# The need for considering multilevel analysis in CSCL research—An appeal for the use of more advanced statistical methods

**Ulrike Cress**



Received: 7 May 2007 / Accepted: 9 November 2007


**Abstract** Per definition, CSCL research deals with the data of individuals nested in groups, and the influence of a specific learning setting on the collaborative process of learning. Most well-established statistical methods are not able to analyze such nested data adequately. This article describes the problems which arise when standard methods are applied and introduces multilevel modelling (MLM) as an alternative and adequate statistical approach in CSCL research. MLM enables testing interactional effects of predictor variables varying within groups (for example, the activity of group members in a chat) and predictors varying between groups (for example, the group homogeneity created by group members' prior knowledge). So it allows taking into account that an instruction, tool or learning environment has different but systematic effects on the members within the groups on the one hand and on the groups on the other hand. The underlying statistical model of MLM is described using an example from CSCL. Attention is drawn to the fact that MLM requires large sample sizes which are not provided in most CSCL research. A proposal is made for the use of some analyses which are useful.

**Keywords** Multilevel models · Hierarchical linear models · Quantitative analysis for CSCL

## Introduction

From its very beginning, CSCL has been an interdisciplinary field to which a broad range of methodological approaches have been applied. In addition to qualitative methods, quantitative methods also play a central role. Many empirical studies compare the effects of varying CSCL environments and analyse their influence on learning or interaction processes. In carrying out such analyses, researchers primarily use well-established methods such as ANOVAs or linear regression models. However, these standard methods do not always meet the special requirements of CSCL research. This paper aims to show

U. Cress (✉)
Knowledge Media Research Center, Konrad-Adenauer-Str. 40, 72072 Tuebinger, Germany
e-mail: u.cress@iwm-kmrc.de

that future CSCL research may have to broaden its focus and make use of more advanced statistical methods in order to deal better with the specific requirements of quantitative research in the field of CSCL.

In general, the use of collaborative learning scenarios is based on the claim that individuals can take advantage of group processes, and that collaboration and social interaction can facilitate learning. Collaborative learning as well as computer-supported collaborative learning thus explicitly takes the interdependency of individuals and their learning processes into account. Consequently, CSCL research has to deal with complex data sets which may contain variables characterizing features of the groups (e.g., the specific setting, the tools, the instruction or the circumstances surrounding learner interaction) and variables describing the individual learners (e.g., their prerequisites, their knowledge acquisition, and their perceptions). If CSCL research aims to analyze the complex interplay of learning settings, individual learning processes, individual outcomes and group outcomes, then it has to deal with the specific requirements of all these complex data.

## Problems occurring in the analysis of multilevel data

Researchers handling data of individuals interacting in groups are confronted with specific problems which can not be tackled using standard methods. The following prototypical example which will be used throughout this paper will describe such a situation.

The sample study aims to analyse the potential of a chat tool for collaborative problem solving in math. To this end, small groups of students discuss a mathematical problem in a chat environment with the task of jointly finding a solution. Each group member's activity during the chat is recorded (variable $X$). After the collaboration each student has to rate his/her satisfaction with this solution by answering a few questions (variable $Y$). The groups differ in their homogeneity (Variable $W$) measured by an index basing on the differences in the group members' maths grades.

A prototypical dataset is shown in Table 1. These data are used throughout the article. The small dataset of $n=17$ units is too small to calculate a real MLM, but it can serve as a prototype for illustrating relevant concepts of MLM.

In the study it is expected that a student's satisfaction (dependent variable) with the jointly found solution corresponds to her/his activity (independent variable). A standard method for describing such relationship between two variables $X$ and $Y$ is the use of a linear regression. With a linear regression a straight line is found on the basis of empirically given pairs ($x_i$, $y_i$), which presents the best estimate of $y_i$ (the estimated values are described with $\widehat{y}_i$) when $x_i$ is given (the index $i=1,\ldots,n$ describes the individuals). The resulting regression line is

**Table 1** Prototypical example of a multilevel dataset

| | Group A | | | | Group B | | | | Group C | | | | Group D | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Homogeneity $W_A=1.7$ | | | | Homogeneity $W_B=2.2$ | | | | Homogeneity $W_C=3.7$ | | | | Homogeneity $W_D=4.4$ | | | | |
| Activity X | 1.8 | 2.5 | 4.7 | 5.1 | 1.3 | 3.5 | 4.4 | 5.1 | 2.0 | 3.7 | 5.1 | 5.4 | 2.1 | 2.7 | 3.7 | 4.2 | 4.5 |
| Satisfaction Y | 1.3 | 1.1 | 0.9 | 0.7 | 1.8 | 1.8 | 2.4 | 2.0 | 4.0 | 4.5 | 5.8 | 6.1 | 4.7 | 5.1 | 6.4 | 6.5 | 7.1 |

The data are used throughout the article. Even if the dataset is too small to calculate a real MLM, it serves as an example.

described by Eq. 1, where $\beta_0$ presents the intercept (the expected $\widehat{y}$ when $x$=0), and $\beta_1$ presents the slope$\left(\Delta X \middle/ \Delta \widehat{Y}\right)$describing the increase of $\widehat{y}$ for x increasing to $x$+1. $\beta_1$ also is termed "regression coefficient", $e_i$ is the residual which is the difference from a predicted $\widehat{y}$ to an observed $y$. Thus $e_i$ describes the error of the prediction. Mostly $\beta_0$ and $\beta_1$ are determined by the ordinary least square algorithm (OLS). This OLS model estimates $\beta_0$ and $\beta_1$ in a way that the sum of the squared differences between the predicted $\widehat{y}_i$ to the observed $y_i$ is minimal

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, \ldots, n \tag{1}$$

Let us now focus on $\beta_1$ which shows the influence of activity on students' satisfaction. At first glance, various possible methods for analyzing are apparent:

1. Possibility: Students can be pooled and the linear regression of their satisfaction with their activity can be computed based on all 17 students without considering that they belong to different groups. This method ignores the fact that the students were parts of different groups. The analysis bases on $n$=17 observations and reveals $\beta_{1\_overall}$=0.35.
2. Possibility: Instead of using the individual measures, it is also possible to use the average activity measures and the average satisfaction of the four groups. This entails aggregating individual measures by calculating the averages of each group. In our example, the regression based on the averages reveals $\beta_{1\_average}$=0.03. This result would suggest that one's activity has almost no influence of her/his satisfaction.
3. Possibility: Regressions can also be calculated separately within each group. This once again provides a very different result: In group A it reveals $\beta_{1\_A}$=0.15, showing a small negative relationship, where the more active people are less satisfied. In the other groups we have positive but quite different correlation coefficients ($\beta_{1\_B}$=0.10; $\beta_{1\_C}$= 0.63; $\beta_{1\_D}$=0.99). These results demonstrate that even when both aggregated and pooled correlations are positive, this can not be assumed to be the case for the individual groups. In group A, activity and satisfaction are negatively correlated, indicating that less active individuals in this group are the most satisfied. But this negative relationship can only be observed when the linear regressions are calculated separately for each group.

This prototypical example illustrates the central problem with collecting data of individuals interacting in groups: Pooling individual data within the groups and handling the data as though they do not come from different groups may lead to results which diverge from those based on aggregating individual data within groups and using average values for each group, or which diverge from analysing the data for all groups separately. These different methods of analysis can lead to very different regression coefficients.

And there is one additional problem, having to do with the different sample sizes. All three variations listed above of calculating the regressions rely on different sample sizes. Thus, they would have different degrees of freedom when testing for significance, and regression coefficients of the same size would probably lead to different significance values.

The problems are caused by the hierarchical structure of the data. Such a hierarchical structure, as shown in Fig. 1, exists whenever a study deals with individuals who in turn are also members of different groups ("nested design"). The example described above comprises measures of individual students (e.g. activity and satisfaction), but these students are also members of different learning groups. It could further be the case that these groups are part of a third hierarchical level, for example, when the members of these learning groups belong to different universities. There could even be a level of measurement beneath

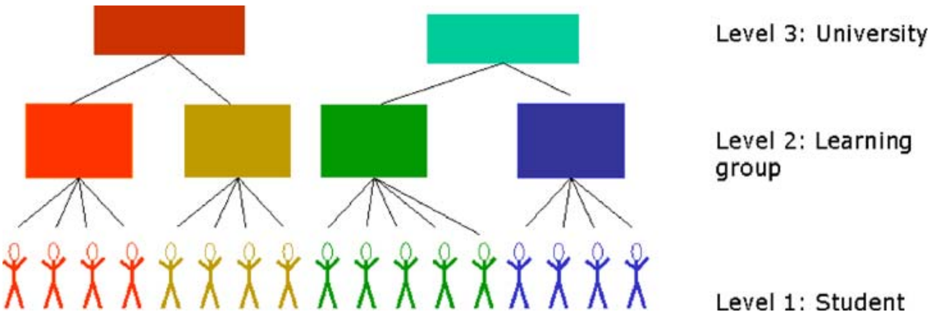




**Fig. 1** Hierarchical data structure

the level of the students, if we had repeated measurements for each student. Then these measurements would be nested within the students and they would provide another level of analysis.

A multilevel structure causes problems such as those described in the prototypical example if individual observations at the lowest level are *stochastically non-independent* and so the individuals are not independently distributed across the groups. This means that the members of a single group may be more (or less) similar to one another than members belonging to different groups. If one repeatedly drew pairs of students randomly, then the people within one and the same group would be more or less similar to each other than to those belonging to different groups.

Such stochastic non-independence can have three different causes: Compositional effects, common fate and reciprocal influences:

*Compositional effects* can occur when observations are similar *before* the study even begins. This can be the case when a CSCL study works with *real groups*, where the learners come from different school classes or different university courses. Compositional effects may therefore occur when it is not possible to randomly assign students to the groups. Due to this methodological aspect, compositional effects are also known as a “design effect.”

Even in randomized studies, however, stochastic non-independence can occur when group members share a *common fate*, which leads them to become increasingly similar over the course of the experiment. This occurs in most CSCL settings. If, for example, learners interact in small groups using a chat or forum, then only participants of a single chat group follow the same discussion. Only these learners are confronted with the same utterances and the same content of discussion; participants of a different chat group follow a different discussion and are confronted with different utterances. At the end of a chat discussion, members of different chat groups have therefore experienced quite different discussions, as a consequence of which only group members of the same group have equivalent conditions. Due to this “common fate” during the experiment, members of a single group become more and more similar than those belonging to different groups. The study described in our prototype example would have to take into account that this effect appears and thus provides statistical non-independence.

There is one further cause of stochastic non-independence. In CSCL, not only members of the same group share a common fate. If we aim to use CSCL settings to promote active interaction among group members, then we have to deal with *reciprocal influence*. This effect is obvious when learners interact in small groups. A single individual can determine the entire interaction process within the group. Just as a creative group member may stimulate the whole group to have an interesting discussion, an unmotivated member with destructive behaviour can destroy all motivation and any form of discussion among the other group members. In each case, learner behaviour is strongly influenced by fellow group members, and the same individual will behave quite differently according to the group to which he/she belongs. Such interactional and reciprocal influences between learners within groups further increase differences between members of different groups.

When comparing the importance of these three effects for CSCL research we should be aware that we can minimize the effect of composition (by randomization of the learners to the different groups), but we can not eliminate common fate and reciprocal influence. We are especially interested in *reciprocal influence* because it is not only unavoidable in CSCL, it is even explicitly intended. If CSCL is meant to stimulate collaboration and support learning by collaboration and interaction between the group members, then such reciprocal influence is desired.

Statistically, the non-independence caused by compositional effects, common fate and reciprocal influence can be measured using *intra-class correlations* (*ICC*). This correlation describes the higher (or lower) similarity of individuals within a group compared to the similarity of people belonging to different groups. It is equal to the average correlation between measures of two randomly drawn lower-level units within the same randomly drawn higher level unit. It can also be calculated by the proportion of variance in the outcome variable which is caused by group membership. If the *ICC* in a given data set is significant (for the use of different test see McGraw and Wong 1996), then it is necessary to deal explicitly with the hierarchical data structure. Standard methods such as the OLS-Regression or the standard Analysis of Variance heavily rely on the assumption of independent observations. If these standard methods are used regardless of a significant *ICC*, then the standard error is systematically underestimated. This underestimation results from the fact that the group composition, the common fate of group members and the effects of reciprocal influence lead to a higher similarity of individuals in the same group than similarity to those in different groups. With non-independence, the variance (which defines the standard error) within the groups will thus be smaller than it would be in groups formed from a stochastically independent sample. This underestimation of the standard error can lead to significant results which would have not achieved significance in a stochastically independent sample (Bonito 2002; Kenney and Judd 1986; Kenny et al. 1998). An alpha-error inflation thus arises in hierarchical data sets. This means that due to the low standard error, significance tests do not test against an alpha-error of 5%, as intended by the researcher, but at a much higher alpha-level depending on the respective ICC. Stevens (1996) showed that alpha-error strongly increases with increasing intra-class correlation and group size. For example, in comparing two conditions with a group size of 30 participants and an intra-class correlation of $ICC$=.30, alpha is equal to $\alpha$=.59. This shows that the alpha-error inflation can be enormously high.



## Some preliminary solutions to the multilevel problem

What is the solution to this problem? How can one correctly deal with hierarchical data? One possibility is to decide at which level the hierarchical data set is to be analyzed, and which level defines the appropriate units of analysis. If the units of analysis are the groups, then the analysis has to be based on aggregated data (i.e. means and standard-deviations of the individuals within each group). At the *group level*, correlations can be calculated between all kinds of aggregated values. Analysis is then, however, based on a much smaller number of units, because only the number of groups and not the number of individuals can be considered. This can be viewed as a waste of data (in our example we had to calculate with the measures of 4 instead of 17 units). A further problem with analyses at the group level is that they do not allow for predictions of processes and relations at the level of the individual. In our example, such a group-level analysis using aggregated measure could only investigate whether more active groups were more satisfied with their solution. Conclusions concerning whether individual learners who are more active are also more satisfied cannot be reached. The failure to distinguish between individual effects and effects found at the group level has been described by Robinson (1950) and has become known as the Robinson-Effect. Therefore, when the aim of a study is to predict individual learning and not the efficacy of a group as a whole, the problem posed by hierarchical data cannot be solved using aggregated data.

If a study focuses on the individual level and uses individual measures as units of analysis, then group effects must be considered and eliminated. A very strict way to do this is by controlling the group interaction in an experimental way. In an experiment we are able to hold constant group behaviour for each individual. This can be done, for example, by the use of trained confederates or by the use of bogus feedback. Then these controlled elements react in exactly the same way for all subjects. Thus, in such an experiment a subject acts as a theoretical part of a group, but there is no real interdependency between group and subject. Because the group's behaviour is faked and controlled, all variance is now caused by the subjects. Thus, by faking, we could eliminate all group effects or systematically vary the group's behaviour as an independent variable. This would be the only approach for a systematic variation of group influences. In this way, the individual level can remain as the unit of analysis. Whereas this strategy of faking has a long tradition in experiments in the field of social psychology, only few CSCL studies have adopted such an approach (e.g. Cress 2005; Kimmerle and Cress, 2007). This is due to the fact that very few factors and short-term processes of social interaction can be analyzed using this method, as a consequence of which the highly complex nature of real group interactions is ignored. By faking the actions of group members, the group interaction under investigation is reduced to a unidirectional effect from (faked) group members to a target person. The bidirectional effect, i.e., the fact that the target person's behaviour also affects the group members' reactions, cannot be considered using this method.

If a study does not intend to take such a reduced and experimentally controlled approach, a potentially effective method could be centring group members' values on the group mean or standardizing them within the group. The individual measure of each person then reflects the difference between his/her individual value and the group mean. While the intra-class correlation is now equal to zero, centring or standardisation within groups completely neglects existing group differences. In applying this method to compare different CSCL settings, a study would therefore only be able to show whether a setting is more or less effective for a person relative to the other group members, and not whether a setting is more or less effective for the average learner. Hence this method also cannot be viewed as a solution to the problem of dealing with hierarchical data.

A further possibility for setting up a model for group effects was proposed by Kenny and colleagues (Kashy and Kenny 2000; Kenny et al. 2002; see an application in Bonito and Lambert 2005) with the actor–partner-interaction model (APIM). This method explicitly takes into account reciprocal influences. The model proposes that a person is affected by his/her own standing on the predictor variable (actor effect), as well as by the average of all other members excluding that person (partner effect). In our example described above a person's satisfaction would be predicted by his/her activity and by the mean activity of his/her team mates. Thus the actor effect is separated from the partner effect and both are part of the prediction. The problem of this method is that, while it takes into account that a person's behaviour is influenced by his or her team -mates, it does not take care of what is motivating the team-mates to act as they do.

Burstein's slopes-as-outcomes approach (Burstein 1978, 1980; Burstein et al. 1989) points the way to an extensive solution for the multi-level problem. This method proposes that a linear regression of a variable $y$ on a variable $x$ in hierarchical data should allow for different groups having different slopes. These slopes represent the different covariances of $x$ and $y$ in the different groups. The method takes into account that the members of one group have equal conditions (are stochastically non-independent) and simultaneously allows different groups to have different conditions, as represented by differential regression functions for the different groups. Burstein's approach used differences in the slopes as outcome variable for a hierarchical analysis. Different slopes thus show different influences of group variables. Figure 2 depicts the linear regressions for the four groups of our example data.



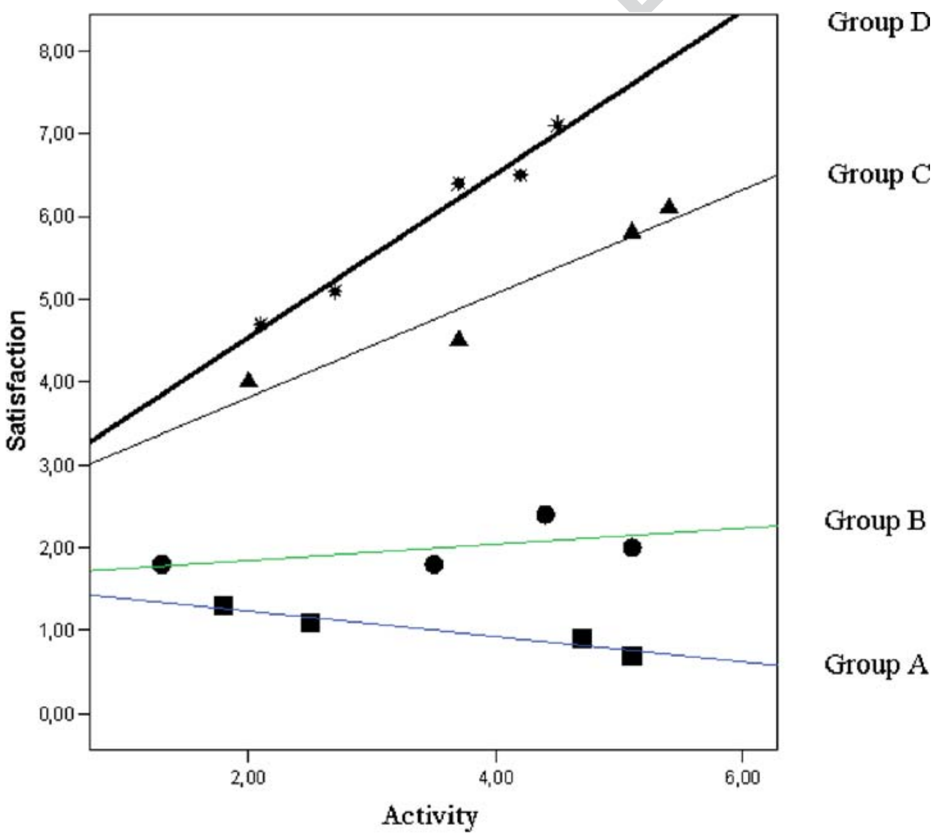




**Fig. 2** Example of the slopes-as-outcome approach for the dataset given in Table 1. There are different regression lines for the four groups. The observed cases of the different groups are marked with *different symbols*

**A short introduction to multilevel modelling (MLM)**

The slopes-as-outcome approach forms the basis of MLM (also called hierarchical linear model), as it was developed by Bryk and Raudenbush in 1992. MLM is also based on linear regression and extends it by allowing the data to be modelled at the group and individual level simultaneously. Instead of only one equation of a normal linear regression (as shown in Eq. 1) this extended model consists of a set of equations which form the linear regression model. The first of its equations (shown in Eq. 2) models the relation between an explanatory variable X and a dependent variable Y at the lowest level (Level 1).

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \tag{2}$$

Eq. 2 is a standard linear regression, with a regression intercept ($\beta_0$), a slope ($\beta_1$) and a residual $e_{ij}$. But in contrast to normal regression equations (shown in Eq. 1), there are two subscripts: the subscript $i=1,\ldots,n$ refers to the individual and the subscript $j=1,\ldots,k$ to the different groups. Eq. 2 thus allows differing regression functions with different intercepts and different slopes for each of the $k$ groups. This means that $\beta_{0j}$ and $\beta_{1j}$ are not constants as in normal regression models, but are variables and are different for each group $j$.

The variables $\beta_{0j}$ and $\beta_{1j}$ are explained by two further equations. These equations describe the processes at level 2. They aim to explain the variables $\beta_{0j}$ and $\beta_{1j}$ by introducing further explanatory variables at the group level. Such predictors (or explanatory variables) are described by $W$. In our prototype example, we could introduce the groups' homogeneity in their pre-knowledge as such an explanatory variable at the group level. Eq. 3 then describes the linear regression with group homogeneity as a predictor of the respective group's intercept, and Eq. 4 describes the linear regression with group homogeneity as predictor $W$ of the respective group's slope.

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \tag{3}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \tag{4}$$

These two linear regressions also have intercepts and slopes. These are described using $\gamma_{00}$, $\gamma_{10}$, $\gamma_{01}$ and $\gamma_{11}$ These gammas are constants with fixed subscripts. Both linear regressions (Eqs. 3 and 4) have residuals $u_{.j}$. They represent the variance which is not explained by the predictor $W$. The residual is group specific, and in the model $u_{0j}$ and $u_{1j}$ are independent of the residuals $e_{ij}$ at the individual level and have a mean of zero. However, the covariance between $u_{0j}$ and $u_{1j}$ is generally not assumed to be equal to zero.

The full hierarchical linear model thus consists of the three equations: Eqs. 2, 3 and 4. Substituting $\beta_{0j}$ in Eq. 2 through Eq. 3 and $\beta_{1j}$ through Eq. 4 results in the following equation:

$$Y_{ij} = \left(\gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij}\right) + \left(u_{1j}X_{ij} + u_{0j} + e_{ij}\right) \tag{5}$$

Eq. 5 comprises two parts. The first part (first bracket) is fixed (or deterministic), with fixed regression coefficients $\gamma_{00}$, $\gamma_{10}$, $\gamma_{01}$ and $\gamma_{11}$. The second part (second bracket) is random (also called the "error part"). This part reflects the fact that group effects are random and that there is some variance which is not explained by the predictors. With this random part, the model assumes that the groups which are part of the study are a random sample of all possible groups. It is due to this random part that multilevel models are also referred to as "random coefficient models." The term $u_{1j}X_{ij}$ shows that the amount of variance which is not explained by the group predictors can vary across groups. This allows for heteroscendasticity, which is a term for the phenomenon that variances of the different groups differ. The homogeneity of variances is a necessary pre-condition for the use of many standard methods, and thus heteroscendasticity would not allow for the use of an OLS-regression.

Figure 3 visually presents this hierarchical regression model of the data given in Table 1. This dataset was constructed in a way that its gammas are $\gamma_{00}$=2, $\gamma_{01}$=0.8, $\gamma_{10}$= 0.4 and $\gamma_{11}$=0.3.

In contrast to Fig. 2 this visualization does not show the regression line for the four observed groups given in Table 1 (these groups would have $\widehat{W}_A = -1.3, \widehat{W}_B = -0.8, \widehat{W}_C = 0.7$ and $\widehat{W}_D = 1.4$ with $\widehat{W}$ describing the z-standardized value of $W$). Instead it shows the effect of a student's activity on her/his satisfaction in a group with a mean homogeneity $\widehat{W} = 0$, with a homogeneity which is a standard deviation above the tested groups $\left(\widehat{W} = 1\right)$, and the group with a homogeneity which is a standard deviation below all tested groups $\left(\widehat{W} = -1\right)$. According to the random part of Eq. 6, these groups do not result from a fixed effect (where $W$ would be varied as an independent variable by establishing three different groups with $\widehat{W} = 0, \widehat{W} = 1$ and $\widehat{W} = -1$). Instead, these groups are rather hypothetical and result from a distribution of groups with all possible values of $W$. From all possible groups, there are

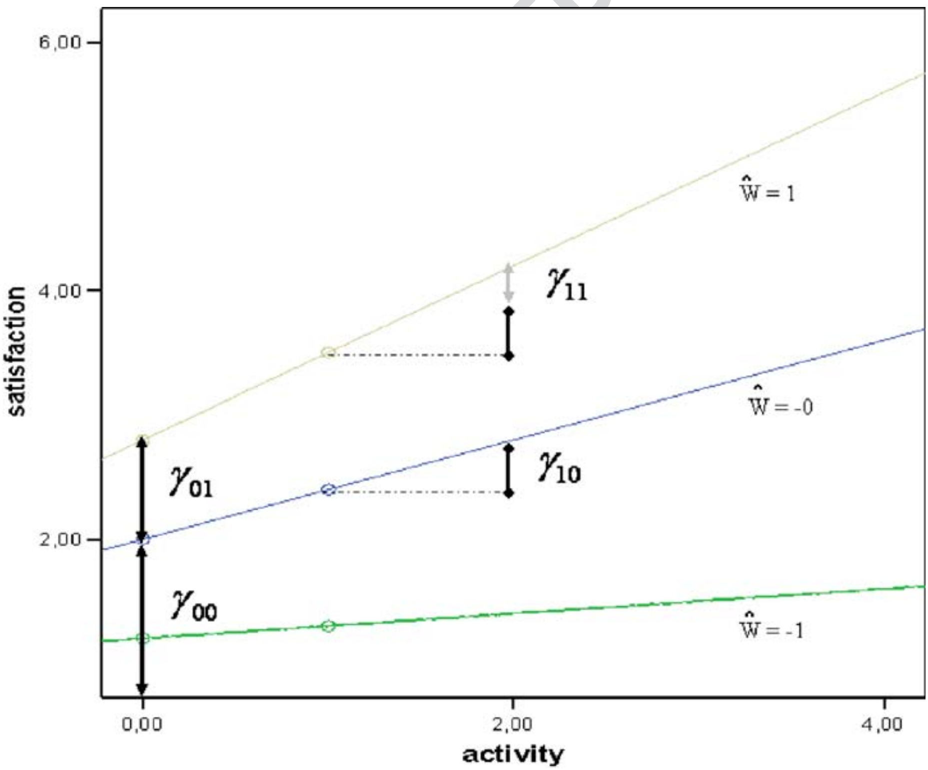




**Fig. 3** Visualization of MLM: The figure shows the regression lines of the group with mean homogeneity ($\widehat{W} = 0$), the groups with a homogeneity of 1 SD above the mean $\widehat{W} = 1$) and below the mean $\widehat{W} = -1$). It illustrates the meaning of the four gammas

three groups in this figure, with $\widehat{W} = 0$, $\widehat{W} = 1$ and $\widehat{W} = -1$. This makes clear that the regressions described in Eqs. 3 and 4 predict regression coefficients and intercepts for all possible $W$, not only for the $W$ given in the dataset.

Eq. 5 estimates the performance of a person i who belongs to the group j. The summands of Eq. 5 are visualized in Fig. 3 and can be described as follows:

$\gamma_{00}$ is the grand mean. It is the satisfaction of an individual in the group with a mean homogeneity of activity $\left(\widehat{W} = 0\right)$, given that this person shows no activity at all. Multilevel models often work with grand-mean-centred models, where $\gamma_{00}$ is zero (see Paccagnella 2006), since regression coefficients are easier to interpret.

$\gamma_{01} W_{\mathrm{j}}$ represents the influence of the homogeneity of activity within the group. The groups of different homogeneity differ in their intercepts. In Fig. 3, $\gamma_{01}$ represents the difference between a person belonging to the group with a homogeneity of $\widehat{W} = 1$ and a person of the group with an average homogeneity of $\widehat{W} = 0$, given that these people show no activity at all.

$\gamma_{10} X_{\mathrm{ij}}$ is the influence of the a student's activity, the explanatory variable at the first level. It represents the slope of the group with $\widehat{W} = 0$.

$\gamma_{11} W_{\mathrm{j}} \mathrm{X}_{\mathrm{ij}}$ represents the cross-level interaction, i.e., the different slopes between the group with homogeneity $\widehat{W} = 0$ and $\widehat{W} = 1$. With a higher $W$ (which also means a higher $\widehat{W}$) the slope is larger. This means that a group member's activity has a stronger influence on his/her satisfaction in homogeneous groups than in heterogeneous groups. Between the homogeneity $W$ and the slope of the linear regression at the first level, there is a linear relationship.

For purposes of clarity, the random parts of the model are not visualized in Fig. 3, although they will be described verbally.

$u_{1j} X_{\mathrm{ij}}$ is part of the random model and takes into account that the slopes cannot be perfectly predicted for each group, i.e., there is some residual in the prediction. This residual $u_{1j}$ can differ across groups, so that heteroscendasticity (different variances in different groups) is allowed. Standard methods including for example ANOVAs do not allow for heteroscendasticity, whereas MLM explicitly deals with and models it. In Fig. 3 this random part of the model would cause the regression slops to be not exactly determined by the gammas.

$u_{0j}$ describes another random part of the model, relating to the residual in the prediction of the groups' regression constants. This means that the explanatory variable at the higher level, $W$, does not perfectly predict the intercepts and that some unexplained error variance remains. This residual is the same for all individuals of the same group. $e_{\mathrm{ij}}$ is an individual specific residual showing that not every person's measure lies directly on the individual's respective regression line.

## Testing the multilevel model

This full hierarchical model is highly complex. Because of the sparsity of theory and data, Hox (2002) suggests that the model be tested using an iterative procedure with five steps.

The *first step* is the intercept-only model (also referred to as "null model" or "empty model"). It includes no explanatory variables at the individual or the group level. The intercept-only model does not explain any variance, but only reveals the proportion of variance caused by the groups.

The intercept-only model is given in Eq. 6.

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{6}$$

In our prototypical example, the model could show whether people's satisfaction depends on the group they belong to. The model is a one-factorial ANOVA with the random factor *u* describing the different groups. This model allows for calculation of the *ICC* which is presented in Eq. 6.

$$\mathrm{ICC} = \frac{\mathrm{Var}(u_0)}{\mathrm{Var}(u_0) + \mathrm{Var}(e_{ij})} \tag{7}$$

In Eq. 7, $u_0$ describes the between-variance on level 2.

Only if the *ICC* is significant must a multilevel model be used. So, if the *ICC* is not significant, we can apply a standard regression without any concern, because there is no group effect in the data.

The *second step* includes the lower-level explanatory variable *X* as fixed variable (i.e., the variance components of the slopes are constrained to zero). This results in the following ANCOVA model with the covariate *X* and a random group factor u:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + e_{ij} \tag{8}$$

In our prototype example this model would predict people's satisfaction by their activity during the chat, and it would take into account that the students are members of four different groups. So it would consider the group effect as a fixed effect. This would allow us to say that the four groups differ, but it would not allow us to make any prediction about groups with other homogeneity than the four measured.

If this model has a significantly better fit than the intercept-only model (which can be tested using a chi-square test), then in a *third step* a model can be chosen which includes the explanatory variables at the group level.

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_{1j} + u_{0j} + e_{ij} \tag{9}$$

In our prototype example we could now additionally predict the different average satisfaction of the groups with the homogeneity of the group (*W*). This would allow us to test if homogeneous groups are in general more satisfied than heterogeneous groups.

The *fourth step* allows for varying slopes in the different groups, as so it is also called "random coefficient model".

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_{1j} + u_{1j}X_{ij} + u_{0j} + e_{ij} \tag{10}$$

In our example this model additionally allows the regression coefficient from satisfaction to activity to be different for the four groups.

In the *fifth step*, a cross-level interaction between the explanatory group level variable W and the individual level explanatory variable *X* is introduced. This enables the different slopes of the groups to be predicted by the group level explanatory variable.

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_{1j} + \gamma_{11}W_{1j}X_{ij} + u_{1j}X_{ij} + u_{0j} + e_{ij} \tag{11}$$

In our prototype example we could now predict the different regression coefficients in the groups with the homogeneity of the group. We could, for example, state that the more homogeneous a group is, the stronger (or the weaker) the influence of one's activity is on her/his satisfaction.

This iterative procedure demonstrates that even when data result from a hierarchical structure, it may not always be necessary to use the full hierarchical model shown in Eq. 5. Less complex models with fewer coefficients are often sufficient. But if we find a significant *ICC* then we have to determine if one of those models is necessary.

The model described thus far is a complex model with two levels and one explanatory variable for each level. According to the experimental design, larger or smaller models can also occur. For example, the appropriate equation for a two-level model which does not include any explanatory variables at the lower level would be:

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + u_{0j} + e_{ij} \tag{12}$$

This model is an ANOVA model with a random effect and can also be calculated using standard software such as SPSS. In our prototype example such a model would be appropriate if we would like to provide a model for the different groups' different effects on students' satisfaction and if we would like to predict these effects with the group homogeneity W.

Of course, it is also possible to calculate models with more than one explanatory variable at the first or the second level. Such models can be found in the MLM literature (e.g., Raudenbush and Bryk 2002; Hox 2002; Snijders and Bosker 1999).

## Hierarchical models in CSCL research

Over the course of the last few years, multilevel models have become part of standard research procedure. A search for the terms "multilevel" or "HLM" in the database PsychInfo reveals that the very first articles appeared in the eighties and that the number of articles has greatly increased to more than 350 over the last five years. In modern educational psychology, hierarchical methods have especially gained a strong position through large-scale studies in the context of evaluating educational systems. In studies such as OECD-PISA, which compare educational systems in different countries, it is obvious that data are nested (learners in classes, classes in schools, and schools in school systems or in countries). A nice example of such a multilevel study can be seen in the work of Marsh and Hau (2003) who evaluated the data of 100,000 students in 4,000 schools, distributed across 26 countries. In this study, the extraordinarily large amount of data permits the analysis of an interesting interaction which considers all three levels: an interaction effect between the selectivity of a school system and the individual self-concepts of the learners in classes with different performance levels (the so-called "big fish little pond effect"). Such an effect can only be addressed by means of MLM. If a study aims to investigate cross-level interaction effects, then an adequately large sample is required, although it is not always necessary to have so much data at one's disposal as, for example, Marsh and Hau (2003). In her simulation studies, Kreft (1996) states that a two-level model requires approximately 30 groups of 30 individuals, 60 groups of 25 individuals or 150 groups of 5 individuals in order to test for cross-level interaction with adequate power. An adequate study should therefore be based on a minimum of approximately 1,000 individuals. Kreft found a rapid decrease in statistical power when the sample size falls below this threshold and a high risk of failing to detect existing cross-level interaction effects. In their simulation studies Maas and Hox (2005) found evidence that such enormous sample sizes are not needed. But they state that a small sample size, especially in level two (less than 50), leads to biased estimates of second-level standard estimates. In simulations with only ten groups they found a bias up to 25%.

This represents a problem within CSCL research, where sample sizes are for the most part considerably smaller. CSCL research often deals with small groups (mostly groups of between two and twelve learners) and studies often do not have the capacity to work with as many groups as would be necessary according to the simulation studies discussed above. In current research on collaborative learning, the predominantly small group sizes thus seem to be a dead-end for the application of MLM. With small sample sizes at the group level, the potential for detecting group level effects and the confidence of the estimated regression coefficient values are low.

Nevertheless, some authors have begun to use multilevel models in CSCL research despite small sample sizes. In the following section, the studies which deal with group influences in collaborative learning will briefly be described and their results with regard to group effects will be summarized.

Strijbos et al. (2004) investigated the effect of roles on group effectiveness in CSCL with 10 groups of approximately four learners each. Strijbos et al. (2007) used a different sample of 13 groups. Piontkowski et al. (2006) studied the effect of a sequencing chat tool based on the participation of 40 groups of three learners each. All three studies found significant intra-class correlations (*ICC* between.32 and.45) and were able to explain some of the group variance using second level factors.

Some studies show a more complex MLM where the dependent variable is measured repeatedly, and where these repeated observations nested within the students serve as the lowest level. For example, Schellens et al. (2005) used such a three-level model to predict learners' knowledge construction in asynchronous discussion groups. Data were collected on four measurement occasions (according to four discussion themes) for each of the 286 students, who where nested in 23 groups. The 3-level hierarchical model revealed significant influence of the student-level predictors (attitude toward the learning environment and engagement in the discussion groups), but no group-level effects.

The follow-up study of De Wever et al. (2007) holds a similar three-level design. Their data sets consist of 14 ten-person groups, with four measurement occasions each. This study confirmed the results of the previous one in revealing no significant group effect. In a four-level model with the levels "message," "theme," "student," "group," the "groups" and the "messages" had a significant effect. But an additionally provided comparison to a unilevel OLS model shows that most parameters, including the *p*-values, were quite similar so that OLS and MLM lead almost to the same conclusion. So, when focusing on possible group effect, the use of MLM would not have been necessary.

In Schellens et al. (2007), 230 students were assigned to 23 asynchronous learning groups to test the influence of student, group and task characteristics on students' final exam scores and their levels of knowledge construction. It revealed that only 6% of the overall variability in the final exam scores is explained by the group characteristics. So in this case also an MLM would not be necessary. With regard to knowledge construction the situation was different. Here about 19% of the variance was explained by differences among groups. Students in active groups which were active in discussion performed at a qualitatively higher level than those belonging to less active groups.

Chiu and Khoo (2003, 2005) analyzed the effect of rudeness and status on group-problem-solving with 80 people belonging to 20 groups. They used a three-level model with "speaker turns" as level 1, "time periods" as level 2 and "group" as level 3. They found significant effects of the group level which explained 12% of the total variance. But when the groups were divided into successful and unsuccessful groups no significant group heterogeneity remained. Thus, here also, MLM was not necessary in analyzing the group effect.

In sum, it seems too early to summarize the results of these studies. But it appears that the amount of variance explained by groups is rather small compared to the amount of variance which is explained through the lower levels of time periods or themes. So, even if in many of these studies the use of MLM could be criticized as inadequate in the case of such small samples sizes on the highest level, it seems nevertheless very important for empirical research in collaborative learning that the influences of the groups be considered explicitly. CSCL studies often implicitly assume that collaboration of learners has an effect, but the data do not always support this assumption. For testing it, MLM would be a potent method. But so far we do not have a clear picture about the biases MLM produces with small samples. For future research in CSCL it would seem desirable in some cases to apply different statistical means, in order to be able to compare their results. In its current state, our research is at the very beginning of a discussion of methodological issues for measuring the effect of collaboration and of establishing an adequate methodology (Snijders and Fischer 2007). Given that no satisfying solution to the multilevel problem in CSCL research has thus far been found, studies with much smaller samples sizes and their critical discussion may help to widen the focus of CSCL research and further direct attention to concurrent existing deficits in its methodology.



## Conclusion and suggestions for further CSCL research

Since CSCL research is explicitly founded on the claim that learning in groups can improve individual learning processes and enhance individual learning outcomes, efforts should be made to find a method which is adequate for testing and identifying such effects. Recent research has often been restricted to traditional methods which are not able to deal with the specific requirements of CSCL research. Some authors are aware of the multilevel problem and subsequently have decided to analyze the processes solely at the group level using exclusively aggregated data (e.g., Hron et al. 2000). This method is too superficial, however, when it comes to analysing the complex combination of individual processes and group influences involved in CSCL settings. Using groups as the unit of analysis is a waste of data and reduces quantitative analyses to a comparison of different CSCL settings without considering that learning is an individual process which, while taking place in a group, is primarily an individual cognitive process. It is precisely the analysis of this interaction between group influences and individual pre-requisitions which should constitute an important goal within CSCL research.

While a consideration of groups as units of analysis is unsatisfying, it is not acceptable to neglect the hierarchical structure of the data and analyze the individual data at the individual level without considering group effects. As shown in the prototype example above, this yields misleading results. Both authors and reviewers of journal submissions should be more aware of this problem. Data can only be analyzed at the individual level given that no significant intra-class correlation exists. This in turn, however, also means that the group has no effect. In dealing with CSCL data, MLM seems to be the method of choice. Intra-class correlations can be used to identify the effect of collaboration, and factors of the learning environment (instruction, tools, roles, content etc.) can be interpreted as mediators and included in a hierarchical linear model as second level predictors. The influence of the instruction, tools, or learning scenario can be modelled as a cross-level interaction. Even if MLM appears to be the optimal method for CSCL research, we must be aware that the enormous sample size required cannot be realized in many studies. Nevertheless, studies with small samples should also consider using multilevel models.

Such studies should report results obtained using traditional methods and those obtained with multilevel methods, in order to allow a comparison of the two. Additionally, future research should focus on simulation studies which make possible an estimation of how much power and reliability correlation coefficients lose in the case of small sample sizes.

As long as no optimal statistical methods exist for the analysis of small sample sizes, CSCL research should continue to attempt multilevel models, even though they may be imperfect. As a minimum standard in CSCL, the *ICC* should be calculated and tested for significance, whenever the sample size is large enough. If a CSCL setting does not produce a significant intra-class correlation, then the groups do not appear to have a systematic impact on people's learning. Indeed, in single groups there may be an influence on the learners, but this influence then remains unpredictable by variables describing the group.

In the case of a significant *ICC*, the slopes of the different groups can be compared if the study includes an individual-level predictor. If a study includes one or more group-level predictors, then the data can be analyzed with a random-coefficient model (ANOVA with varying instead of fixed factors), given that the groups' different intercepts are of interest. All of these methods can be used with smaller sample sizes and are adequate for many CSCL studies which do not apply a full hierarchical design with individual level predictors, group level predictors and cross-level interactions.

In general, CSCL research should address the hierarchical structure of its data in a more explicit manner. We might change our point of view so as not to interpret groups only as a source of unintended error variance, but we should also be interested in group effects and cross-level interactions as important outcome variables.

**Acknowledgments** The author would like to thank three anonymous reviewers for their helpful comments on an earlier version of this article.



## References

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

Bonito, J. A., & Lambert, B. L. (2005). Information similarity as a moderator of the effect of gender on participation in small groups: A multilevel analysis. *Small Group Research*, *36*, 139–165.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Burstein, L. (1978). Assessing differences between grouped and individual-level regression coefficients. *Sociological Methods & Research*, *7*, 5–28.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, *8*, 158–232.

Burstein, L., Kim, S. S., & Delandshere, G. (1989). Multilevel investigations of systematically varying slopes: Issues, alternatives, and consequences. In D. Bock (Ed.) *Multilevel analysis of educational data* (pp. 233–279). San Diego: Academic.

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 modeling. *Small Group Research*, *36*, 600–631.

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

De Wever, B., Van Keer, H., Schellens, T., & Valcke, M. (2007). Applying multilevel modelling to content analysis data: Methodological issues in the study of role assignment in asynchronous discussion groups. *Leaning & Instruction*, *17*, 436–447.

Hox, J. J. (2002). *Multilevel analysis: techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hron, A., Hesse, F. W., Cress, U., & Giovis, C. (2000). Implicit and explicit dialogue structuring in virtual learning groups. *British Journal of Educational Psychology*, *70*, 53–64.

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*. Cambridge, UK: Cambridge University Press.

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

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.) *The handbook of social psychology* (4th ed., Vol. 1, pp. 233–265). New York: McGraw Hill.

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 (2007). Group Awareness and Self-Presentation in Computer-Supported Information Exchange. *International Journal of Computer-Supported Collaborative Learning*. Q2

Kreft, I. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Retrieved March, 20, 2007 from http://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html.

Maas, C. J. M., & Hox, J. (2005). Sufficient samples sizes for multilevel modeling. *Methodology*, *1*, 86–92.

Marsh, H. W., & Hau, K. T. (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26-Country) test of the negative effects of academically selective schools. *American Psychologist*, *58* (5), 364–376.

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

Paccagnella, O. (2006). Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Evaluation Review*, *30*, 66–85.

Piontkowski, U., Keil, W., & Hartmann, J. (2006). Analyseebenen und Dateninterdependenz in der Kleingruppenforschung am Beispiel netzbasierter Wissensintegration. *Zeitschrift für Sozialpsychologie*, *37*, 41–50.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: applications and data analysis methods*. Thousand Oaks: Sage.

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

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.

Schellens, T., Van Keer, H., Valcke, M., & De Wever, B. (2007). Learning in asynchronous discussion groups: A multilevel approach to study the influence of student, group and task characteristics. *Behaviour & Information Technology*, *26*, 55–71.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis*. London: Sage.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Strijbos, J. W., & Fischer, F. (2007). Methodological. challenges in collaborative learning research. *Learning & Instruction*, *17*, 389–393. Q3

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*, *25*, 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.

# AUTHOR QUERIES

**AUTHOR PLEASE ANSWER ALL QUERIES.**

Q1. Snijders and Fischer 2007 was cited in the text but was not found in the reference list. Please provide complete bibliographic information.

Q2. Please provide update on the publication status Kimmerle, J. & Cress (2007).

Q3. Strijbos, J. W., & Fischer, F. (2007) was found in the reference list but was not found in the text. Please provide complete bibliographic information.