses variables measured at different levels. The two independent variables she wants to investigate are measured at different levels. Group member familiarity is measured at the level of the student (i.e., each student rates his or her familiarity with the other group members). In contrast, group size is measured at the level of the group. The number of unique observations for group member familiarity will therefore be higher than for group size. Again, to cope with this problem MLA is needed.

COMMON STRATEGIES FOR ANALYZING COLLABORATIVE LEARNING DATA

In this section we will describe two strategies that researchers employ when they are confronted with datasets such as the ones described previously. We will also highlight the dangers of these strategies.

Ignoring Nonindependence

A common strategy is to ignore the hierarchical structure of the dataset, the nonindependence, and the different units of analysis (Cress, 2008; Kenny, Kashy, & Cook, 2006; Kenny, Mannetti, et al., 2002). When CL researchers ignore the nonindependence in their dataset, they use statistical analyses such as *t*-tests, (M)ANOVA's, or (multiple) regression analysis. When this is the case, they run an increased risk of committing Type I or Type II errors (Kashy & Kenny, 2000). Nonindependence needs to be addressed when conducting statistical analyses, because it distorts estimates of error variances. This distortion makes standard errors, *p*-values, and confidence intervals invalid, when it is not taken into account (Kenny, 1995; Kenny & Judd, 1986; Kenny, Kashy, & Cook, 2006). Whether the chance to falsely reject (Type I error) or falsely accept (Type II error) the null hypothesis is increased, depends on the sign of the ICC (either positive or negative), and the type of dependent variable for which the ICC was calculated (see Kashy & Kenny, 2000 for a detailed discussion).

Using data from a study on the effects of an awareness tool on students' online collaboration, Janssen, Erkens, Kirschner, and Kanselaar (2011) highlighted the dangers of ignoring the nonindependence in their dataset. In this study half of the groups were given access to an awareness tool (treatment condition), whereas the other half were not given access to this tool (control condition). Janssen et al. found an ICC of .41 for the dependent variable, online collaboration, indicating a considerable influence of the group on the dependent variable and the presence of nonindependence in their data. When Janssen et al. analyzed the data using multiple regression analysis they found no effect of condition on the collaborative process. In contrast, when MLA was used, a different conclusion was drawn. In this case, a significant effect of condition on the collaborative process was found.

This example highlights the dangers of ignoring nonindependence. It should be noted however, that not in all cases will the differences between using analyses such as regression analysis or ANOVA and MLA will be this dramatic (i.e., drawing different conclusions when MLA is used). On the other hand, running an increased chance of committing Type I or Type II errors is something researchers should avoid.

Aggregating or Disaggregating Data

Another common strategy is to *aggregate* individual data to the level of the group (Kenny, Kashy, & Cook, 2006; Snijders & Bosker, 1999). In this case, a researcher who for example wishes to investigate the effect of prior knowledge on levels of participation during group discussion, would compute the average prior knowledge for the group members and use this group level measure as an independent variable instead of the prior knowledge measured at the level of the student. A drawback of this strategy is that it ignores the fact that prior knowledge is in essence an individual level variable (although it may be affected by group level variables). Another drawback is that by aggregating data, a researcher uses fewer observations for prior knowledge than are actually available (Cress, 2008). This means the power to detect a significant effect of the independent variable on the dependent variable is lowered, which in turn increases the risk of committing a Type II error (Snijders & Bosker, 1999).

The opposite strategy—*disaggregation* of group level data to the individual level—is also used in CL research. This means a researcher treats data measured at the level of the group as if they were measured at the level of the individual. Consider, for example, the study by Savicki, Kelley, and Lingenfelter (1996) about the effects of gender group composition on students' satisfaction with the collaborative process. Group composition was measured at the group level (all male, all female, or mixed groups), while satisfaction was measured at the individual level (students completed a questionnaire individually). In total, their sample consisted of six unique groups and 36 students. Savicki et al. conducted an analysis of variance to examine whether group composition affected satisfaction. However, this analysis does not take into account that group composition was measured at the group level. Thus, Savicki et al.'s analysis uses 36 observations for the group composition variable, while in fact there are only six observations for this variable. This led to an exaggeration of the actual sample size for this variable and increased the chance of committing a Type I error (Snijders & Bosker, 1999).

Using data from a study on the effects of representational guidance on student performance on a knowledge posttest, Janssen et al. (2011) showed the dangers of aggregating data to the level of the group. The dataset they used contained two conditions: in one condition students used a graphical version of a tool to construct an external representation, while in the other condition students used a textual version of this tool. In this dataset the assumption of nonindependence was violated, because the ICC of the dependent variable, posttest performance, was .32. Furthermore, this dataset contained different units of analysis: condition was measured at the level of the group (some groups used the graphical tool, others the textual tool) and posttest performance was measured at the level of the individual student. When Janssen et al. aggregated their data to the level of the group by calculating, for each group, the average posttest score of the individual group members, and then analyzed the differences between the two conditions using analysis of covariance, they found no significant differences between the graphical and textual tool. In contrast, when they used MLA, a significant difference between the graphical and textual tool for posttest performance was found.

This example highlights the dangers of aggregating data to the level of the group to deal with the problems described in the previous section. In the next section we will describe MLA in more detail by explaining its foundations and assumptions.

MULTILEVEL ANALYSIS

Foundations and Assumptions of Multilevel Analysis

Multilevel analysis is based on linear regression analysis, but extends this technique by allowing researchers to model the data at the level of the individual and the group simultaneously (Hox, 2003; Snijders & Bosker, 1999). Unlike regression analysis, which can only incorporate fixed coefficients (e.g., intercept or slopes), ML models can incorporate random coefficients as well. This means that in ML models the intercepts or slopes of the regression lines can vary for different groups of students. Such a model with varying

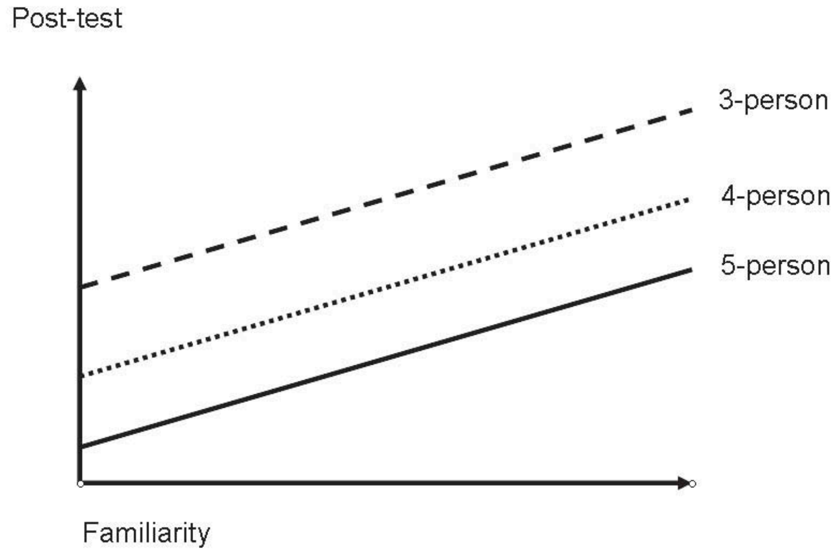

Figure 6.1 Depiction of a random intercept model.

intercepts is called a "random intercept model" (Hox, 2003; Snijders & Bosker, 1999). Figure 6.1 shows an example for a random intercept model for the fictional example introduced at the beginning of this chapter. Figure 6.1 shows not one regression line, but three different regression lines: one for each group in the sample. As can be seen, the intercepts of the three-person group and the four- and five-person groups differ.

The ML model can be extended by also allowing the slopes of the regression lines to vary. This is called a "random intercept, random slope model" (Hox, 2003; Snijders & Bosker, 1999). A graphical depiction of such a model for our fictional example is shown in Figure 6.2. In this case not only the intercept varies for the different groups in the

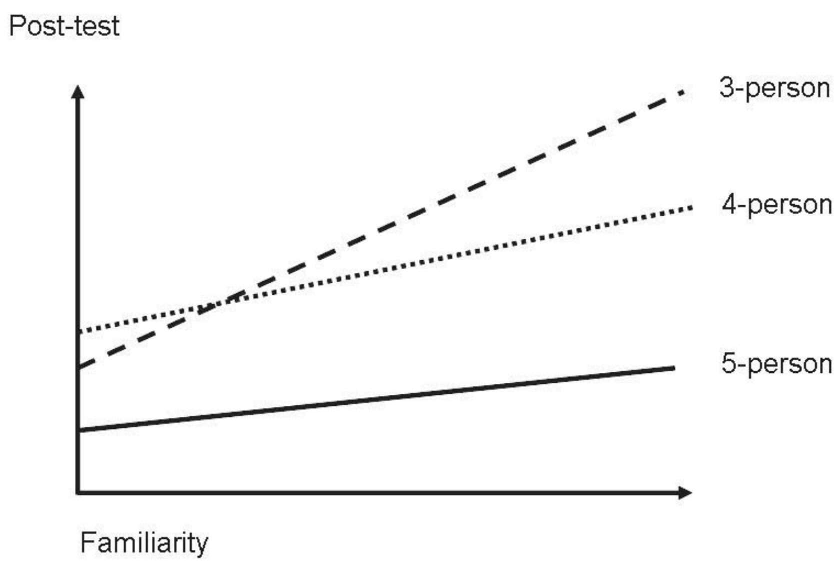

Figure 6.2 Depiction of a random intercept, random slope model.

dataset, but the slopes of the groups vary as well. Figure 6.2 shows that familiarity has a different effect on posttest performance for a three-person group than for four- and five-person groups (i.e., the slope is steeper for the three-person group).

ML models are estimated in two steps (Cress, 2008; Dedrick et al., 2009; Kenny, Kashy, & Cook, 2006; Snijders & Bosker, 1999). In the first step, the dependent variable (posttest performance in our example) is predicted by one or more *X* variables. *X* variables are measured at the lowest level, usually the individual. In our example group member familiarity is a level-1 variable and is thus used as an *X* variable. It is important to note that a *separate regression equation is computed for each level-2 unit*. In our example, groups are the level-2 unit. The first-step regression equation for group member *j* in group *i* is then

$$Y_{ij} = b_{0i} + b_{1i}X_{ij} + e_{ij} \tag{1}$$

In equation 1, b_{0i} represents the average post-test performance for group members in group *i*, and b_{1i} is the coefficient for the effect of *X*—group member familiarity—on posttest performance for group *i*.

During the second step of the MLA, the slopes and intercepts from the first-step analysis are treated as outcome variables in two separate regression models (Kenny et al., 2006). In these regression analyses, the regression coefficients b_{0i} and b_{1i} from equation 1 are assumed to be a function of a group-level predictor variable *Z* (group size, the level-2 variable in our example).

$$b_{0i} = a_0 + a_1Z_i + d_i \tag{2}$$

$$b_{1i} = c_0 + c_1Z_i + f_i \tag{3}$$

In equation 2, the intercepts from the first step are treated as a function of the *Z* variable, group size. This means that in equation 2 the average post-test performance, b_{0i}, in group *i* is predicted by group size. Parameter a_0 can be thought of as the grand mean of post-test performance across all students and groups, while parameter a_1 represents the effect of group size on post-test performance.

Equation 3 treats the effect of *X*—group member familiarity—as a function of the *Z* variable, group size. Thus, the effect of group member familiarity on posttest performance for group *i*, b_{1i}, is predicted by group size, Z_i. Parameter c_0 represents the overall effect of group member familiarity on post-test performance. Parameter c_1 represents the cross-level interaction between group member familiarity and group size. In other words, this parameter indicates the degree to which the effect of group size varies for students who know their group members well versus students who do not know their group members well.

In sum, the parameters a_0, a_1 and c_0 and c_1 describe *fixed effects*. The *random effects* are represented by the three parameters e_{ij}, d_i, and f_i. First, e_{ij} is the error component in the first regression equation (see equation 1). In our example, e_{ij} represents how post-test performance varies from student to student within groups, controlling for both group member familiarity and group size. Second, d_i represents the variance in the intercepts which cannot be explained by *Z* (see equation 2). In our example, this is the group variance in posttest performance that cannot be explained by group size. Finally, f_i is the group variance in the effects of *X* on *Y* (see equation 3). Thus, in our example f_i

represents how the effect of group member familiarity on posttest performance varies from group to group.

MLA usually involves estimating at least two models: an empty model, which includes no predictor, and just reveals the individual- and the group-level variances, and a model including one or more predictor variables. For both models, the deviance (a measure of the goodness of fit of the model) can be computed. By comparing the deviance of the latter model with the empty model, a decrease in deviance can be calculated. When this decrease in deviance is significant (tested with a χ^2-test), it can be concluded the latter model is a better model. Furthermore, the estimated parameters of the predictors can be tested for significance by dividing the regression coefficient by its standard error. This so-called *t*-ratio has approximately a standard normal distribution (Snijders & Bosker, 1999).

Table 6.1 shows a fictitious example of a MLA of the effects of group member familiarity (level-1 student predictor) and group size (level-2 group predictor) on posttest performance. This example was described at the beginning of this chapter and used earlier. Model 0 represents the empty model without level-1 or level-2 predictors. As can be seen, a large part of the variance in posttest performance is determined by the group (1.33 of a total of 1.33 + 2.83). The empty model also provides the ICC (.32). Model 1 contains group member familiarity as a level-1 predictor (*X*) and group size as a level-2 predictor (*Z*), as well as the cross-level interaction between these variables. It can be concluded from Table 6.1 that both group member familiarity and group size significantly affect posttest performance. As group member familiarity increases, posttest performance increases also. In contrast, when group size increases, post-test performance decreases. Finally, in this example no significant interaction effect is found between group member familiarity and posttest performance. This means that the effect of group member familiarity does not depend on the size of the group (i.e., the effect of group member familiarity on posttest performance is similar for small and large groups). Table 6.1 also shows the deviance of Model 1 is significantly lower than the deviance of the empty model, indicating the new model fits the data better than does the empty model.

Table 6.1 Fictitious Example of Multilevel Analysis of the Effect of Group Member Familiarity and Group Size on Post-Test Performance

	Parameter	Model 0		Model 1	
		β	SE	β	SE
Fixed effect					
Intercept	a_0	6.70	0.27	5.60	1.07
Group member familiarity (X)	a_1			0.28**	0.10
Group size (Z)	c_0			–0.42*	0.22
X * Z	c_1			0.28	0.26
Random effect					
Group level	σ_d^2	1.33	0.65	0.44	0.45
Individual level	σ_e^2	2.83	0.56	2.98	0.58
Deviance		355.70		344.63	
Decrease in deviance				11.07**	

* $p < .05$. ** $p < .01$.

It is important to note that MLA makes assumptions about the errors at each level in the ML model (Dedrick et al., 2009). The first-level errors, e_{ij} in equation 1, are assumed to be independently and normally distributed. Furthermore, it is also assumed that the second-level errors, d_i in equation 2 and f_i in equation 3, are also normally distributed. In most cases their covariance is not assumed to be equal to zero. Finally, all the assumptions that are made for linear regression analysis, such as linearity and no multicollinearity also apply to MLA (Field, 2009).

Applications of Multilevel Analysis to Collaborative Learning Data

Although MLA is a relatively new statistical technique in the field of CL research, several researchers have used MLA to answer their research questions. In this subsection we will shortly describe several studies that employed MLA to highlight its potential to answer meaningful research questions in the field of CL.

Paletz et al. (2004) examined how ethnic composition of groups affected satisfaction with the collaborative process using MLA. Their study contrasted groups that were composed of mostly Caucasians with groups that were composed of mostly ethnic minorities (e.g., Hispanic or African American). Students worked in groups of three. In total, their sample consisted of 108 students working in 36 groups. They found that 40% of the variance in satisfaction with the collaborative process was due to group-level factors. Furthermore, Paletz et al. were able to show that ethnic composition affected satisfaction: groups composed of mostly ethnic minorities reported higher satisfaction with the collaborative process than did groups composed of mostly Caucasian students.

Strijbos et al. (2004) investigated how role assignment affected online collaborative learning. Their sample consisted of 33 students collaborating in 10 groups. Strijbos et al. compared the perceived group efficiency (PGE) of groups with role assignments to the PGE of groups without role assignments. In Strijbos et al.'s study, a considerable part of the variance in PGE was due to group-level factors (ICC was found to be .47). MLA showed that in the role condition, groups reported higher PGE than in the nonrole condition.

Whereas the studies by Paletz et al. (2004) and Strijbos et al. (2004) used ML models with two levels (individual and group), Schellens et al. (2005) used a three-level model. Schellens et al. measured students' level of knowledge construction in asynchronous discussion groups at four different measurement occasions. Their study is therefore a good example of how MLA can also be used on repeated measure data and to investigate trends or developmental patterns in the data. In Schellens et al.'s study, measurement occasions were nested within students, and students were in turn nested within groups. In total their sample consisted of 286 students, collaborating in 23 groups. In contrast to the two previously described studies, the study conducted by Schellens et al. found only a small effect of group-level variables: 0.54% of the variance in knowledge construction could be attributed to differences between the groups.

In sum, the examples provided in this section highlight the usefulness of MLA for CL research. Although not all studies report large and significant ICCs, it seems that a moderate part of the variance in the dependent variables CL researchers investigate can be explained by group-level factors. This justifies the use of MLA for CL research. Moreover, due to the relatively small sample sizes in CL researcher (especially at the level of the group) the ICC may not be significant, while it is actually large enough to bias standard errors, *p*-values, and so on. Kenny et al. (2002) therefore propose assuming group data are nonindependent and use MLA even though the ICC is not significant.

Limitations of Multilevel Analysis

Of course not all data-analytic problems that CL researchers encounter are solved by using MLA. Furthermore, MLA has its own limitations. First, MLA is mostly used when the dependent variable is measured at the interval level of measurement. Sometimes however, researchers may be interested in dichotomous (e.g., success or failure of group work) or categorical dependent variables (e.g., levels of knowledge construction). Although MLA techniques have been developed to incorporate these kinds of dependent variables (multilevel logistic regression; see Snijders & Bosker, 1999), they are rarely adapted to CL data.

Second, for an adequate analysis of collaborative learning using MLA, it is often suggested that large sample sizes at all levels (individual as well as group) are necessary (Cress, 2008; Maas & Hox, 2005). Maas and Hox, using a simulation study, demonstrated that only a small sample size at the group level (less than 50 groups) is problematic and leads to biased estimates. A small sample size at the individual level (groups consisting of five group members or less), does not appear to be problematic. This means that, in order to use MLA confidently for CL data, researchers should collect data about at least 50 groups. CL researchers, however, often employ less than 50 groups in their studies (see, for example, the sample studies cited above). Given the complexity of CL research and how time-consuming data collection and analysis often are, a sample size of at least 50 groups places a heavy burden on researchers. We agree with Cress (2008) that as a minimum standard in CL research, the ICC should be calculated and tested for significance. This will give the researcher and his/her audience insight into the effect of group-level variables on the dependent variable and of the existence of nonindependence in the dataset. When this is the case, researchers can take other precautions when application of MLA seems problematic due to a small sample size at the highest level (less than 50 groups). One such precaution might be to use a more stringent alpha level (i.e., .01 instead of .05) or to aggregate all individual variables to the group level (cf., Kenny, Kashy, & Cook, 2006).

Software Applications for Multilevel Analysis

For the researcher who wants to conduct MLA on his/her dataset, several software applications are available (see Robert & McLeod, 2008 for an excellent overview). A commonly used application is MLwiN (http://www.cmm.bristol.ac.uk/MLwiN/index.shtml) developed by the team at the Centre for Multilevel Modelling at the University of Bristol. Another commonly used application for MLA is HLM (Raudenbush, Bryk, Cheong, Congdon, & Du Toit, 2004; see also http://www.ssicentral.com/hlm/index.html).

Conventional statistical software such as SPSS and SAS also incorporate procedures for carrying out MLA. For researchers who want to perform advanced MLA, using, for example, binary or binomial distributions, especially SPSS may be less suited than specialized applications such as MLwiN or HLM. On the other hand, SPSS and SAS are more user friendly with respect to data manipulation than MLwiN and HLM. For SAS, Campbell and Kashy (2002) illustrate how the MIXED procedure can be used to perform MLA on CL data. For SPSS, Field (2009) describes how the Mixed Models option can be used to perform MLA. In addition, the book written by Kenny, Kashy, and Cook (2006) and the text by Robert and McLeod (2008) contain detailed examples, including syntax, of how both SPSS and SAS can be used to perform MLA.

CONCLUSION

In this chapter we discussed the data analytical problems CL researchers frequently encounter, namely hierarchically nested datasets, nonindependence of dependent variables, and different units of analysis. We argued that, in order to take these problems into account, MLA should be used. We also demonstrated that alternative analysis strategies such as ignoring nonindependence or aggregating data can lead to different results and possibly to mistakes regarding the significance or non-significance of these results. We therefore strongly advocate the use of MLA in CL research. Fortunately, more and more CL researchers are beginning to use this technique to answer their research questions.

As noted before, not all data-analytic problems that CL researchers encounter are solved by using MLA because it is mostly suited for analyses of interval-level dependent variables. Sometimes, however, researchers may be interested in dependent variables measured at other levels (e.g., dichotomous or categorical). MLA can be adapted to those situations as well, but they are seldom applied in CL research. One of the key challenges for CL research is to develop and incorporate these more advanced applications of MLA as well.

Second, the issue of sample size and power addressed earlier can pose a problem for CL researchers who wish to use MLA to answer their research questions. Often, time and budget considerations do not allow researchers to collect data for large numbers of groups. An option to deal with this problem may be to focus on smaller groups, such as dyads or triads. Because the sample size at the group level seems to be the bottleneck with respect to sample size and power, using dyads or triads instead of larger groups limits the number of students needed to reach an adequate sample size (e.g., 100 students when dyads are used, 150 students when triads are used). Field (2009) however, concludes that many factors affect the power of MLA. It is therefore advisable that CL researchers use sample size and power calculators to determine the optimal sample size for their model. HLM will for example allow power calculations for a ML model. Additionally, Tom Snijders' homepage (http://stat.gamma.rug.nl/multilevel.htm) contains a program that performs power analysis for two-level models.

CL researchers are sometimes interested in data over time. An example might be how familiarity with group members affects trust development in groups over time. To investigate this question, a researcher would collect data about trust levels on different occasions. This adds even more problems to the data analysis. The effects of familiarity on trust may not be the same at every measurement occasion (e.g., its effects may be greater at the beginning of the collaboration). Furthermore, the level of trust at measurement occasion 1 may also have an effect on the level of trust at occasion 2 (if trust was high at occasion 1, this may affect trust at occasion 2). This creates a new type of nonindependence: autocorrelation (Kenny, Kashy, & Cook, 2006). Again, MLA techniques have been developed to analyze time-series data (cf., Chiu & Khoo, 2003, 2005; Kenny, Kashy, & Cook, 2006), but they are not often used in CL research. CL researchers should therefore begin to investigate the possibilities of using MLA for time-series data.

Finally, MLA will not be a suitable technique to answer all research questions. As was stated at the beginning of this chapter, quite a lot of CL research focuses on capturing the interactive processes that unfold between group members (i.e., process-oriented research). In some cases researchers are interested in providing "thick" or "rich" descriptions of the collaborative process (Baker, 2003; Hmelo-Silver & Bromme, 2007). In such cases, it is not necessary to apply MLA. Furthermore, it has been argued that studying intersubjective meaning making or group cognition should be the focus of CL research (Stahl, 2006; Suthers, 2006). This involves studying "how people make sense of situations and of each other" (Suthers, p. 321). Researchers with such a perspective on CL research could object to disentangling group and individual aspects of CL. They would argue that in order to understand the collaborative process, the group should be the unit of analysis, not the individual. Again, if one has such an approach to studying CL, using MLA will not be a sensible strategy.

CL research can still make progress by incorporating MLA in its repertoire of analysis techniques. It is an encouraging development that CL researchers are increasingly turning toward MLA. It is our hope and expectation that this development will continue and that CL researchers are going to find new ways to deal with the complex data analytical problems they are faced with. Ultimately, this will lead to a better understanding of how group-level factors (e.g., group composition), student-level factors (e.g., prior knowledge, motivation), and their interaction, affect the collaborative process and student learning. Furthermore, when researchers combine MLA with qualitative analyses in a mixed methods design (cf., Leech & Onwuegbuzie, 2009) an even more complete picture of the CL process is possible.

REFERENCES

Baker, M. (2003). Computer-mediated argumentative interactions for the co-elaboration of scientific notions. In J. Andriessen, M. Baker, & D. Suthers (Eds.), *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments* (pp. 47–78). Dordrecht, Netherlands: Kluwer Academic.

Bonito, J. A. (2002). The analysis of participation in small groups: Methodological and conceptual issues related to interdependence. *Small Group Research, 33*, 412–438.

Campbell, L., & Kashy, D. A. (2002). Estimating actor, partner, and interaction effects for dyadic data using PROC MIXED and HLM: A user-friendly guide. *Personal Relationships, 9*, 327–342.

Chiu, M. M., & Khoo, L. (2003). Rudeness and status effects during group problem solving: Do they bias evaluations and reduce the likelihood of correct solutions? *Journal of Educational Psychology, 95*, 506–523.

Chiu, M. M., & Khoo, L. (2005). A new method for analyzing sequential processes: Dynamic multilevel analysis. *Small Group Research, 36*(5), 600–631.

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.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., … Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*, 69–102.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In H. Spada & P. Reimann (Eds.), *Learning in humans and machine: Towards an interdisciplinary learning science* (pp. 189–211). Oxford, England: Elsevier.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Hmelo-Silver, C. E., & Bromme, R. (2007). Coding discussions and discussing coding: Research on collaborative learning in computer-supported environments. *Learning and Instruction, 17*, 460–464.

Hox, J. (2003). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Janssen, J., Erkens, G., & Kanselaar, G. (2007). Visualization of agreement and discussion processes during computer-supported collaborative learning. *Computers in Human Behavior, 23*, 1105–1125.

Janssen, J., Erkens, G., Kanselaar, G., & Jaspers, J. (2007). Visualization of participation: Does it contribute to successful computer-supported collaborative learning? *Computers & Education, 49*, 1037–1065.

Janssen, J., Erkens, G., Kirschner, P. A., & Kanselaar, G. (2011). Multilevel analysis in CSCL research. In S. Puntambekar, G. Erkens, & C. Hmelo-Silver (Eds.), *Analyzing interactions in CSCL: Methodologies, approaches and issues* (pp. 187-205). New York: Springer.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 38*, 365–379.

Kashy, D. A., & Kenny, D. A. (2000). The analysis of data from dyads and groups. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 451–477). Cambridge, England: Cambridge University Press.

Kenny, D. A. (1995). The effect of nonindependence on significance testing in dyadic research. *Personal Relationships, 2*, 67–75.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99*, 422.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford.

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology, 83*, 126–137.

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). A cognitive load approach to collaborative learning: United brains from complex learning. *Educational Psychology Review, 21*, 31–42.

Klein, J. D., & Pridemore, D. R. (1992). Effects of cooperative learning and need for affiliation on performance, time on task, and satisfaction. *Educational Technology Research and Development, 40*, 39–47.

Leech, N. L., & Onwuegbuzie, A. J. (2009). A typology of mixed methods research designs. *Quality & Quantity, 43*, 265–275.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modelling. *European Journal of Research Methods for the Behavioral and Social Sciences, 1*, 85–91.

Mevarech, Z. R., & Kramarski, B. (2003). The effects of metacognitive training versus worked-out examples on students' mathematical reasoning. *British Journal of Educational Psychology, 73*, 449–471.

Nichols, J. D. (1996). The effects of cooperative learning on student achievement and motivation in a high school geometry class. *Contemporary Educational Psychology, 21*, 467–476.

Paletz, S. B. R., Peng, K. P., Erez, M., & Maslach, C. (2004). Ethnic composition and its differential impact on group processes in diverse teams. *Small Group Research, 35*, 128–157.

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & Du Toit, M. (2004). *HLM6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Robert, J. K., & McLeod, P. (2008). Software options for multilevel models. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 427–467). Greenwich, CT: Information Age.

Saleh, M., Lazonder, A. W., & De Jong, T. (2005). Effects of within-class ability grouping on social interaction, achievement, and motivation. *Instructional Science, 33*, 105–119.

Savicki, V., Kelley, M., & Lingenfelter, D. (1996). Gender and group composition in small task groups using computer-mediated communication. *Computers in Human Behavior, 12*, 209–224.

Schellens, T., Van Keer, H., & Valcke, M. (2005). The impact of role assignment on knowledge construction in asynchronous discussion groups: A multilevel analysis. *Small Group Research, 36*, 704–745.

Schepers, J., de Jong, A., Wetzels, M., & de Ruyter, K. (2008). Psychological safety and social support in groupware adoption: A multi-level assessment in education. *Computers & Education, 51*, 757–775.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. Cambridge, MA: MIT Press.

Strijbos, J.-W., & Fischer, F. (2007). Methodological challenges for collaborative learning research. *Learning and Instruction, 17*, 389–393.

Strijbos, J.-W., Martens, R. L., Jochems, W. M. G., & Broers, N. J. (2004). The effect of functional roles on group efficiency: Using multilevel modeling and content analysis to investigate computer-supported collaboration in small groups. *Small Group Research, 35*, 195–229.

Strijbos, J.-W., Martens, R. L., Jochems, W. M. G., & Broers, N. J. (2007). The effect of functional roles on perceived group efficiency during computer-supported collaborative learning: A matter of triangulation. *Computers in Human Behavior, 23*, 353–380.

Suthers, D. D. (2006). Technology affordances for intersubjective meaning making. *International Journal of Computer Supported Collaborative Learning 1*, 315–337.

Van der Linden, J. L., Erkens, G., Schmidt, H., & Renshaw, P. (2000). Collaborative learning. In P. R. J. Simons, J. L. Van der Linden, & T. Duffy (Eds.), *New learning* (pp. 1–19). Dordrecht: Kluwer Academic.