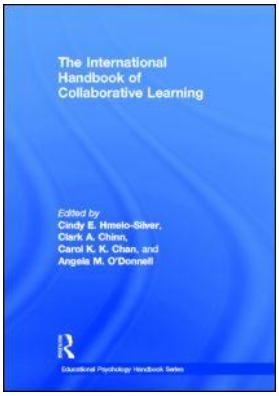


The International Handbook of Collaborative Learning

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

Quantitative Methods for Studying Small Groups

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203837290.ch5
Ulrike Cress, Friedrich Wilhelm Hesse
Published online on: 04 Feb 2013

II

Studying Collaborative Learning

5

QUANTITATIVE METHODS FOR STUDYING SMALL GROUPS

ULRIKE CRESS AND FRIEDRICH WILHELM HESSE

Knowledge Media Research Center, Tübingen, Germany

Quantitative methodology includes various different methods of gathering and analyzing empirical data. What these methods have in common is that they do not just categorize phenomena but measure them and express their relationships in terms of quantity. If applied to collaborative learning, these methods allow researchers to analyze both learning processes and learning outcomes of individuals, interaction processes among group members, and learning outcomes of groups. The aim of quantitative research is to go beyond simply describing processes and outcomes, but to make and test predictions about the interrelation of these processes and about factors that trigger them. For understanding the quantitative research approach, we will first describe some fundamental concepts of quantitative methodology: its hypothetico-deductive approach, its experimental logic, operationalization, measurement, and use of inferential statistics. On this basis, we will present some quantitative methods which are suited for collaborative learning (CL) studies. But quantitative research is a wide field, and this chapter will not be able to give an exhaustive overview of its methods. Here we would refer readers to standard literature like Everitt and Howell (2005), or Keeves (1997). The focus of this chapter is on showing quantitative approaches for analyzing data of collaborative learning. These consist of different types: data that describe interactional processes, data that describe individuals within groups, and data that describe groups. Considering this specific structure of data, we will introduce the concept of *units of analysis* and describe *events*, *interactions, persons,* and *groups* as relevant units in the context of analyzing collaborative learning. For each of these levels of analysis, we will explain some typical quantitative methods and provide examples from CL-relevant studies. In addition, one specific example will be cited throughout the whole chapter. This will make it easier to concentrate on relevant aspects of all the methods which are presented here.

FUNDAMENTAL CONCEPTS OF QUANTITATIVE RESEARCH

Theory-Based Approach

Quantitative research describes phenomena in the world and their interrelations in terms of quantity. In many cases it takes a *theory-based approach*: It starts with some theory about interrelations between constructs, and deduces from that a prediction about what will happen in the observable world. Such predictions, which serve as hypotheses, are then tested by observing empirical phenomena. With the help of statistical analyses, the researcher then decides whether or not the observed empirical phenomena are congruent with these predictions. This leads to the acceptance or rejection of the hypotheses and allows maintaining or falsifying the proposed theory (Popper, 1963). Figure 5.1 shows this hypothetico-deductive approach.

Theories mostly focus on causal relationships and try to explain phenomena as effects of well-defined causes. The fewest assumptions a theory needs to explain a phenomenon (principle of parsimony) the better it is. Based on this approach, quantitative research often aims at explaining as much as possible of occurring variation with as few predictors as possible.

The hypothetico-deductive model leads to the distinction between independent and dependent variables. The first describes causes, the second effects. The causal interrelation between both can best be tested with experiments, where different experimental conditions are established by experimental manipulation, which only differ with regard to the value of the independent variable. The subjects are assigned randomly to different experimental conditions. This ensures that differences between subjects in different conditions will not result from any other feature than the value of the independent variable. Only experiments can test causal relationships, because only randomization guarantees that variation in the dependent variable can be attributed—in a logically correct way—to variation in the independent variable.

Let us clarify these fundamental concepts of the quantitative approach by referring to a fictitious empirical study (which will serve as an example throughout this chapter). The example is simplified, but describes a prototypical research question in CL (as, for example, in Haake & Pfister, 2010; Rummel, Spada, & Hauser, 2008; Fischer et al. this volume; O'Donnell this volume): A researcher would like to know (the research question) if scripts are an effective way of instructing learners. Her research interest results from a theory—it may be an implicit underlying theory—that scripting has some influence on

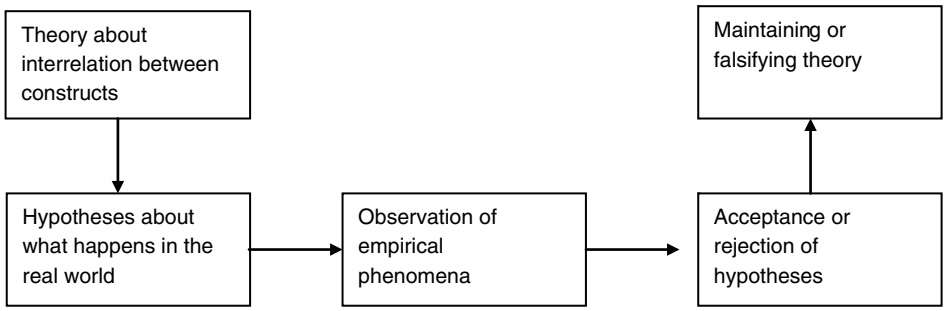


Figure 5.1 Hypothetico-deductive approach of quantitative research.

cooperative learning. So she has a theoretical assumption about some causal relationship. This assumption includes, at least, an idea of what makes cooperative learning effective, and how an interaction script might support such effective learning. As a first step, our researcher makes this theoretical consideration explicit, and finds an adequate operationalization of the script (independent variable) and a way of measuring effective cooperative learning (dependent variable). She then develops a prototypical script, with all relevant features, and plans an experiment with two conditions. In the experimental condition, groups of three learners (triads) have to work with this script, in the other condition (control condition), the triads do not use a script.

Operationalization and Measurement

As quantitative research typically starts with some theoretical assumption, the issue of adequate *operationalization* is crucial. What needs to be determined is the way in which theoretical constructs and their interrelations are made observable. Which variables serve as independent or dependent variables, how are they defined and measured? In the example about effects of scripting, the researcher first has to define what a script is, what its relevant features are, and how its effect on cooperative learning can be measured. The operationalization of the script (as independent variable) is good if the script which is used in the experiment is a prototypical realization of the underlying theoretical idea what scripts are. The goodness of the operationalization of the dependent variable can be described with its *reliability* and *validity* (Carmines & Zeller, 1979). A measurement is reliable if it is exact and if, for example, a retest turns out similar results, and it is valid if it really measures what it claims to measure. Testing the validity of a measurement may be difficult. A valid measure should show high correlations with other measures of the same concept, and low correlations with measurements of other concepts (construct validity). In our example, the researcher also has to decide how to measure the dependent variable. She may decide to test "effective cooperative learning" in three different ways: (a) by categorizing the learners' contributions during their interaction into the four categories questions, answers, statements, and off-task; (b) by measuring each learner's performance in an individual postexperimental knowledge test; (c) by measuring performance of the triads by checking the time they need for solving a following collaborative transfer task. Each of these measurements covers a different aspect of the idea of "effective collaborative learning," and each of these operationalizations leads to a different dependent variable. Taken as a whole, this may be a valid operationalization of the researcher's idea of collaborative learning.

Internal, External, and Statistical Validity

Not only single measurements should be valid. The study as a whole should be valid as well. The term *external validity* describes whether the results of a study can be generalized. In order to allow generalizations about other people, the participants of a study have to be a random sample of all learners covered by the theory. The participants of the study have to be *stochastically independent*, which means that each person of a population has the same chance of being part of the sample (this may be a problem in CL studies as we will describe in the next sections). External validity also requires that the results of the study can be generalized to other situations. This may be a problem with experimental studies. Experimental situations are sometimes very constructed and artificial, and generalizing their results to real-life situations may be critical. In our example, a study

may show that a script leads to effective cooperative learning in the experiment. But it will still remain an open question whether this effect would also occur in a real school setting. It may be possible that in real life, where learning is affected by many different circumstances, scripts have no strong effect, or that the effect of a script is quite different from the experimental situation, in which all other influences were kept under control. Because experimental research works according to the principle of parsimony, and only focuses on the effect of a few independent variables on a few dependent variables, it does not take into account the complexity of all possible influencing factors. So it may lack external validity. But this tight focus is necessary for ensuring high *internal* validity. This kind of validity describes that an observed effect in the dependent variable can be attributed in a logically correct way to the independent variable. So, if we try to prove some causality, we need a high degree of internal validity. And this can only be taken for granted if the experimental conditions differ systematically *only* in the dependent variable, and in no other feature. And this is exactly what experiments with their randomized research designs ensure. But, of course, the results of an experiment can only then be generalized if its conclusions are statically valid. This means that *inferential statistics* must be used. Only then will data provided by a random sample allow valid statements about the respective population.

What would these fundamental concepts of quantitative research mean if applied to our example, a prototypical study about scripts? The researcher would have to select a representative sample of students as participants in her experiment. She would then have to assign them randomly to both conditions (with and without scripts). After conducting the experiment, she would compare the number of task-relevant statements, the learner's performance in the knowledge test, and problem solving time between both conditions. Using inferential statistics, she would decide if differences between the two conditions are statistically significant, that is to say, that they are unlikely to have occurred just by chance, and can be generalized to the population. On the basis of these results, she would maintain her theory that scripts are effective tools for cooperative learning, or consider it falsified.

EXPERIMENTAL AND NONEXPERIMENTAL RESEARCH

Most quantitative research is theory-based research, but it is not necessarily experimental. Quantitative research also plays a role in real settings, with people who act in their real everyday environment. In so-called *quasi-experimental* studies, people belong to naturally given conditions. Here, no randomization is possible, and people who belong to different conditions will not just differ in the value of one independent variable, but also in many other characteristics. Such studies raise the question of internal validity, because differences in the dependent variable have not necessarily been affected by differences of the independent variable. In our example about the effects of scripting, a quasi-experimental study would deal with naturally occurring groups who are working with or without scripts. Such groups might, for example, consist of different school classes. But these classes would not only differ in their use of scripts. There may be other differences as well: teachers who are using scripts might be more experienced in instructing pupils for cooperative work than teachers who have never used such scripts. So the use of scripts and the teachers' experience may be confounded. But when such data have been obtained, quantitative researchers try in many cases to find evidence

for causality by statistically controlling for covariates. For example, Dewiyanti, Brand-Gruwel, Jochems, and Broers (2007) predicted students' satisfaction of CSCL courses by referring to characteristics of the individuals and to courses in which they had participated. Campbell and Stanley (1963) describe experimental and nonexperimental designs, their threads for internal and external validity, and ways to handle them.

However, finding causal effects is not always the aim of quantitative research. Some studies just examine relationships between variables. This is done by using correlations or regression models. They describe relationships between variables, without differentiating between cause and effect. For example, Hijzen, Boekaerts, and Vedder (2006) predicted the quality of cooperative learning by referring to different student goals. This type of studies also requires operationalization, but they do not distinguish between independent and dependent variables. They just show a covariation between variables. Such an exploratory approach may be necessary if a theory is not developed well enough to make a prediction. Quantitative research may even sometimes take an *inductive* approach and take phenomena into account that emerge from empirical data. This would apply, for example, when a researcher conducts a post hoc search for factors, which could explain a pattern found in the data. For example Dehler, Bodemer, Buder, and Hesse (2009) and Hmelo-Silver, Chernobilsky, and Jordan (2008) contrasted effective and noneffective groups in order to identify relevant processes. But in the logic of quantitative research, such exploratory analyses are not an end in itself, they provide just a starting point for generating hypotheses, which would then need to be tested in a hypothetico-deductive approach. Only then will it be possible to generalize results and make predictions about relationships.

METHODS ACCORDING TO DIFFERENT LEVELS OF ANALYSIS

CL research deals with different levels and granularities:

- the level of single events (which are the elements of interaction)
- the level of interactions (which are the sequences of such events)
- the level of learners
- the level of the group

It will always depend on the focus of a study and the underlying theory which of these units of analysis has to be considered. In turn, the choice of the unit of analysis will determine which statistical method is appropriate. For each of the four units of analysis, we provide some typical methods and refer to our example about the effects of scripts. In addition, we provide prototypical examples from previously published CL studies. A handbook chapter like this one cannot, of course, provide an exhaustive survey; what is intended here is an overview of a variety of different research questions according to the different levels.

Methods for Describing and Quantifying Single Events

The lowest level of analysis of CL deals with single activities of individual learners. These are the elements of the collaborative processes. *Quantitative content analysis* (Berelson, 1952; Rourke & Anderson, 2004) identifies important events that occur during the learners' interaction, and analyzes the frequencies of these events. In CL,

such important events may consist of communicative acts, like utterances, questions, answers, or written messages, as well as nonverbal behavior, like gestures and other cue actions. It depends on the focus of study and underlying theory regarding which events are seen as important, and how these events are made observable and measurable. In order to produce reliable and valid measurements, several conditions must be fulfilled: (a) The procedure which creates the learning and interacting situation must be described in such a way that other researchers are able to replicate the study. In order to be able to categorize the content of communication, people's communication and interaction have to be segmented into units. For such segmentation, *manifest units* may be used (e.g., words or messages), but underlying *latent units* (e.g. meaning units, see Rourke, Anderson, Garrison, & Archer, 2000) may also be suitable. (b) The categorization should be based on a well-defined coding schema which guarantees a high degree of reliability. (c) The codes which are used should be derived from the research question, in the sense that these codes are valid operationalizations of the underlying theoretical concepts and constructs.

The coding schema itself provides categorical codes, which are in most cases (but not necessarily) mutually exclusive and exhaustive. This means that the use of a coding schema produces nominal data. The reliability of coding should be measured with the interrater reliability; this describes to what extent different people, who have coded the same content independently from each other, achieve similar results. The easiest way of measuring interrater reliability is the relative amount of agreement (Holsti, 1969), but this will not take into account the fact that different raters may also have agreed by chance. Measurements which correct this random agreement are *Cohen's kappa* (Cohen, 1960) and *Krippendorff's alpha* (Krippendorff, 2004). Cohen's kappa calculates the chance-corrected agreement between two raters, while Krippendorff's alpha deals with any number of raters and also takes into account the magnitude of misses. *Intraclass correlation* may also be used for calculating interrater correlation for more than two raters (Shrout & Fleiss, 1979).

If codes are assigned to latent units, not only the classification of content, but also its segmentation might jeopardize reliability (Strijbos, Martens, Prins, & Jochems, 2006). When referring to "meaning units," segmentation and coding are intertwined. Different raters may segment differently, which leads to overlapping units with different codes. This causes serious problems. Strijbos and Fischer (2007) state that finding the adequate unit of analysis is one of the big methodological challenges of CL research, and Strijbos et al. (2006) propose a segmentation procedure which is systematic and independent of the coding categories.

The *validity of a coding schema results* from adequate operationalization of the underlying theory (Rourke & Anderson, 2004). At the conceptual level, the coding schema should capture all relevant aspects of the theoretical construct. But at the empirical level, it will often be difficult to check validity. One possible way of estimating the validity of a coding schema is by comparing classification results of different coding schemas. A valid coding schema will lead to measurements that are consistent with those obtained through other methods. Another indicator for high validity is measurement that is sensitive to group differences or experimental interventions (e.g., Boxtel, van der Linden, & Kanselaar, 2000; Hron, Cress, Hammer, & Friedrich, 2007).

Strijbos and Stahl (2007) have drawn attention to the following paradox of reliable and valid coding in exploratory CL research: If research attempts to focus on unique

interactions, it cannot use preexisting categories, because these will not be able to capture new, context-specific phenomena. For capturing these phenomena, new categories have to be defined in an inductive way. But these new categories lack reliability and validity. From a methodological point of view, this paradox cannot be solved by a single study. Coding schemas can only be tested for reliability and validity if the critical events occur frequently. So the paradox can only be solved in a deductive way: It requires, first of all, an assumption about the occurrence of the relevant critical events, and on the basis of this assumption a researcher can then create a situation in which these events are expected to occur frequently. A predefined coding schema will then be able to capture these critical events, and the reliability and validity of this schema can be measured. This means that, from the point of view of quantitative research, new and surprising events cannot be treated adequately in one study. When new phenomena have been observed, further studies are needed in order to be able to measure them reliably.

Because of this need to obtain many data and carry out multiple studies in order to test reliability and validity of a coding schema, we recommend reuse or adaptation of an existing coding schema where possible. In the last years, several suitable schemas have been introduced. De Wever, Schellens, Valcke, and Van Keer (2006) give an overview of 15 instruments which were used for analyzing online asynchronous discussion groups. The authors provide information about their theoretical foundation, reliability, and validity. Apart from that, there exist many other well-established coding schemas; for example, the multidimensional coding schema for individual actions in interactive group processes by Chiu (2001), or the functional category system by Poole and Homes (1995).

How can the data which result from coding schemas be analyzed? As described above, categorizing single events leads to nominal data. Such data can just be analyzed by counting the occurrence of the single categories, or by depicting the percentage of categories (e.g., Kanuka, Rourke, & Laflamme, 2007). This may be done for each learner or for the learning group. If we want to use inductive statistics and compare different situations, we should be aware that the single events are not stochastically independent. If we code, for example, learners' utterances by categorizing questions, answers, and statements, we might find that the learning partners show a symmetric interaction: If one learner asks many questions, this may induce many answers by the other learning partners. This means that different learners will not behave independently from each other. So the frequency of single events describes interaction within the group, not the behavior of independent learners. What we can do is to compare frequencies of events between different groups. But then the level of analysis is that of the group, not of the event. In our example about scripting, we might, for example, compare triads working with and without scripts with regard to the frequency of off-task talk. In such an analysis, the data will be based on a categorization of single events, but these will then be pooled within groups, and different groups will be compared. Many CL studies have taken this approach (e.g. Hron et al., 2007; Meier, Spada, & Rummel, 2007; Wolfe, 2007).

Methods that Focus on Interactions

In collaborative learning, it is often the interaction between learners that is of interest. Analysis of interaction processes is, in most cases, based on a categorization of single events, as described in the previous section. But it is mainly the sequence of these

events that is meaningful for considering interaction. One way of analyzing interaction quantitatively is sequence analysis for categorical data (Bakeman & Gottman, 1997). It includes different methods; for example, lag-sequential analysis or log-linear methods. Even if these have been not used very often in CL research so far, they will be presented here briefly, because of their great potential for dealing with sequential processes.

Lag-sequential analysis describes sequences of events as Markov chains, in which current events determine the probability of events in the next period (Bakeman, Adamson, & Strisik, 1995; Bakeman & Gottman, 1997; Faraone & Dorfman, 1987; Gottman & Roy, 1990). So lag-sequence analysis focuses on transitional probabilities (probability of one event following the other), which describe typical patterns of event sequences. A typical research question is presented in Hou, Chang, and Sung's (2008) study, which searched for typical sequences of problem-solving events like "proposing a problem," "proposing a solution," "comparing," and "forming a conclusion." Jeong (2005) illustrates the use of lag-sequential analysis nicely. As a first step, events must be classified, using a coding schema. Then frequency with which each type of event category is followed by other event categories has to be analyzed. We might describe this process using prototypical data from our example about scripting. Here a coding schema could classify the events into (a) questions, (b) answers, and (c) statements. A prototypical sequence could be as follows:

a b c a b b c a b c b a b b c

Out of such a sequence, a frequency matrix may be constructed which shows how often each one of the event types succeeds the others. For constructing it, a moving two-unit window may be used which slides across the sequence.

(a b) c a b b c a b c b a b b c
a (b c) a b b c a b c b a b b c
a b (c a) b b c a b c b a b b c

The window moves across the stream of codes and we can tally the frequency of each event pair. This leads to the frequency matrix shown in Table 5.1. It shows, for example, that a question has never been succeeded by another question, but four times by an answer.

As a second step, these absolute frequencies are converted into relative frequencies. These describe the transitional probability that a question, answer or statement is succeeded by a question, answer or other statement (see Table 5.2).

Table 5.1 Frequency Matrix

	Succeeded by			
	Question (a)	**Answer (b)**	**Statement (c)**	**Total**
Question (a)	0	4	0	4
Answer (b)	1	2	4	7
Statement (c)	2	1	0	3
Total	3	7	4	14

Table 5.2 Transitional Probability Matrix

	Succeeded by		
	Question (a)	**Answer (b)**	**Statement (c)**
Question (a)	.00	1.00	.00
Answer (b)	.14	.29	.57
Statement (c)	.67	.33	.00

These probabilities are *conditional* probabilities. They describe the probability that an event category occurs if a preceding event has taken place. This transition matrix is the base of the transitional state diagram shown in Figure 5.2. Here the thickness of a pointed arrow represents transition probability.

In the last few years, the use of such transition state diagrams has become quite common in CL research, as a way of detecting and visualizing patterns of event sequences (e.g. Erkens, Kanselaar, Prangsma, & Jaspers, 2003; Hou, Sung, & Chang, 2009; Jeong, 2003; Jeong & Joung, 2007; Kanselaar et al., 2002).

Lag-sequential analysis may not only be applied to two-event sequences, but also to sequences containing more than two events. With three events, the window that slides over the event sequences will have three positions, and all possible chains of three events have to be tallied. This leads to a three-dimensional frequency matrix.

The longer the sequence of events, the (exponentially) larger the matrix of possible transitions will be. So with three-event sequences and a code of five categories, we would get a matrix of 125 cells (5^3). This means that with many codes, and longer sequences, the expected frequencies and probabilities will be very small. According to Bakeman and Quera (1995) and Tabachnick and Fidell (1989), the sample size of events should exceed five times the number of cells, and 80% of the cells should have an expected probability of greater than 5. This means that for longer sequences enormously large amounts of data are required to achieve valid results.

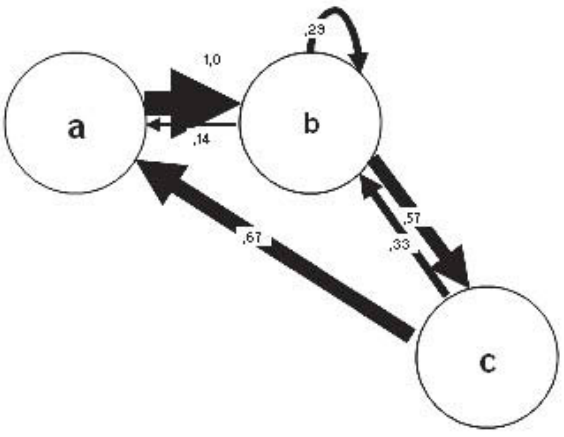
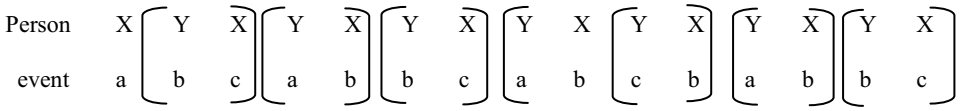


Figure 5.2 Transition state diagram for event sequences in (1).

Lag-sequential analysis is easy to use and flexible. It may be used for different types of sequences. The example shown above used overlapping sequences, in which the window slides in such a way that each event is observed twice: first as a prior event, and then as a succeeding event. But nonoverlapping chains are also possible. If, in our example, the script provided different roles, and one student acted as a mentor (X) and the other as a mentee (Y), we could analyze how the mentor reacts to the mentee. In order to tally the appropriate chains, we would have to use nonoverlapping windows.



This would lead to a frequency matrix which is different from the one shown in Table 5.1.

Furthermore, not just a lag of 1 may be of interest, but also events that follow later. We can assume, for example, that with a larger group, a mentor will not be able to answer all questions immediately. The mentor's answers may follow after two, three, or even more events. This would require lags of 2, 3, or more in our analysis. For this purpose, a moving window with varying sizes of lags might be considered, and it could be tallied how often the tutor (mentor) answers within the next 2 or 3 events after a question.

These different examples show that a specific research question defines the way of tallying the relevant chains of a sequence, in order to obtain the adequate frequency matrix. It strongly depends on the research question and the underlying theory which type of sequence chain is of interest. For this reason, lag-sequential analysis is a very typical statistical procedure for the hypothetico-deductive model, where first a theoretical model is stated and, accordingly, the appropriate operationalization is found.

With lag-sequential analysis the use of inferential statistical may be problematic. Early articles have pointed out that binomial *z scores* can be used to test individual pairs of events for significance (Bakeman, 1978; Sackett, 1979). In this sense, in Table 5.2 and Figure 5.2, *z-values* could be calculated for all nine transition probabilities. This would make it possible to detect if some of them occur more frequently than would have been expected by chance. Several more recent articles have expressed some concern about such an exploratory procedure (Allison & Liker, 1982; Bakeman & Quera, 1995; Faraone & Dorfman, 1987), because the use of too many binomial tests may lead to too many significant results, which occur only by chance simply because of the large number of significance tests.

Table 5.3 Frequency Matrix for the Mentor's Teaction on the Mentee's Actions

Y	Succeeded by X's			
	Question (a)	**Answer (b)**	**Statement (c)**	**Total**
Question (a)	0	3	0	3
Answer (b)	0	0	2	2
Statement (c)	0	2	0	2
Total	0	5	2	7

Bearing in mind this problem, the use of *log-linear models* has become common for some types of research. These provide a whole-table view and allow the use of omnibus tests instead of many binomial tests (Bakeman & Quera, 1995). Log-linear models are the standard procedure for analyzing multidimensional contingency tables. As the frequency matrixes of lag-sequential methods are, in fact, such contingency tables, log-linear models are well suited to analyze event sequences. Log-linear models are rarely used in CL research, but we will present their basic principles here.

Log-linear models are ANOVA-like procedures, which interpret the rows and columns of a contingency table as factors. Log-linear models define the observed cell frequencies as a function of these factors. In a two-way table, we can model cell frequencies in terms of the average count (grand mean), a column effect, a row effect, and row-by-column interactions. In a first step, we have to specify a model, which means we have to specify if the observed cell frequencies are explained by just the grand mean, additionally by a row-and-column effect, or additionally by an interaction between both. Then the cell frequencies, which are expected under this model, are calculated. A statistical test (Pearsons chi-square or likelihood-ratio chi-square) then analyzes how well this model fits the observed data. The log-linear procedure compares different models and their goodness-of-fit value, in order to find the least complex model that, nonetheless, achieves a good fit to the data.

Let us explain this with the data from Table 5.1 (even if the very low cell frequencies make these data not very suited for the use of log-linear modelling). Let us name the *observed* cell frequencies f_{ij}, with i describing the rows and j describing the columns. Different models provide *expected* cell frequencies m_{ij}, which are then compared with the observed f_{ij} (which are the cell frequencies given in Table 5.1). For each model, the following Pearson chi-square evaluates how well the model fits the observed data.

$$X^2 = \sum_i \sum_j \frac{(f_{ij} - m_j)^2}{m_j} \qquad \text{Eq. 1}$$

The least complex model is the *null model*. It states that all cell f_{ij} frequencies differ only randomly from the grand mean. So under the null model, we would expect that questions, answers and statements are equiprobable and that all cells will just randomly differ from 14/9 = 1.56 (overall-total through number of cells). With Eq. 1 this leads to a chi-square of $X^2 = 20.45$ with df = 8. The high value shows that the model does not fit the data very well.

A more complex model would be one that states that the event types (questions, answers, statements) have different probabilities, but that succeeding events are independent from the preceding ones. This would allow two main effects (varying row totals and varying column totals), but no interaction. The expected cell frequencies of such an independency model can be calculated by the formula:

$m_j = \frac{f_{i+} f_{+j}}{f_{++}}$ with f_{i+} being the total of the regarding row i and f_{+j} the total of the regarding column j, and f_{++} the overall total. Table 5.4 shows the frequencies which would be expected under this model.

With Equation (1) this model give a chi-square of $X^2 = 10.7$ with df = 4. Even if this chi-square is lower than the one of the null model, also this chi-square is significant.

Table 5.4 Expected Frequencies under the Main Effects Model

	Succeeded by		
	Question (a)	**Answer (b)**	**Statement (c)**
Question (a)	0.9	2	1.1
Answer (b)	0.6	1.5	0.9
Statement (c)	0.6	1.5	0.9

This shows that this model does not fit the data very well either. This means that a model must be accepted here which assumes that rows and columns are associated. In our example, this would be true of the *saturated model*. Saturated models take into account all possible effects: row effects, column effects and interaction effects. In our example, the saturated log-linear model allows that the three rows have different totals (row effect), that also the three columns have different totals (column effect), and that all cells have different means (interaction). In a saturated model, the expected frequencies are identical with the observed frequencies, so it cannot be tested for significance. But because of the large chi-square of the independency model, we must assume that columns and rows are associated and that preceding events determine succeeding ones. In order be more precise and to determine if some type of events occur quite frequently after other types, the *adjusted residuals* may be used. This statistical method gauges the extent to which a specific f_{ij} differs from its expected value m_{ij}. So it is a kind of post hoc test after a model has turned out a significant result in a preliminary omnibus test (for a further discussion see Bakeman & Quera, 1995).

For tables of higher dimension we can specify more main factors and interactions. Then the log-linear models will become more complex, and it is the task of the researcher to specify in detail which sequence patterns are expected. As in the example, significance tests estimate the probability of the observed data under each specified model. The larger this probability is, the better will the model fit the data. An example of the use of log-linear models is the study of Marttunen (1997). He studied e-mails with counterarguments and examined the associations between four different features: *level of counterargumentation* (good, moderate, or poor), *time of sending the message* (first/second half of the study), and *mode* (discussion vs. seminar). He started from a saturated model in which all the possible main and interaction effects of the four variables were included. Then all those parameters that were not statistically significant were dropped from the model step by step by starting from higher-order terms, ending at a minimal acceptable model. This final model had a main effect of *mode* (more counterarguments in discussion mode than in seminar mode), and an interaction between *mode* and *level of counterargumentation*, showing that there is more good or moderate counterargumentation in discussion mode than in seminar mode.

Log-linear modeling can be done with standard procedures in SPSS or SAS. Lag-sequential methods may not be applied so easily, but Bakeman, Adamson, and Strisik (1995) and O'Connor (1999) describe relevant SPSS and SAS procedures, and specific software tools were developed by Bakeman and Quera (http://www2.gsu.edu/~psyrab/gseq/index.html). For lag-sequential analysis and the construction of a transition state diagram, Jeong provides the easy-to-use Discussion Analysis Tool (http://myweb.fsu.edu/ajeong/dat/).

All described methods for analyzing interaction depend on frequency tables, which aggregate the relevant event chains across the whole sequence. So the models assume that event sequences are homogeneous across time and will not systematically differ between periods of time. But homogeneity is not always what occurs during collaboration. Group work often will be carried out in different phases, in which people interact differently. An analysis of event sequences with lag-sequential methods or log-linear models requires data from a stationary process. If nonstationary processes with phases of different interaction patterns are considered, an inductive approach for defining the breakpoints that divide the stationary sequences is described by Chiu (2008).

Methods that Focus on Individuals

The unit of analysis which most psychological or educational studies address is that of individual learners. Whenever the (potential) influence of a tool, an instruction, or a feature of a learning environment on learners' behavior, beliefs, or performance is the topic of a study, it will have to focus on the individual learner as the relevant unit of analysis. But in CL research this may be a pitfall. CL has to do with learners who collaborate, so these learners are generally nested within groups. We have to be aware that individual learners within a group are not stochastically independent. They do interact and this will influence their behavior and learning results.

In our example (triads work with and without scripts) we may be interested in finding out if a learner who works with a script shows less off-task talk and achieves higher scores in the knowledge post-test than a learner without a script. But it would not be permissible to compare the means of all learners working with and without scripts by using a standard method like t-test or ANOVA. This is not possible because all learners worked in groups, so the group mates have influenced each other. The resulting stochastical interdependency can be measured by intraclass correlation. It describes the higher (or lower) similarity of individuals within a group, compared to the similarity of people who belong to different groups. This is identical with the proportion of variance in the outcome variable which is caused by group membership. If the intraclass correlation in a given data set is significant (for the use of different test see McGraw & Wong, 1996), it will be necessary to deal explicitly with the hierarchical data structure and use multilevel methods (see Janssen et al., this volume). It would not be possible in this case just to pool individual learners across all groups and compare their means. Standard methods, such as OLS Regression or standard Analysis of Variance, rely heavily on the assumption of independent observation. If these standard methods are used and individuals pooled across different groups, then the standard error is systematically underestimated, and this will lead to an alpha-error inflation with an overestimation of significant results (Bonito, 2002; Kenny & Judd, 1986).

That means that a study which has its focus at the individual level and aims at predicting individual behavior or individual learning outcome *must* consider group effects. One approach, which is common in social psychology but little used in CL, is to deal with the group effect in an experimental way and to hold group behavior constant for each individual in the experiment. This may be achieved by using confederates (stooges), who simulate group mates and act in exactly the same way vis-à-vis each subject. In virtual settings, in which the group members do not interact face-to-face, the behavior of the group mates can be staged easily: A person is then told to be part of a group, but all online activities of the other "group" members are simulated on the basis

of a standardized protocol. Then the individual subject behaves like being in a group, responds to the "group," but all the participants are statistically independent, because they do not influence each other. Simulating other group members eliminates group effects. In addition, the researcher can systematically vary the "group's" behavior as an independent variable and measure its influence on the behavior of individual learners. Using this method, Cress (2005) analyzed whether a group's sharing behavior influenced the individuals' behavior in a knowledge-sharing task, and Kimmerle and Cress (2008) showed that a group awareness tool was differentially effective with people in a group with a high and low degree of interpersonal trust. By simulating groups, it is possible to study aptitude treatment interactions, in which the aptitude variable is an individual-level variable and the treatment is a group-level one. Buder and Bodemer (2008) worked with simulated ratings of group members. Learners with a correct view of a Physics controversy, but in the role of a minority in a fictitious group, were confronted with a majority who presented a plausible but incorrect view of the matter. The experiment then compared learners working with and without a group awareness tool. In an analogous experiment, Dehler, Bodemer, Buder, and Hesse (2011) manipulated the knowledge distribution among dyads and measured the quality of a learner's explanation to questions of their learning partners. These questions were also faked.

But there are only few factors and short-term processes of social interaction that can be analyzed by experimentally controlling group effects. By faking activities of group members, group interaction is reduced to unidirectional effects from (simulated) teammates to target persons. *Real* interactions, however, will not only consist of unidirectional effects; people's behavior may be influenced by the group, but they also influence the group itself. This complex interaction cannot be considered by faking group interaction.

Another way of dealing with group effects is the actor-partner-interaction model (APIM), as proposed by Kenny and colleagues (Kashy & Kenny, 2000; Kenny, Mannetti, Pierro, Livi, & Kashy, 2002), but it has to our knowledge not been used in CL research so far.

Methods that Focus on Groups

CL is based on the assumption that learning in groups will not only influence the way in which learners interact, but also their learning results. We may even go as far as stating that CL research will naturally have to focus on groups as its unit of analysis. It assumes synergistic processes, which describe that what is going on in the group is not simply determined by individual characteristics of the group members; instead it proposes that the group situation introduces new processes and shapes the individuals' behavior and learning in specific ways. The interesting point in many cases is not the behavior of the different group members, but what happens with the group as a whole. Does the group construct new knowledge? Does collaborative meaning making happen?

For analyzing these processes at the level of the group, we can use all standard methods of inferential statistics, because different groups are stochastically independent. With t-tests or ANOVAs, we can test if groups from different experimental conditions differ significantly. As dependent variables we can use measures of the whole group, but also aggregated measures from group members, or events which have taken place within the group. In our example about scripting, possible dependent variables could be the learning time that dyads spent on a transfer group task (group measure), the mean of the

learners' performance in the posttest (aggregated measure), or the number of different events which occurred during interaction (pooled measure of event frequencies). At the level of the group, all aggregated lower-level measures can serve as dependent variables: frequencies, means, sums, standards deviations, or any other transformed measures about events or individual scores. We could, for example, describe the heterogeneity of a group (group level variable) by the standard deviation of the prior-knowledge test scores of the group members (aggregated individual level variable). If we do not have an experimental design, we may use correlations to describe unidirectional relations, and we can use structural equation models to model causal relationships in nonrandomized settings. So once the group is used as the unit of analysis, all statistical methods are open. But it should be mentioned that the number of units which define our degree of freedom here is the number of *groups*, not the number of individual learners or distinct events. This leads to the practical problem that we need many more participants than we would need for analyses which consider the individual learner as their unit of analysis. With dyads, we need twice as many learners, with groups of n, we need the n-fold number of learners to perform tests with adequate power. Looking at results of any group-level analysis, we have to be aware that they describe groups, not individuals. Such results cannot simply be applied to the individual level. If a study has determined, for example, that groups who are more active achieve higher learning outcomes, we may *not* conclude likewise that the same correlation exists at the level of individuals and that a learner's activity predicts his or her individual learning outcome. This failure of transferring group-level effects to individual-level effects is known as the Robinson effect (Robinson, 1950), and was one of the reasons for developing specific methods for handling nested data. We will briefly describe multilevel analysis in the next section.

An additional, and also more practical, problem may be that we are really wasting data if we have obtained many data from group members during experiments, but just analyzed them at the level of the group. So, if possible, we will prefer *multilevel methods*, which are able to consider more than one level (see chapter 6 of this volume).

Apart from classical methods of inferential statistics, which can be used for all levels of analysis (if the units are stochastically independent), one method is particularly useful for describing groups and becoming more and more common in both CL and CSCL research: *social network analysis* (SNA). This method is based on analyzing interactions between group members, and it visualizes and measures these relations. For example, the e-mail traffic within a group constitutes a network, in which people may be represented as nodes, and the number of messages exchanged between them may be represented by weighted links. Each group member can be described by his or her location within the network. SNA provides a variety of measures to describe the centrality of individuals: *degree centrality* describes how many connections one individual has; *betweenness centrality* describes the extent to which a person connects different parts of the network; and *closeness centrality* describes the mean shortest distances to other persons. *Network centrality* describes centralization of the network as a whole. De Laat, Lally, Lipponen, and Simon (2007) use SNA to describe the interaction pattern between learners in a learning task during three different phases. Kimmerle, Moskaliuk, Harrer, and Cress (2010) use a two-mode SNA to describe the coevolution of a wiki artifact and its authors (Cress & Kimmerle, 2008). This two-mode SNA is based on links between webpages as well as links between authors and webpages. The dynamic SNA depicts the dynamic process of the growing network of webpages and the activities of its authors. SNA is an

exploratory method, which allows looking at a network as a whole. It has frequently been used for purposes of evaluation (Martinez et al., 2006; Nurmela, Lehtinen, & Palonen, 1999), but, so far, not for testing hypotheses.

INTEGRATION OF DIFFERENT LEVELS

Statistical analyses may be conducted at each of four level of analysis (events, interactions, persons, groups), and the previous sections of this chapter have cited studies which focused on one of these levels in different ways. When we use data from individuals who interact in groups, we have to deal with a hierarchical structure of data. Events and interactions take place within groups, and the learners belong to groups. Learners will influence each other, shape each other's behavior, and influence each other's learning outcomes. Statistically speaking, learners are nested within their learning groups, and events are nested within learners. As we have emphasized before, data which describe events, interactions, or individuals are not statistically independent. With intraclass correlation, this independence can be measured, and if it is significant (see McGraw & Wong, 1996 for the respective tests), multilevel analysis will have to be applied. This method will be described in the following chapter of this handbook.

CONCLUSIONS

This chapter has shown that quantitative methods in CL vary according to the level of analysis. Different statistical methods are available to analyze events, event sequences, individual measures, or group measures. When we analyze events, it is clear that these are not independent. Lag-sequential analysis and log-linear approaches take this into account explicitly in the search for patterns between events. These patterns capture the independency of events. If we are interested in individuals, we can deal with the multilevel structure by eliminating group effects (for example, by faking a group) or by using multilevel analysis. Analyses at group level are much easier to perform because the groups are stochastically independent and we can use standard methods here—as long as the groups themselves are not nested within larger units (which may be the case in quasi-experimental designs, where we may, for example, work with learner groups from different schools).

In all cases, quantitative methodology works best, if the underlying assumption of the study provides a clear prediction of where the focus should be. The aim of considering all processes at all levels will only at first glance appear to be a good choice. If we try to take into account the full complexity of processes that take place in CL, then everything seems to be related to everything. For such a complex situation, there is no "good" method which would allow to separate systematic effects from all types of random effects. Only if we can work with clear predictions, will we be able to find an adequate method which allow for testing them. The clearer our predictions are, the better we can operationalize the theoretical constructs, and the easier it is to test them adequately.

REFERENCES

Allison, P. D., & Liker, J. K. (1982). Analyzing sequential categorical data on dyadic interaction. *Psychological Bulletin, 91,* 393–403.

Bakeman, R. (1978). Untangling streams of behavior: Sequential analyses of observational data. In G. P. Sackett (Ed.), *Observing behavior* (Vol. 2). Baltimore, MD: University Park Press.

Bakeman, R., Adamson, L. B., & Strisik, P. (1995). Lags and logs: Statistical approaches to interaction (SPSS version). In J. M. Gottman (Ed.), *The analysis of change* (pp. 279–308). Mahwah, NJ: Erlbaum.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, England: Cambridge University Press.

Bakeman, R., & Quera, V. (1995). Log-linear approaches to lag-sequential analysis when consecutive codes may and cannot repeat. *Psychological Bulletin, 118*(2), 272–284.

Berelson, B. (1952). *Content analysis in communication research*. Glencoe, IL: Free Press.

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

Boxtel van, C., van der Linden, J., & Kanselaar, G. (2000). Collaborative learning tasks and the elaboration of conceptual knowledge. *Learning and Instruction, 10*, 311–330.

Buder, J., & Bodemer, D. (2008). Supporting controversial CSCL discussions with augmented group awareness tools. *International Journal of Computer-Supported Collaborative Learning, 3*(2), 123–139.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Carmines E. G., & & Zeller, A. R. (1979). *Reliability and validity assessment* (Qualitative Applications in Social Sciences, Vol. 17). Beverly Hills, CA: Sage.

Chiu, M. M. (2001). Analyzing group work processes: Towards a conceptual framework and systematic statistical analysis. In F. Columbus (Ed.), *Advances in psychology research* (Vol. 6, pp. 1–29). Huntington, NY: Nova Science.

Chiu, M. M. (2008). Flowing toward correct contributions during group problem solving: A statistical discourse analysis. *Journal of the Learning Sciences, 17*(3), 415–463.

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

Cress, U. (2005). Ambivalent effect of member portraits in virtual groups. *Journal of Computer-Assisted Learning, 21,* 281–291.

Cress, U., & Kimmerle, J. (2008). A systemic and cognitive view on collaborative knowledge building with wikis. *International Journal of Computer-Supported Collaborative Learning*, *3*(2), 105–122.

Dehler, J., Bodemer, D., Buder, J., & Hesse, F. W. (2009). Providing group knowledge awareness in computer-supported collaborative learning: Insights into learning mechanisms. *Research and Practice in Technology Enhanced Learning, 4*(2), 111–132.

Dehler, J., Bodemer, D., Buder, J., & Hesse, F. W. (2011). Guiding knowledge communication in CSCL via group knowledge awareness. *Computers in Human Behavior, 27*(3), 1068–1078.

De Laat, M., Lally, V., Lipponen, L., & Simons, R. J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis. International *Journal of Computer-Supported Collaborative Learning, 2*(1), 87–103.

De Wever, B., Schellens, T., Valcke, M., & van Keer, H., (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers and Education, 46,* 6–28.

Dewiyanti, S., Brand-Gruwel, S., Jochems, W., & Broers, N. J. (2007). *Computers in Human Behavior, 23*(1), 496–514.

Erkens, G., Kanselaar, G., Prangsma, M. E., & Jaspers, J. G. M. (2003). Computer support for collaborative and argumentative writing. In E. De Corte, L. Verschaffel, N. Entwistle, & J. van Merriënboer (Eds.), *Powerful learning environments: Unraveling basic components and dimensions* (pp. 157–176). Amsterdam, Netherlands: Pergamon/Elsevier Science.

Everitt, B. S., & Howell, D. C. (Eds.). (2005). *Encyclopedia of statistics in behavioral science*. Chichester, England: Wiley.

Faraone, S. V., & Dorfman, D. D. (1987). Lag sequential analysis: Robust statistical methods. *Psychological Bulletin, 101*(2), 312–323.

Gottman, J. M., & Roy, A. K. (1990). *Sequential analysis—A guide for behavioral researchers.* Cambridge, England: Cambridge University Press.

Haake, J., & Pfister, H.-R. (2010). Scripting in distance-learning university course: Do students benefit from net-based scripted collaboration? *International Journal of Computer-Supported Collaborative Learning 1*(6), 155–175.

Hijzen, D., Boekaerts, M., & Vedder, P. (2006). The relationship between the quality of cooperative learning, students' goal preferences, and perceptions of contextual factors in the classroom. *Scandinavian Journal of Psychology 4* (1), 9–21.

Hmelo-Silver, C. E., Chernobilsky, E., & Jordan, R. (2008). Understanding collaborative learning processes in new learning environments. *Instructional Science, 36*(5–6), 409–430.

Holsti, O. (1969). *Content analysis for the social sciences and humanities*. Don Mills, ON: Addison-Wesley.

Hou, H.-T., Chang, K.-E., & Sung, Y.-T. (2008). Analysis of problem-solving-based online asynchronous discussion pattern. *Educational Technology & Society, 11*(1), 17–28.

Hou, H., Sung, Y.-T., & Chang, K.-E. (2009). Exploring the behavioral patterns of an online knowledge sharing discussion activity among teachers with problem-solving strategy. *Teaching and Teacher Education, 25,* 101–108.

Hron, A., Cress, U., Hammer, K., & Friedrich, H. F. (2007). Fostering collaborative knowledge construction in a video-based learning setting: Effects of a shared workspace and a content-specific graphical representation. *British Journal of Educational Technology, 38*, 236–248.

Jeong, A. C. (2003). The sequential analysis of group interaction and critical thinking in online threaded discussions. *The American Journal of Distance Education, 17*(1), 25–43.

Jeong, A. C. (2005). A guide to analyzing message-response sequences and group interaction patterns in computer-mediated communication. *Distance Education 26*(3), 367–383.

Jeong, A. C., & Joung, S. (2007). Scaffolding collaborative argumentation in asynchronous discussions with message constraints and message labels. *Computers and Education, 48*(3), 427–445.

Kanselaar, G., Erkens, G., Andriessen, J., Prangsma, M., Veerman, A., & Jaspers, J. (2002). Designing argumentation tools for collaborative learning. In P. A. Kirschner, S. J. B. Shum, & C. S. Carr (Eds.), *Visualization argumentation: Software tools for collaborative and educational sense-making* (pp. 51–74). London: Springer-Verlag.

Kanuka, H., Rourke, L., & Laflamme, E. (2007). The influence of instructional methods on the quality of online discussion. *British Journal of Educational Technology, 38*(2), 260–271.

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.

Keeves, J. P. (1997). *Educational research, methodology and measurement: An international handbook* (2nd ed.). New York: Pergamon.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99,* 422–431.

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.

Kimmerle, J., & Cress, U. (2008). Group awareness and self-presentation in computer-supported information exchange. *International Journal of Computer-Supported Collaborative Learning, 3*(1), 85–97.

Kimmerle, J., Moskaliuk, J., Harrer, A., & Cress, U. (2010). Visualizing co-evolution of individual and collective knowledge. *Information, Communication and Society, 13*(8), 1099–1121.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Martínez, A., Dimitriadis, Y., Gómez-Sánchez, E., Rubia-Avi, B., Jorrín-Abellán, I., & Marcos, J. A. (2006). Studying participation networks in collaboration using mixed methods in three case studies. *International Journal of Computer-Supported Collaborative Learning, 1*(3), 383–408.

Marttunen, M. (1997). Electronic mail as a pedagogical delivery system: An analysis of the learning of argumentation. *Research in Higher Education, 38*(3), 345–363.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *International Journal of Computer-Supported Collaborative Learning, 2,* 63–86.

Nurmela, K., Lehtinen, E., & Palonen, T. (1999). Evaluating CSCL log files by social network analysis. In C. Hoadley & J. Roschelle (Eds.), *Proceedings of the Third International Conference on Computer Support for Collaborative Learning (CSCL'99)* (pp. 434–444). Mahwah, NJ: Erlbaum.

O'Connor, B. P. (1999). Simple and flexible SAS and SPSS programs for analyzing lag-sequential categorical data. *Behavior Research Methods, Instrumentation, and Computers, 31*, 718–726.

Poole, M. S., & Holmes, M. E. (1995). Decision development in computer-assisted group decision making. *Human Communication Research, 22*(1), 90–127.

Popper, K. R. (1963). *Conjectures and refutations*. London: Routledge & Kegan Paul.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.

Rourke, L., & Anderson, T. (2004). Validity in quantitative content analysis. *Educational Technology Research and Development, 52*(1), 5–18.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2000). Methodological issues in the content analysis of computer conference transcripts. *Journal of Artificial Intelligence in Education, 12,* 8–22.

Rummel, N., Spada, H., & Hauser, S. (2008) Learning to collaborate while being scripted or by observing a model. *International Journal of Computer-Supported Collaborative Learning, 4*(1), 69–92.

Sackett, G. P. (1979). The lag sequential analysis of contingency and cyclicity in behavioral interaction research. In J. Osofsky (Ed.), *Handbook of infant development* (pp. 623–649). New York: Wiley.

Shrout, P., & Fleiss, J. L., (1979) Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

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

Strijbos, J. W., Martens, R. L., Prins, F. J., & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computers and Education, 46,* 29–48.

Strijbos, J. W., & Stahl, G. (2007). Methodological issues in developing a multi-dimensional coding procedure for small-group chat communication. *Learning and Instruction, 17*(4), 394–404.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: Harper & Row.

Wolfe, J. (2007). Annotations and the collaborative digital library: Effects of an aligned annotation interface on student argumentation and reading strategies. *International Journal of Computer-supported Collaborative Learning, 3*(2), 141–164.