Cluster-randomized Studies in Educational Research: Principles and Methodological Aspects

Abstract

An increasing number of studies are being performed in educational research to evaluate new teaching methods and approaches. These studies could be performed more efficiently and deliver more convincing results if they more strictly applied and complied with recognized standards of scientific studies. Such an approach could substantially increase the quality in particular of prospective, two-arm (intervention) studies that aim to compare two different teaching methods. A key standard in such studies is randomization, which can minimize systematic bias in study findings; such bias may result if the two study arms are not structurally equivalent. If possible, educational research studies should also achieve this standard, although this is not yet generally the case. Some difficulties and concerns exist, particularly regarding organizational and methodological aspects. An important point to consider in educational research studies is that usually individuals cannot be randomized, because of the teaching situation, and instead whole groups have to be randomized (so-called "cluster randomization"). Compared with studies with individual randomization, studies with cluster randomization normally require (significantly) larger sample sizes and more complex methods for calculating sample size. Furthermore, cluster-randomized studies require more complex methods for statistical analysis. The consequence of the above is that a competent expert with respective special knowledge needs to be involved in all phases of cluster-randomized studies.

Studies to evaluate new teaching methods need to make greater use of randomization in order to achieve scientifically convincing results. Therefore, in this article we describe the general principles of cluster randomization and how to implement these principles, and we also outline practical aspects of using cluster randomization in prospective, two-arm comparative educational research studies.

Keywords: cluster randomization, structural equivalence, educational research, study, sample size calculation, statistical analysis

Jens Dreyhaupt[1]
Benjamin Mayer[1]
Oliver Keis[2]
Wolfgang Öchsner[2,3]
Rainer Muche[1]

1 Ulm University, Institute of Epidemiology and Medical Biometry, Ulm, Germany

2 Ulm University, Office of the Dean of Studies for Medicine, Ulm, Germany

3 University Hospital Ulm, Department of Cardiac Anesthesiology, Ulm, Germany

1. Introduction

An increasing number of studies are being performed to evaluate new teaching methods and approaches in educational research, particularly in the field of medicine [1], [2], [3], [4]; the increase has been particularly noticeable in Germany in recent years [5]. These studies should also comply with recognized standards and methods of scientific research. The methods of clinical studies in drug development are well developed [6], [http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html], and the procedure is standardized worldwide to ensure that meaningful study results are achieved. These standards should therefore be established and applied as far as possible also in educational research studies. In addition to being important for observational studies, these standards are relevant mainly for prospective two-arm (intervention) studies that aim to compare two different teaching methods.

Studies in clinical research require a control arm that is statistically comparable to the test arm (structural equivalence) in order to prove the efficacy or superiority of a treatment. Randomization ensures that all confounders – known and unknown – are distributed equally across the control and test arms and that only random differences, if any, are present at the start of treatment [7]. In the context of clinical studies, randomization does not mean that the patient is simply assigned to a treatment without any obvious criterion, but that a random experiment is formally performed that is independent of the participating clinician. If any other procedure is used to allocate volunteers or patients to the control and test arms, systematic distortions of the results are likely. Randomization is also an accepted method for educational research studies [8].

Educational research studies have some features and special requirements that are relevant for randomization. One of these features is that it is often not meaningful or possible to randomize individual students, for example because students are not always available as a result of their individual semester schedules. Furthermore, studies are usually performed in the natural learning situation of teaching in groups, such as joint lectures or seminars supervised by a lecturer. As a consequence of the above, in educational research studies it makes sense and is necessary to randomize whole groups (cluster randomization) rather than individual students.

In order to obtain scientifically convincing results in the sense of "evidence-based didactics," educational research needs to make more use of randomization in studies to evaluate new teaching methods and approaches. To help increase the use of randomization, this paper describes the principles of cluster randomization and explains practical aspects in order to facilitate its use in prospective, two-arm comparative studies in educational research.

The first section presents the rationale of randomization. Subsequently, the distinctive features of scientific studies in educational research are discussed. The third section first describes the principles of cluster randomization and then discusses sample size calculation and approaches to analysing cluster-randomized studies, both of which differ from those of studies with individual randomization. Finally, a sample design for an educational research study with cluster randomization is presented as an example of its application.

2. The reason for randomization: structural equivalence

In order to show why randomization is necessary, we will first explain some concepts that are relevant to educational research studies.

Necessity of a control arm: In educational research, if a new method is used in a study with only one study arm (test arm), one cannot rule out that an observed effect could have arisen also without the new method. Thus, not all observed results in the test arm can be attributed to the new method. The efficacy of a new method only becomes evident if the specific result is better with this method than without it [9]. Therefore, to obtain proof of efficacy a control arm, in which the previously used method is applied, is essential.

Statistical comparability: Ensuring that differences in results observed when comparing the control and test arms really can be attributed only to the new method requires statistical comparability of the following:

1. the structure of the control and test arms (structural equivalence);
2. the interactions with students, with the exception of the specific new method being evaluated (equivalence of treatment conditions); and
3. the observations (equivalence of observations).

In this context, "statistically comparable" means that the control and test arms differ as little as possible and at the most by chance.

Structural equivalence: Structural equivalence is given when the composition of the test and control arms is statistically comparable with respect to potential so-called "confounders" [10]. Examples of such confounders are age and gender, which often influence results. In educational research, additional factors are personality factors, such as educational background and special skills, as well as interests and activities. Randomization is one approach to achieve structural equivalence. Random allocation of students to the control or test arm allows one to achieve a similar distribution of known and unknown confounders across both arms, or at least to assume that it has been achieved. In addition, stratification can be used to balance (a few) important known confounders between the control and test arms (see Section 4.2); however, the feasibility of stratification is limited in educational research.

Equivalence of treatment conditions and observations: Equivalence of treatment conditions is achieved if all students in all groups receive and experience the same treatment conditions, apart from the new method being assessed. These treatment conditions (e.g. the same time for seminars, the same conditions for written examinations) should be specified in as much detail as possible in a study protocol – which should be written also for educational research studies – so that when the results are interpreted it is clear "what" was compared and under what conditions the observed effect arose. The control and test arms are considered to have "equivalence of observations" when situations are always observed and assessed according to the same rules (standardized conditions), e.g. the same evaluators are used to grade examinations. In clinical research, one procedure to achieve equivalence of treatment conditions and observations is the masking (blinding) of participants and investigators as to the treatment being received. In educational research studies, however, blinding is usually not possible and is only conceivable for evaluations, e.g. results in problem-based learning (PBL) can be evaluated by independent people who are not involved in the study.

If an educational research study has the above mentioned characteristics and if the success in the test arm is that much greater than in the control arm that the difference cannot be a coincidence, the study has shown that only the new method can be responsible for the outcome. However, if the control and test arms differ in other respects (i.e. they do not have structural equivalence or equivalence of treatment conditions or observations), the findings are difficult to interpret because an observed effect can no longer be exclusively attributed to the new method (blended effects). In such a situation, the specific effect of the new teaching method cannot be calculated and the study results usually cannot be interpreted according to the study question. Consequently, randomiza-

tion is an important instrument to achieve structural equivalence in studies of the efficacy of new methods in educational research.

3. Conditions and requirements of educational research studies

Educational research studies take place in a special context. They usually face a natural cluster structure, because teaching is usually performed in groups of students. Examples are seminar groups, PBL groups, and groups of students attending a joint lecture. The group size varies greatly and ranges from 5-8 (PBL groups) to over 100 students in a joint lecture. The size of seminars varies, whereas in Germany one can assume because of legal specifications that in medicine a seminar group has up to 20 students [http://www.gesetze-im-internet.de/_appro_2002/index.html]. All students within a group are exposed to the same conditions, such as the same teacher or the same facilities and times. Consequently, the intervention being studied (e.g. the use of a new teaching method by the lecturer) takes place at the level of the whole group. Thus, the outcomes of students in the same group are usually more similar than those of students in different groups. Furthermore, in addition to students' individual semester schedules an additional aspect to be considered in educational research studies is the time availability of appropriate resources (such as lecturers, seminar rooms, laboratories, lecture halls, computer pools). Also, if the number of students is limited and the group size is predefined, then a limited number of possible groups is available for a particular study. The above mentioned aspects mean that individual randomization can hardly be used in educational research studies, but that cluster randomization is meaningful and feasible. In cluster randomization, student groups or lecturers (who supervise groups of students) are randomized to the test and control arms, whereby the special characteristics mentioned above have to be taken into account. Various outcome variables are conceivable in educational research studies:

- Evaluation results (e.g. student satisfaction, self-assessment of the achieved competence)
- Results of course assessments/examinations (level of competence achieved)
- Measurement of the necessary learning effort (e.g. tracking of study time)
- Accompanying effects of studying the learning material (e.g. motivation curves, enthusiasm for the topic, awakening of interest, career planning)

In the following, we consider metric outcome variables (e.g. examination scores), which are assumed to be approximately normally distributed. We do not cover binary outcome variables (e.g. passed: yes/no) in this paper because studies examining such variables usually require a significantly larger sample size and are thus almost impossible to conduct in the field of educational research.

4. Cluster randomization and its use in educational research

The following sections explain the principles of cluster randomization and how they can be applied in the field of educational research. Furthermore, they provide information about study implementation, sample size calculation, and statistical analysis.

4.1. Definition and motivation

In a cluster-randomized study (also called a "group-randomized study," "community randomization study," or "community intervention study"), entire social groups or clusters of individuals are randomized, rather than individuals (see Figure 1).

To date, this type of study has been performed particularly to evaluate non-therapeutic interventions, such as training programs, prevention programs, and health promotion measures. For the reasons mentioned in the last section, it seems to make sense to use cluster-randomized studies also in the field of educational research. Published cluster-randomized studies include studies with both small and large clusters: clusters are defined frequently by households, families, neighborhoods, municipalities, school classes, employers, hospitals, and doctors' practices. Thus, the number of individuals per cluster can range from 2 to several thousands. Various cluster sizes are also conceivable in educational research (see Section 3).

The main motivation for conducting a cluster-randomized study is wanting to avoid or reduce a contamination bias. If individual randomization was used in an educational research study, such a bias (distortion or systematic error) could be caused by interactions between individuals in different study arms. For example, if students in the control arm were individually randomized they could easily be encouraged by students in the test arm to carry out the specific methods of the test arm whose efficacy is being tested. Cluster randomization does not entirely eliminate this risk, however, but only reduces it. The widespread use of social media, such as Facebook, plays a role in contamination bias in educational research studies. Another important reason for the use of cluster randomization in educational research is the existence of natural clusters, because learning is usually done in groups (see Section 3).

Table 1 presents important advantages and disadvantages of cluster randomization in the context of educational research.

4.2. Designs

In cluster-randomized studies, a distinction can be drawn between a completely randomized design, a stratified randomized design, and a matched design. In a completely randomized design, the clusters are assigned randomly to the groups and are neither stratified nor matched. An example is an educational research study

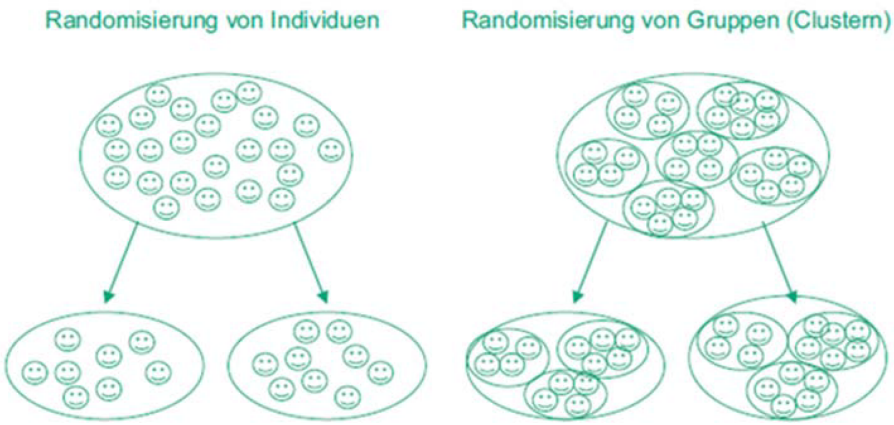

Figure 1: Randomization of individuals vs. randomization of clusters (reproduced from [15]).

Table 1: Advantages and disadvantages of cluster randomization in educational research studie (adopted from [15])

Advantages	Disadvantages
- Takes into account the natural teaching situation (teaching in groups) - Takes into account the conditions and requirements of scientific studies in educational research - Reduces contamination bias	- Usually requires (significantly) larger sample sizes than studies with individual randomization - Statistical dependence of students within groups: more complex statistical methods are required for sample size calculation and analysis

in which seminar groups are randomized into either the test arm or the control arm. In the stratified randomized design, randomization is stratified on the basis of (a few) important known confounders so that the distribution of the confounders is similar in the test and control arms. Stratification is performed according to factors that are strongly associated with the outcome variables, such as cluster size, gender, or day. An example is an educational research study in which the day on which a seminar takes place is assumed to influence the outcome variable (i.e. it is a confounder). In this case, stratification can first be performed according to the day (e.g. Monday/Wednesday/Friday) and then seminar groups can be randomized into either the test or the control arm within each day. In this way, the confounder "day" is distributed almost equally across both arms. In the matched design, pairs of clusters are formed that are as similar as possible with respect to important factors that affect the outcome variable. One cluster of the pair is randomized to the test arm and the other to the control arm. This represents a good way to balance confounders (e.g. characteristics from the baseline evaluation, such as gender, semester, previous grade) between the two arms and thus to make the arms comparable. Not too many criteria should be used for the matching, however, because it may then become impossible to find a cluster that can form a pair with another.

As a result of the conditions and requirements described in Section 3 (specifications for cluster size; limited number of students and consequently limited number of clusters; availability of resources; individual semester schedules), educational research studies often can be assumed to have a relatively small number of clusters with a more or less fixed cluster size. In this context, stratified and matched designs are probably only feasible under special conditions. An example would be a multicenter study conducted at different institutions. For this reason, the completely randomized design will prevail in educational research.

4.3. Practical implementation

Educational research studies usually have to be reported to the responsible ethics committee. However, often it is not necessary to obtain written informed consent from the participating students and sufficient just to inform them about the study [11].

Inclusion and exclusion criteria must be defined at both the individual level (students) and the cluster level (teachers). One problem is that blinding usually is not possible in educational research studies. Consequently, there is a risk of bias in the outcome variable. This risk should be countered by measures to achieve equivalence of treatment conditions and observations. Examples are a strong standardization of the general approach and perhaps a blinded assessment of the outcome, e.g. by a third evaluator who is not involved in the study and who has no knowledge of the respective student's assignment to the test or control arm.

4.4. Sample size calculation

4.4.1. Why is sample size calculation different?

Cluster randomization generates a special data structure, whereby observations within the clusters usually are more similar than observations from different clusters (i.e. there is statistical dependence). In the context of educational

research studies, this means that the results of students within the same seminar group (e.g. grades in the written exam) are more similar than the results of students in different seminar groups. This results in a loss of efficiency and power, which affects sample size calculation: the effective sample size of a cluster-randomized study (i.e. the number of truly statistically independent individual observations) is lower than the actual sample size (i.e. the number of recruited students). Therefore, standard procedures that assume the statistical independence of all observations are unsuitable for calculating sample size for cluster-randomized studies and evaluating data from these studies. The use of standard procedures to calculate sample size would lead to studies with too little power in which the chance of proving a difference between the study arms that is actually present would be (significantly) lower than assumed in the calculation. In educational research, this may result in a new teaching method that is actually better not being recognized as such by the study, for example.

4.4.2. Determination of similarity - the intracluster correlation coefficient (ICC)

The intracluster correlation coefficient (synonym: intraclass correlation coefficient, ICC; ρ) is used to quantify the similarity of observations within a cluster compared with observations from different clusters. The ICC can be defined in various ways [12]; for metric outcome variables, it is often defined as a quotient of variances [13], [14]:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} = \frac{\sigma_b^2}{\sigma^2}$$

whereby σ_b^2 is the variance between the clusters, σ_w^2 the variance within the same cluster, and $\sigma^2 = \sigma_b^2 + \sigma_w^2$ the total variance. This definition allows the ICC to be interpreted as the share of the total variance accounted for by the variance between the clusters, assuming that the variance σ_w^2 is constant in each cluster. With this definition, the ICC can take values between 0 and 1. Its size is a measure of the strength of the similarity of the observations within the cluster compared with the similarity of the observations between the clusters. If the ICC is 1, the observations within each cluster are the same. In the context of educational research studies this would mean, for example, that in each seminar group all the students have the same examination grade (but not necessarily that all seminar groups in the study have the same grade). The ICC has the value 0 if all observations are statistically independent. In the case of educational research studies, this would mean, for example, that the students' examination grades within the same seminar group are not dependent, i.e. the seminar group has no influence on the examination grades.
Estimating the ICC a priori is often a challenge. The ICC can be calculated from data from a pilot study or from the literature, for example. Therefore, publications of cluster-randomized studies should include the post hoc calculation of their ICC, so that it is available for similar studies [15], [16]. Furthermore, the ICC is only an estimate from a sample and thus subject to uncertainty (confidence interval [17]). This is of particular importance for educational research studies because often only small studies with few clusters can be performed, so that the ICC cannot be reliably estimated.
In addition, different calculation methods can influence the value of the ICC. An overview of ICC calculation methods suitable for metric outcome variables is given in [18]. For binary outcome variables, corresponding methods are available in [19] and [20].

4.4.3. The design effect (DE)

In order to achieve the same power in a cluster-randomized study as in a study with individual randomization, usually more individuals have to be recruited in the former. The sample size required for a cluster-randomized study is calculated by multiplying the sample size of a study with individual randomization with the design effect (DE), which is calculated from the ICC ρ and the fixed cluster size m:
DE = 1+ρ·(m-1)
The result is a total sample size and subsequently a number of clusters (with fixed cluster size) for a given power. For educational research studies, this means that initially a total number of students is calculated and then, on the basis of this, the number of seminar groups (with fixed group size m).
If cluster sizes are unequal, m can be replaced by the arithmetic mean or by the maximum cluster size. The use of the arithmetic mean of the cluster size is useful when there is little variability in cluster size [12]; the use of the maximum cluster size is a conservative approach. If the ICC is ρ=0 (statistical independence of the observations, see above), DE=1, which means that the cluster-randomized study has the same sample size as a corresponding study with individual randomization. In this case, the formation of clusters has no influence on the sample size. In practice, most ICCs are between 0.00 and 0.20, although there is a very wide range [21].

4.4.4. Procedure for calculating sample size

In general, two approaches can be considered when planning a study. One is an exploratory approach in which a minimum effect can be calculated for the given maximum sample size with a given power and cluster size or a power can be calculated for a given minimum effect and cluster size [22]. This is particularly useful when only a limited number of observations are available. Figure 2 shows the schema for calculating the power or minimum effect for a given sample size in educational research studies.
The other is a confirmatory approach: a sample size (i.e. the number of students and a resulting cluster number) is calculated for a given power and a predetermined

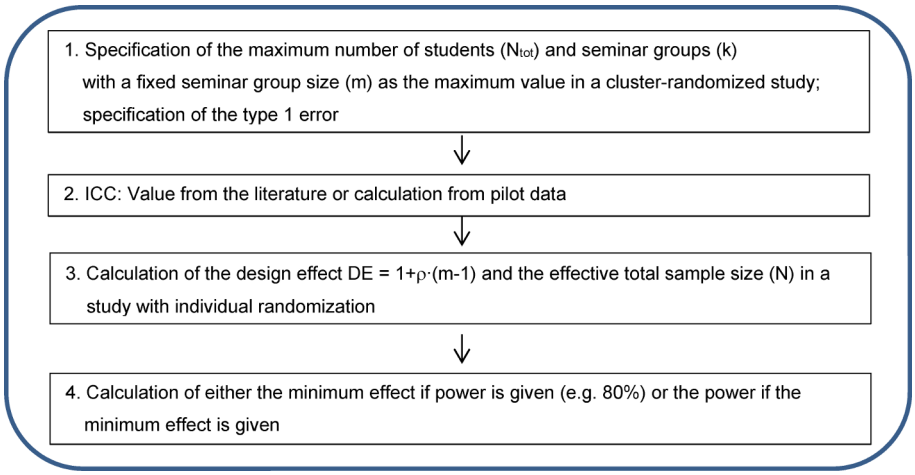

Figure 2: Schema for calculating power or minimum effect in educational research studies if the sample size is predetermined

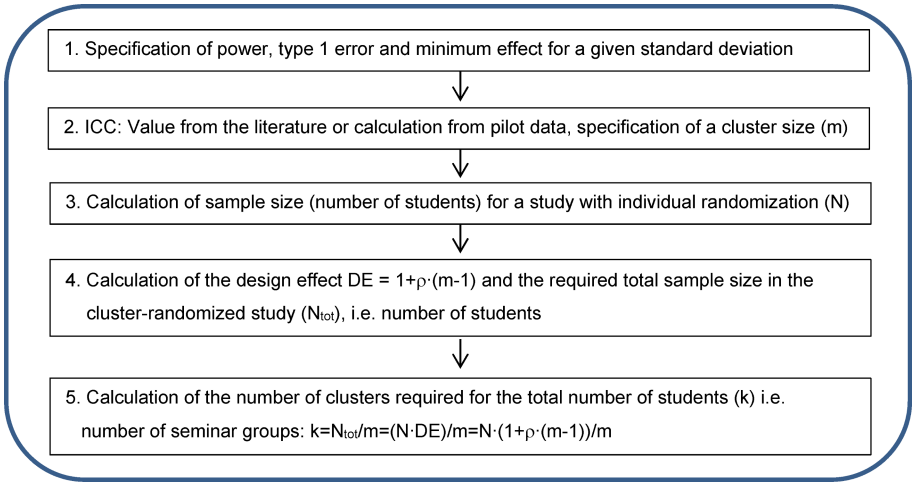

Figure 3: Schema for calculating sample size in educational research studies with a given power and minimum effect

minimum effect. Figure 3 shows the schema for calculating sample size in educational research studies with a given power and minimum effect.

However, because of the special conditions in educational research (limited number of students and therefore of clusters and a given cluster size, see above and Section 3) confirmatory studies often cannot be implemented.

If additional covariates are to be included in the planning of a cluster-randomized study, the definition of the ICC can be expanded according to [14]. Simulation is another option, especially for complex study designs (e.g. if several covariates are to be considered in a longitudinal design, presence of additional hierarchical levels) (e.g. [23], [24], [25], [26]).

4.5. Statistical Analysis

In order to take into account the statistical dependencies within the clusters (in educational research studies, an example would be the dependence of examination results of students in the same seminar group), a cluster adjustment has to be performed during the analysis [12]. A so-called "naive analysis" (cluster adjustment is not performed; standard methods are applied, such as a two-sample t-test) can result in the estimated confidence intervals and p values being too small [27], [28]. In educational research studies, this would result in studies being falsely reported as significant and thus new teaching methods appearing to be better than they actually are.

The methods used to plan a study should also be used to analyze it [23], [28], whereby the methods depend on the study design (see above). In the statistical analysis, a distinction can be made between the analysis at the cluster level and at the individual level [28], [13]. Because of the complexity of the statistical methods, the support of a competent expert (e.g. statisticians with appropriate special knowledge) is recommended in particular for the analysis. Almost all medical faculties in Germany that have a medical or dental school are connected with methodically versed institutes (such as biometry departments), which could provide expertise accordingly.

Cluster-level analysis is the simplest evaluation method for a cluster-randomized study and can be viewed as a two-step process: initially, a composite measure (cluster-level summary) is calculated for each cluster (first stage), and then the composite measures are compared with a suitable statistical test (second stage), see e.g. [16]. In educational research studies, for example, the mean cluster values (e.g. the mean grade in each seminar group) can be used in the analysis (e.g. ordinary two-

sample t-test) instead of the students' individual results. Covariates can be considered in a simplified way via regressions [13]. Analysis at the cluster level is robust, especially if the number of clusters is small, but has the disadvantage that it does not take into account the variability within the clusters. An alternative is to adjust univariate test statistics (e.g. the t value in the t-test) by considering the design effect, whereby the individual results may be evaluated as being statistically independent [15], [29].

Individual-level analysis is an alternative approach that is especially relevant for strongly varying cluster sizes, a situation in which cluster-level analysis is less efficient. The adjusted two-sample t-test is one simple statistical procedure that also allows an analysis to be performed at the individual level [28]. If additional covariates are to be considered, regression models with random effects, mixed effects regression models, or generalized estimating equations (GEE models) can be applied. These approaches also allow factors to be considered as potential influencing variables in the event that stratification was not possible during cluster randomization, even though known prognostic factors were present. The use of one of these approaches has advantages over using cluster-level analysis methods because the effects of covariates can be examined on the same level as the effect of the study arm (as a regression coefficient with a confidence interval and p value). Individual-level analysis methods have the disadvantage, however, that they are less robust when there is a small number of clusters. One recommendation, therefore, is to use cluster-level analysis methods if there are fewer than 15 to 20 clusters per study arm [13]. In studies with a larger number of clusters, individual-level analysis methods can have advantages, especially if cluster size is highly variable.

4.6. Reporting

The CONSORT Statement was developed for the reporting of randomized clinical trials but has been extended to cluster-randomized studies by Campbell et al. [30]. The extended CONSORT statement considers the special characteristics of a cluster-randomized study and should be considered when publishing such a study. The stipulations include the following:

- Describe the reasons for using cluster randomization
- Name the unit of randomization and the intervention
- In addition to the number of individuals, state the number of clusters and their size
- Show structural equivalence not only at the individual level but also at the cluster level
- Calculate and report the ICC (see above)
- Analyze the drop-outs on both the individual and the cluster level
- Draw a flowchart to show the number of study participants and clusters in the course of the study

5. Example of application

In this section, we will use an example to outline the planning, implementation, and analysis of a cluster-randomized study in educational research. The example is based on the NANA study [31], which is used to illustrate studies in clinical research. The study is conducted as a two-arm, prospective observational study and compares people with a sweet tooth ("NAschkatzen" in German) with people who like to nibble ("NAgetiere" in German) with regard to parameters such as the body mass index. The name of the study also is of relevance to Ulm University, which has a large NANA statue in front of it (see Figure 4).

Figure 4: NANA in front of Ulm University

The example cluster-randomized study is supposed to evaluate whether the use of a new "active seminar concept" (as part of the NANA study) influences the test results in the biometry education of medical students. The "active seminar concept" (planning, implementation, and analysis of a small empirical study during the seminar) is to be compared with the previous standard concept (working on practice exercises in the form of a "classical seminar"). For the study (balanced, prospective, cluster-randomized), whole seminar groups are to be randomized to either a test arm ("active seminar concept") or a control arm ("classical seminar"). The study is to be carried out at the medical faculty of Ulm University during a winter semester.

A maximum total of about 320 students can be assumed. The students are supervised in seminar groups, each with one lecturer and approximately 20 students. This results in a maximum number of 16 clusters (i.e. seminar groups) in the overall study (i.e. 8 clusters per study arm), making it a smaller study [32].

Figure 5 shows a possible result of the cluster randomization for the example study.

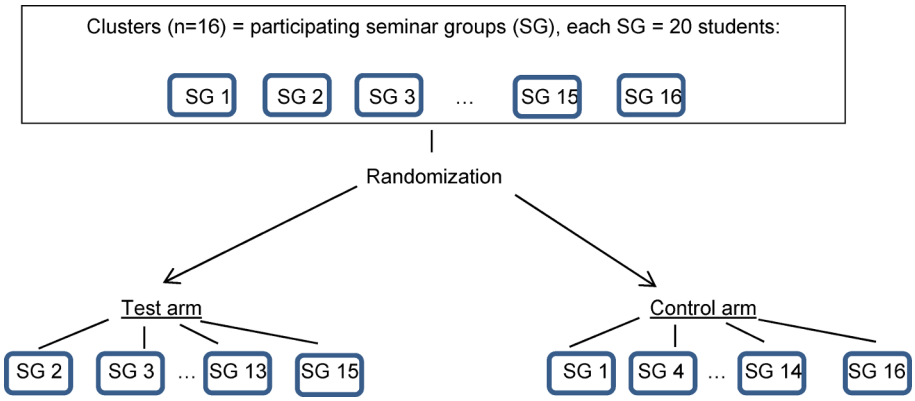

Figure 5: Sample result of cluster randomization for the sample study

The primary outcome variable is the number of points achieved in the examination, assessed for the individual students (i.e. on the individual level). Because the composition of the groups plays a role, in addition to the influence of the lecturer, we can assume that the examination results of the students in the same seminar group are more similar than the results of students in different seminar groups. The outcome variable is assumed to be metric and approximately normally distributed.

Pilot data: The results of the cohort from the winter semester 2015/2016 are available as pilot data: an arithmetic mean of 92.5 points (maximum score: 120) was determined with a standard deviation of 9.16 points (see Table 2).

The results of the pilot data are used for the control arm. For the test arm, the score is assumed to improve by a mean of 3 points (i.e. from 92.5 to 95.5). The ICC for the outcome variable "points" was estimated by a linear mixed effects regression model [12]: the model fit resulted in the variances σ_b^2 = 1.67 and σ_w^2 = 82.36, so that the ICC is estimated to be ρ=1.67/(1.67+82.63)=0.02. The design effect (DE=1+ρ·(m-1)) is thus calculated as DE=1+0.02·(20-1)=1.38, whereby m represents the mean cluster size (here: number of students per seminar group; m=20).

Study planning: Both of the approaches mentioned in Section 4.4 (exploratory and confirmatory approaches) will be applied below to calculate the sample size for the example study.

Exploratory approach: Calculation of the power or minimum effect for a particular sample size

The implementation of steps 1 to 4 of the schema shown in Figure 2 is described below for the example study. Assuming a maximum number of 320 students per semester, a maximum of 16 clusters (seminar groups with 20 students each) are possible in the planned study. For a design effect of 1.38 (calculation: see above), the maximum number of 320 students in the cluster-randomized study corresponds to an effective sample size of a maximum of approximately 320/1.38 ~ 232 students (even number because of 1: 1 randomization) in a study with individual randomization (i.e. 116 students per study arm). For this sample size, in order to achieve a power of 80% with a two-sided type 1 error of 5% the two-sample t-test requires a minimum difference of 3.4 points (with a standard deviation of 9.16), which corresponds with an effect size of 0.37 according to Cohen (small effect). If a difference of 3 points is assumed (original planning), a two-sided type 1 error of 5% will achieve a power of only 70% (with a given standard deviation of 9.16). Table 3 shows the impact of different sizes of the ICC on the minimum effect and power. The calculations were performed with the two-sample t-test, assuming the same variances in both arms.

The planning and implementation of the example study on the basis of this exploratory approach is pragmatic. This approach often appears to be more realistic than the confirmatory approach, because the latter often results in unfeasibly high sample sizes.

Confirmatory approach: Calculation of the sample size for a given power and minimum effect

The implementation of steps 1 to 5 of the schema shown in Figure 3 is described below for the example study. The outcome variable is assumed to be metric and approximately normally distributed. For the test arm, the score is assumed to change by a mean of 3 points (i.e. from 92.5 to 95.5, see above). The calculations are performed with the two-sample t-test, whereby the same variances are assumed in both arms (9.16, see above). For sample size calculation, a power of 80% and a type 1 error of 5% (two-sided) are assumed. First, this information is used to calculate the sample size for a study with individual randomization (e.g. [12]): the calculation shows that a total of 148 students per study arm (296 students in total) would have to be included in a study that randomizes individuals (i.e. does not consider the clustering). This number now has to be corrected by the design effect (DE=1.38, calculation see above): this correction indicates that 148·1.38~ 205 students would have to be included in each study arm (total study: 409 students, which would mean a total of approximately k=21 seminar groups). Table 4 shows the impact of the size of the ICC and seminar groups on the total sample size and the number of seminar groups for the above mentioned effects of the example study. The total sample size was rounded to the nearest whole number, and the number of seminar groups was rounded to the nearest even number because the example study will use a 1:1 randomization. This ap-

Table 2: Results from the winter semester 2015/2016 cohort: arithmetic mean and standard deviations of the score in the total group and in the individual course groups

Group	Arithmetic mean	Standard deviation
A1	90.5	6.86
A2	93.1	5.78
A3	94.6	8.97
A5	94.1	8.02
A6	88.7	7.77
A8	95.3	6.38
B1	90.4	16.92
B2	89.7	10.34
B3	96.4	5.38
C1	92.7	9.92
C2	92.3	8.47
C3	93.4	7.50
Total	92.5	9.16

Table 3: Impact of the size of the intracluster correlation coefficient (ICC) on the minimum effect (a) and power (b) for a predefined number of 320 students in 16 seminar groups with 20 students each. ESS = effective sample size (italics = study situation)

a					b			
ICC	DE	ESS	Minimum effect (for 80% power)		ICC	DE	ESS	Power (for minimum effect of 3 points)
0.01	1.19	270	3.2		0.01	1.19	270	77
0.02	*1.38*	*232*	*3.4*		*0.02*	*1.38*	*232*	*70*
0.03	1.57	204	3.7		0.03	1.57	204	64
0.05	1.95	166	4.0		0.05	1.95	166	56

Table 4: Impact of the size of the intracluster correlation coefficient (ICC) and the size of the seminar groups on the total sample size and the number of seminar groups in the overall study (italics = study situation)

ICC (ρ)	Size of seminar groups (m)	Design effect (DE)	Total sample size (N_{tot})	Number of seminar groups (k)
0.01	20	1.19	353	18
0.01	15	1.14	338	24
0.01	10	1.09	323	34
0.02	*20*	*1.38*	*409*	*22*
0.02	15	1.28	379	26
0.02	10	1.18	350	36
0.03	20	1.57	465	24
0.03	15	1.42	421	30
0.03	10	1.27	376	38
0.05	20	1.95	578	30
0.05	15	1.70	504	34
0.05	10	1.45	430	44

proach results in a higher actual total sample size than the total sample size given in the column N_{tot}, which means that the power reaches values higher than 80%. Because of the given framework conditions (maximum of 320 students, seminar group size m=20), it would not be possible to complete the study in one semester. However, it would not be advisable to perform the study over several semesters or as a multicenter study because of the considerable differences between different academic years (students and lecturers, other framework conditions) and universities. Consequently, a confirmatory study is not realistic in this setting. In such a situation, the exploratory approach therefore appears to be advisable, i.e. calculation of power or minimum effect for a fixed given sample size.

A modification of the design would be the use of a stratified cluster randomization by weekday (Tuesday, Thursday, Friday).

Statistical analysis: Because the study is rather small and has a small number of clusters and because the cluster size is almost constant, cluster-level analysis is recommended (see Section 4.5). One option to perform such an analysis would be to calculate cluster mean values from the study results and use the two-sample t-test. Further examples with practical examples of analyses of cluster-randomized studies are given in [30], [32], and [33].

6. Discussion and recommendations

In addition to studies with other designs (e.g. observational studies), prospective two-arm (intervention) studies are frequently used in educational research to compare different teaching methods. These studies should adhere to recognized standards and methods of scientific research, in particular the presence of a control arm and the achievement of statistical equivalence (i.e. structural equivalence – achieved by randomization and possibly stratification – and equivalence of treatment conditions and observations). Unless there is a legitimate reason not to do so, comparative scientific studies should no longer be performed without a control arm. Quasi-experimental studies with a control arm but without randomization should also be avoided. A major criticism of the results of such studies is the lack of structural equivalence combined with the risk of confounded effects. The best approach to avoid such problems is to assign the study participants to the study arms randomly, by either individual or cluster randomization. Because of the advantages of randomization, if at all possible the extra effort should be made, especially because the effort required to perform randomization is small compared with the effort associated with the entire study: the implementation of a study usually requires many resources, whereas randomization requires comparatively few. However, randomization results in large gains in the interpretability and validity of the study findings.

Compared with studies from other fields, however, studies in educational research have some special conditions and requirements that affect their planning, implementation, and analysis. Because of the existence of natural clusters, cluster randomization usually is the only option to perform randomization if there is a limited number of students and a given approximately constant cluster size. Furthermore, the time- and location-related availability of different resources, such as lecturers, seminar rooms, laboratories, lecture halls, and computer pools, must be considered. When calculating sample size, it is necessary to take the cluster structure into account because the outcome is more similar among students within a cluster (e.g. within seminar groups) than among students from different clusters (e.g. different seminar groups). Depending on the strength of this similarity (measured by the ICC), in cluster-randomized studies the sample size required to achieve a certain power can be significantly higher than the sample size of a corresponding study with individual randomization. For this reason, i.e. because of their limited maximum sample size, many studies in educational research have only exploratory character (for reasons of feasibility). Structural equivalence is particularly important so that differences found in the study can be attributed to the method being studied. When performing the statistical analysis of cluster-randomized studies, one should also ensure that an adequate statistical methodology is being used that gives appropriate consideration to the dependencies resulting from the cluster structure. Because of the complex statistical methods required in all phases of a cluster-randomized study, support from a competent expert with appropriate specialist knowledge is recommended during the practical implementation of such studies. Such experts may be scientific staff at biometric institutions, for example, which are part of most universities with a medical faculty.

In addition to the disadvantages mentioned in Table 1, compared with studies with individual randomization studies with cluster randomization have a higher risk of not achieving structural equivalence at the individual level, which may jeopardize the internal validity. Another reason to critically scrutinize internal validity is that cluster-randomized studies usually are not blinded [12]. Consequently, an adjustment for the unequally distributed characteristics must be made during the statistical analysis, e.g. by a suitable regression method [12], [13]. As with all clinical studies, even if cluster-randomized studies have internal validity external validity can only be established heuristically. This is probably more difficult in educational research studies than in clinical studies because the conditions at the various teaching institutions are so different. Because of the larger sample sizes and the more complex methodology, one should consider at the planning stage of educational research studies whether a cluster randomization is justified and necessary [34].

Finally, the following recommendations summarize important measures that take clusters into account and ensure the quality of prospective, two-arm studies in educational research:

1. Teaching is usually performed in groups of students, so that a natural cluster structure is given and cluster randomization is the most appropriate approach.
2. Attention must be paid to cluster randomization during study design, sample size calculation, analysis, and reporting.
3. A cluster-randomized study should not include too few clusters, i.e. no fewer than 8-10 [32].
4. In the case of very few or very different clusters, it may make sense to match clusters.
5. Blinding usually is not possible. The use of outcome variables that are as objective as possible and a blind assessment, for example the evaluation of PBL outcomes by independent people who are not participating in the study, is therefore recommended and serves to improve the internal validity.

6. As far as possible, structural equivalence should be maintained by creating the same conditions for the groups, for example the same times and seminar rooms for the study arms being compared.

On the basis of our experience and the arguments presented here, we recommend the use of control arms and suitable randomization in prospective, two-arm comparative educational research studies in order to achieve good and convincing results also in studies in this field. Cluster randomization can be a crucial building block in this context, and therefore it should be increasingly used in educational research studies.

Acknowledgements

The authors thank Jacquie Klesing, Board-certified Editor in the Life Sciences (ELS), for editing assistance with the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Buss B, Wagner R, Bauder M, Fenik Y, Riessen R, Lammerding-Köppel M, Gawaz M, Fateh-Moghadam S, Weyrich P, Celebi N. Student tutors for hands-on training in focused emergency echocardiography – a randomized controlled trial. BMC Med Educ. 2012;12:101. DOI: 10.1186/1472-6920-12-101
2. Herter DA, Wagner R, Holderried F, Fenik Y, Riessen R, Weyrich P, Celebi N. Effect of supervised students' involvement on diagnostic accuracy in hospitalized medical patients--a prospective controlled study. PLoS One. 2012;7(9):e44866. DOI: 10.1371/journal.pone.0044866
3. Werner A, Holderried F, Schäffeler N, Weyrich P, Riessen R, Zipfel S, Celebi N. Communication training for advanced medical students improves information recall of medical laypersons in simulated informed consent talks - a randomized controlled trial. BMC Med Educ. 2013;1:13-15. DOI: 10.1186/1472-6920-13-15
4. Herrmann-Werner A, Nikendei C, Keifenheim K, Bosse HM, Lund F, Wagner R, Celebi N, Zipfel S, Weyrich P. Best practice" skills lab training vs. a "see one, do one" approach in undergraduate medical education: an RCT on students' long-term ability to perform procedural clinical skills. PLoS One. 2013;8(9):e76354. DOI: 10.1371/journal.pone.0076354
5. Ackel-Eisnach K, Raes P, Hönikl L, Bauer D, Wagener S, Möltner A, Jünger J, Fischer MR. Is German Medical Education Research on the rise? An analysis of publications from the years 2004 to 2013. GMS Z Med Ausbild. 2015;32(3):Doc30. DOI: 10.3205/zma000972
6. Schumacher M, Schulgen G. Methodik Klinischer Studien, Methodische Grundlagen der Planung, Durchführung und Auswertung. 3. Auflage. Heidelberg: Springer Verlag; 2008.
7. Armitage P. The role of randomization in clinical trials. Stat Med. 1982;1:345-352. DOI: 10.1002/sim.4780010412
8. Boet S, Sharma S, Goldman J, Reeves S. Review article: medical education research: an overview of methods. Can J Anaesth. 2012;59(2):159-170. DOI: 10.1007/s12630-011-9635-y
9. Fisher LD. Ethics of Randomized Trials. In: Armitage P, Colton T (Hrsg). Encyclopedia of Biostatistics. Chichester: Wiley & Sons Ltd; 1998. P.1394-1398.
10. Gaus, W, Muche, R. Medizinische Statistik. Stuttgart: Schattauer Verlag; 2013.
11. Korzilius H. EU-Verordnung über klinische Prüfungen: Kompromiss verabschiedet. Dtsch Ärztebl. 2014;5.
12. Eldridge SM, Kerry S. A Practical Guide to Cluster Randomised Trials in Health Services Research. Weinheim: Wiley; 2012. DOI: 10.1002/9781119966241
13. Hayes RJ, Moulton LH. Cluster Randomised Trials. Oxford: Oxford University Press; 2009. DOI: 10.1201/9781584888178
14. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. Int Stat Rev. 2009;77:378-394. DOI: 10.1111/j.1751-5823.2009.00092.x
15. Chenot JF. Cluster-randomisierte Studien: eine wichtige Methode in der allgemeinmedizinischen Forschung. Z Evid Fortbild Qual Gesundheitswes. 2009;103(7):475-480. DOI: 10.1016/j.zefq.2009.07.004
16. Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. BMJ. 1998;316(7142):1455. DOI: 10.1136/bmj.316.7142.1455
17. Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. Stat Med. 2002;21:3757-3774. DOI: 10.1002/sim.1330
18. Donner A. A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model. Int Stat Rev. 1986;54(1):67-82. DOI: 10.2307/1403259
19. Ridout MS, Demétrio CG, Firth D. Estimating intraclass correlation for binary data. Biometrics. 1999;55(1):137-148. DOI: 10.1111/j.0006-341X.1999.00137.x
20. Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp Clin Trials. 2012;33(5):869-880. DOI: 10.1016/j.cct.2012.05.004
21. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. J Clin Epidemiol. 2004;57(8):785-794. DOI: 10.1016/j.jclinepi.2003.12.013
22. Hemming K, Girling AJ, Sitch AJ, Marsh J, Lilford RJ. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. BMC Med Res Methodol. 2011;11:102. DOI: 10.1186/1471-2288-11-102
23. Dang Q, Mazumdar S, Houck PR. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. Comput Methods Programs Biomed. 200;91(2):122-127.
24. Dreyhaupt J. Instrumente für Power- und Fallzahlberechnungen bei komplexen hierarchischen Studiendesigns in der Versorgungsforschung. Monit Versorgungsforsch. 2015;6:49-54.
25. Dreyhaupt J. Generelle Fallzahl- und Powerabschätzung über Simulation bei Studien mit komplexen hierarchischen Daten als Unterstützung der Studienplanung in der Versorgungsforschung. Ulm: Universität Ulm; 2015. Zugänglich unter/available from: URL: http://vts.uni-ulm.de/query/longview.meta.asp?document_id=9509

26. Landau S, Stahl D. Sample size and power calculations for medical studies by simulation when closed form expressions are not available. Stat Methods Med Res. 2013;22(3):324-345. DOI: 10.1177/0962280212439578

27. Bland JM, Kerry SM. Trials randomised in clusters. BMJ. 1997;315(7108):600. DOI: 10.1136/bmj.315.7108.600

28. Donner A, Klar N. Design and Analysis of Cluster Randomization trials in Health Research. Weinheim: John Wiley & Sons, Ltd; 2010.

29. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. Fam Pract. 2000;17(2):192-196. DOI: 10.1093/fampra/17.2.192

30. Campbell MK, Piaggio G, Elbourne DR, Altman DG; CONSORT Group (2012). Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012. DOI: 10.1136/bmj.e5661

31. Mayer B, Danner B. Von Naschkatzen und Nagetieren – Eine interaktive Einführung in die Medizinische Biometrie mit der NANA-Studie. In: Rauch G, Muche R, Vonthein R (Hrsg). Zeig mir Biostatistik! Ideen und Material für einen guten Biometrie-Unterricht. Heidelberg: Springer Verlag; 2014. S.3-14. DOI: 10.1007/978-3-642-54336-4_1

32. Eldridge SM, Costeloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? Stat. Methods Med Res. 2016:1039-1056. DOI: 10.1177/0962280215588242

33. Campbell MK Analysis of cluster randomized trials in primary care: a practical approach. BMJ. 1998;316:1455.

34. Kuß O, Jahn P, Renz P, Landenberger M. Cluster-randomisierte Studien in der Pflegewissenschaft. Halle Beitr Gesundheit Pflegewissenschaft. 2009;8(1):302-310.

Corresponding author:
Dr. Jens Dreyhaupt
Ulm University, Institute of Epidemiology and Medical Biometry, Schwabstr. 13, 89075 Ulm, Germany, Phone: +49(0)731/50-26895, Fax: +49(0)731/50-26902
jens.dreyhaupt@uni-ulm.de

Please cite as
Dreyhaupt J, Mayer B, Keis O, Öchsner W, Muche R. Cluster-randomized Studies in Educational Research: Principles and Methodological Aspects. GMS J Med Educ. 2017;34(2):Doc26. DOI: 10.3205/zma001103, URN: urn:nbn:de:0183-zma0011038

This article is freely available from
http://www.egms.de/en/journals/zma/2017-34/zma001103.shtml

Received: *2016-08-16*
Revised: *2016-11-17*
Accepted: *2016-12-29*
Published: *2017-05-15*

Cluster-randomisierte Studien in der Lehrforschung: Grundlagen und methodische Aspekte

Jens Dreyhaupt[1]
Benjamin Mayer[1]
Oliver Keis[2]
Wolfgang Öchsner[2,3]
Rainer Muche[1]

1 Universität Ulm, Institut für Epidemiologie und Medizinische Biometrie, Ulm, Deutschland

2 Universität Ulm, Studiendekanat Medizin, Ulm, Deutschland

3 Universitätsklinik Ulm, Abteilung Kardioanästhesie, Ulm, Deutschland

Zusammenfassung

In der Lehrforschung werden immer häufiger Studien zur Evaluation neuer Lehrmethoden und Ansätze durchgeführt, wobei diese Studien bei einer strengeren Anwendung und Einhaltung von anerkannten Standards und Methoden wissenschaftlicher Untersuchungen noch effizienter durchgeführt werden und überzeugendere Ergebnisse liefern könnten. Insbesondere bei prospektiven zweiarmigen (Interventions)-Studien, in denen zwei verschiedene Lehrmethoden verglichen werden sollen, könnte eine entsprechende Vorgehensweise zu einer substanziellen Qualitätssteigerung führen. Ein wesentlicher Standard ist dabei die Randomisierung, mit der systematische Verzerrungen der Studienergebnisse durch Strukturungleichheiten in den zu vergleichenden Studienarmen weitestgehend ausgeschlossen werden können. Dieser Standard sollte möglichst auch bei Studien in der Lehrforschung erreicht werden, wo er sich allerdings aktuell noch nicht allgemein durchgesetzt hat. Es gibt hierbei einige Schwierigkeiten und Vorbehalte, vor allem organisatorische und methodische Aspekte. Insbesondere muss beachtet werden, dass bei Studien in der Lehrforschung bedingt durch die Lehrsituation meist keine individuelle Randomisierung sondern eine Randomisierung ganzer Gruppen (sogenannte Cluster-Randomisierung) vorgenommen werden muss. Im Vergleich zu individuell randomisierten Studien sind bei cluster-randomisierten Studien meist (deutlich) höhere Fallzahlen sowie eine komplexere Methodik der Fallzahlplanung notwendig. Weiterhin erfordern cluster-randomisierte Studien umfassendere Methoden zur statistischen Auswertung. Dies hat zur Konsequenz, dass die praktische Anwendung cluster-randomisierter Studien in allen ihren Phasen der Unterstützung durch einen kompetenten Experten mit entsprechenden Spezialkenntnissen bedarf.
Eine verstärkte Anwendung der Randomisierung in Studien zur Beurteilung neuer Methoden in der Lehre ist notwendig, um wissenschaftlich überzeugende Ergebnisse zu erzielen. Um dazu beizutragen, werden in diesem Beitrag allgemeine Grundlagen der Cluster-Randomisierung beschrieben, deren Umsetzung und praktische Aspekte der Durchführung im Kontext von prospektiven zweiarmigen vergleichenden Studien in der Lehrforschung erläutert.

Schlüsselwörter: Cluster-Randomisierung, Strukturgleichheit, Lehrforschung, Studie, Fallzahlplanung, Auswertung

1. Einleitung

In der Lehrforschung, insbesondere im medizinischen Umfeld, werden immer häufiger Studien zur Evaluation neuer Lehrmethoden und Ansätze durchgeführt [1], [2], [3], [4], wobei gerade in Deutschland in den letzten Jahren ein Ansteigen beobachtet wurde [5]. Eine Einhaltung anerkannter Standards und Methoden wissenschaftlicher Untersuchungen sollte zweifelsohne auch bei diesen Studien erfolgen. In der Arzneimittelentwicklung sind die Methoden klinischer Studien weit entwickelt [6], [http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html]. Das Vorgehen ist weltweit standardisiert, um zu aussagekräftigen Studienergebnissen zu führen. Diese Standards sollten daher möglichst auch bei Studien in der Lehrforschung etabliert werden und zur Anwendung kommen. Neben Beobachtungsstudien betrifft dies vor allem prospektive zweiarmige (Interventions)-Studien, in denen zwei verschiedene Lehrmethoden verglichen werden sollen.
In der klinischen Forschung ist für den Wirksamkeitsnachweis oder den Überlegenheitsnachweis einer Therapie ein Kontrollarm notwendig, mit welchen der Testarm im statistischen Sinne vergleichbar ist (Strukturgleichheit).

Eine Randomisierung bewirkt, dass sich alle Störgrößen – bekannte und unbekannte – gleichmäßig auf Kontroll- und Testarm verteilen und sich bei Behandlungsbeginn höchstens zufällige Unterschiede ergeben [7]. Im Kontext klinischer Studien bedeutet Randomisierung nicht, dass der Patient ohne ersichtliches Kriterium einfach zugeteilt wird, sondern dass formal ein Zufallsexperiment durchgeführt wird, welches unabhängig vom beteiligten Kliniker ist. Bei Verwendung jedes anderen Zuteilungsverfahrens der Probanden oder Patienten auf Kontroll- und Testarm muss mit systematischen Verzerrungen der Ergebnisse gerechnet werden. Auch für Studien in der Lehrforschung ist die Randomisierung als Methode anerkannt [8].

Bei Studien in der Lehrforschung gibt es darüber hinaus einige Besonderheiten und spezielle Anforderungen beim Einsatz der Randomisierung. So ist es oft nicht sinnvoll oder möglich, einzelne Studierende individuell zu randomisieren, da beispielsweise durch individuelle Semesterpläne die Studierenden nicht zu jedem Termin verfügbar sind. Weiterhin ist hier meist die natürliche Studiensituation der Lehre in Gruppen vorgegeben, wie beispielsweise gemeinsame Vorlesungen oder Seminare, die von einem Dozierenden betreut werden. Aufgrund dieser Besonderheiten ist es sinnvoll und notwendig, bei Studien in der Lehrforschung eine Randomisierung ganzer Gruppen vorzunehmen (Cluster-Randomisierung) anstelle individueller Randomisierung einzelner Studierender.

Um wissenschaftlich überzeugende Ergebnisse im Sinne einer "evidence based didactics" zu erhalten, ist es wichtig, auch in der Lehrforschung verstärkt randomisierte Studien zur Evaluation neuer Lehrmethoden und Ansätze durchzuführen. Um zu einer stärkeren Nutzung der Randomisierung beizutragen, werden in dieser Arbeit Grundlagen der Cluster-Randomisierung beschrieben und praktische Aspekte erläutert, um ihren Einsatz in der Lehrforschung im Kontext von prospektiven zweiarmigen vergleichenden Studien zu erleichtern.

Im ersten Abschnitt wird die Rationale der Randomisierung dargestellt. Anschließend wird auf die besonderen Bedingungen bei wissenschaftlichen Studien in der Lehrforschung eingegangen. Im dritten Abschnitt wird das Prinzip der Cluster-Randomisierung dargestellt, danach wird auf Fallzahlplanung und Auswertungsansätze cluster-randomisierter Studien eingegangen, die sich von individuell randomisierten Studien unterscheiden. Zuletzt wird eine exemplarische Studienplanung mit Cluster-Randomisierung in der Lehrforschung als Anwendungsbeispiel vorgestellt.

2. Der Hintergrund der Randomisierung: Strukturgleichheit

Um aufzeigen zu können, warum eine Randomisierung notwendig ist, sollen zunächst einige Begrifflichkeiten im Kontext von Studien in der Lehrforschung erläutert werden.

Notwendigkeit eines Kontrollarms: Wird im Rahmen einer Studie in der Lehrforschung bei nur einem Studienarm eine neue Methode angewendet (Testarm), kann nicht ausgeschlossen werden, dass ein beobachteter Effekt auch ohne die neue Methode entstanden sein könnte. Es können also nicht alle beobachteten Erfolge im Testarm der neuen Methode zugerechnet werden. Die Wirksamkeit einer neuen Methode ist erst dann evident, wenn der spezifische Erfolg dieser Methode größer ist als ohne diese [9]. Deshalb ist für einen solchen Wirksamkeitsnachweis ein Kontrollarm unabdingbar, in welcher die bisherige Methode angewendet wird.

Statistische Vergleichbarkeit: Damit beobachtete Unterschiede hinsichtlich des Erfolgs im Vergleich zwischen Kontroll- und Testarm tatsächlich nur auf die neue Methode zurückgeführt werden können, muss statistische Vergleichbarkeit vorliegen:

1. hinsichtlich der Struktur von Kontroll- und Testarm (Strukturgleichheit),
2. im Umgang mit den Studierenden mit Ausnahme der spezifischen zu evaluierenden neuen Methode (Behandlungsgleichheit) und
3. auch hinsichtlich der Beobachtung (Beobachtungsgleichheit).

Dabei bedeutet statistisch gleich, dass sich Kontroll- und Testarm möglichst wenig, jedoch höchstens zufällig unterscheiden.

Strukturgleichheit: Strukturgleichheit ist gegeben, wenn Test- und Kontrollarm in ihrer Zusammensetzung hinsichtlich möglicher sogenannter "Störgrößen" statistisch gleich sind [10]. Beispiele für solche Störgrößen sind in der Lehrforschung Alter und Geschlecht, die oft einen Einfluss auf das Ergebnis haben. In der Lehrforschung sind darüber hinaus Persönlichkeitsfaktoren zu nennen, wie beispielsweise Vorbildung und spezielle Fähigkeiten sowie Interessen und Aktivitäten. Ein Instrument zum Erreichen der Strukturgleichheit stellt die Randomisierung dar. Sie bewirkt, dass man durch die Zufallszuteilung der Studierenden zu Kontroll- und Testarm eine ähnliche Verteilung bekannter und unbekannter Störgrößen auf beide Arme erreichen bzw. annehmen kann. Darüber hinaus kann ein Gleichgewicht zwischen Kontroll- und Testarm hinsichtlich (weniger) wesentlicher bekannter Störgrößen durch Schichtung erzeugt werden (siehe Abschnitt 4.2), was allerdings im Rahmen der Lehrforschung nur sehr eingeschränkt praktikabel ist.

Behandlungs- und Beobachtungsgleichheit: Behandlungsgleichheit liegt vor, wenn alle Studierenden aller Gruppen bis auf die zu beurteilende neue Methode die gleiche Behandlung erhalten und erfahren. Diese Behandlungen (z. B. gleiche Uhrzeiten für Seminare, gleiche Bedingungen für das Schreiben von Klausuren) sollten – auch bei Studien in der Lehrforschung – in einem Studienprotokoll möglichst detailliert festgelegt werden, damit bei der Interpretation der Ergebnisse klar ist, "was" miteinander verglichen wird und unter welchen Bedingungen der beobachtete Effekt entstanden ist. Kontroll- und Testarm gelten als "beobachtungsgleich", wenn gleiche Sachverhalte stets nach gleichen Regeln beobachtet und beurteilt werden (standardisierte Bedingungen), z. B. gleiche Be-

werter für Klausuren. Ein Instrument zum Erreichen von Behandlungs- und Beobachtungsgleichheit stellt in der klinischen Forschung die Maskierung (Verblindung) von Therapien dar, was allerdings bei Studien in der Lehrforschung meist nicht möglich ist. Denkbar wäre hier lediglich eine verblindete Bewertung, z. B. die Bewertung von Ergebnissen im Problem-basierten Lernen (PBL) durch unabhängige und nicht an der Studie beteiligte Personen. Hat eine Studie in der Lehrforschung die oben genannten Eigenschaften und ist der Erfolg im Testarm um so viel größer als im Kontrollarm, dass dieser Unterschied nicht mehr mit dem Zufall vereinbar ist, wurde gezeigt, dass nur die neue Methode den Erfolg bewirkt haben kann. Unterscheiden sich Kontroll- und Testarm jedoch in weiterer Hinsicht (ist also Struktur-, Behandlungs- oder Beobachtungsgleichheit nicht erfüllt), treten Interpretationsprobleme auf, da ein beobachteter Effekt nicht mehr ausschließlich der neuen Methode zugeschrieben werden kann (vermengte Effekte). Der spezifische Effekt der neuen Lehrmethode kann in so einer Situation nicht berechnet werden, und in der Regel können die Studienergebnisse auch nicht entsprechend der Fragestellung interpretiert werden. Die Randomisierung ist daher als Verfahren zur Erreichung der Strukturgleichheit ein wichtiges Instrument bei der Wirksamkeitsuntersuchung neuer Methoden in der Lehrforschung.

3. Bedingungen und Anforderungen an Studien in der Lehrforschung

Studien in der Lehrforschung finden in einem besonderen Kontext statt. Es gibt meist eine natürliche Clusterstruktur, da die Lehre in der Regel in Gruppen von Studierenden durchgeführt wird. Beispiele sind Seminargruppen, PBL-Gruppen oder Gruppen von Studierenden, die eine gemeinsame Vorlesung besuchen. Die Gruppengröße ist sehr unterschiedlich und reicht von 5-8 (PBL-Gruppen) bis über 100 Studierende in einer gemeinsamen Vorlesung. Die Gruppengröße in Seminaren ist unterschiedlich, wobei in der Medizin von bis zu 20 Studierenden pro Seminargruppe ausgegangen wird [http://www.gesetze-im-internet.de/_appro_2002/index.html]. Alle Studierenden innerhalb einer Gruppe sind denselben Bedingungen ausgesetzt, wie beispielsweise derselben Lehrperson oder denselben Räumlichkeiten und Zeiten. Dies hat zur Konsequenz, dass bei Studien die Intervention (z. B. Anwendung einer neuen Lehrmethode durch den Dozierenden) auf Ebene der gesamten Gruppe erfolgt. Somit ist das Ergebnis von Studierenden einer Gruppe in der Regel ähnlicher als das Ergebnis von Studierenden verschiedener Gruppen. Darüber hinaus ist bei Studien in der Lehrforschung neben den individuellen Semesterplänen der Studierenden die zeitliche Verfügbarkeit geeigneter Ressourcen (wie Dozierende, Seminarräume, Labore, Hörsäle, Computerpools) als weitere Rahmenbedingung zu beachten. Durch eine limitierte Anzahl der Studierenden ergibt sich -bei vorgegebener Gruppengröße- auch eine limitierte Anzahl möglicher Gruppen für Studien in der Lehrforschung. Die genannten Aspekte führen dazu, dass bei Studien in der Lehrforschung kaum individuelle Randomisierung erfolgen kann, sondern eine Cluster-Randomisierung sinnvoll und möglich ist. Dies bedeutet, dass Gruppen von Studierenden bzw. dass Dozierende (welche Gruppen von Studierenden betreuen) in Test- und Kontrollarm randomisiert werden, wobei die hier genannten Besonderheiten Berücksichtigung finden müssen.

In Studien der Lehrforschung sind verschiedene Zielgrößen denkbar:

- Evaluationsergebnisse (z. B. Zufriedenheit der Studierenden, Selbsteinschätzung der erreichten Kompetenz)
- Ergebnisse von Leistungsnachweisen/Prüfungsergebnissen (erreichter Kompetenzgrad)
- Überprüfung des erforderlichen Lernaufwands (z. B. Tracking von Lernzeiten)
- Begleiteffekte der Beschäftigung mit dem Lernstoff (z. B. Motivationskurven, Begeisterung für das Fach, Interessensweckung, Karriereplanung)

Im Folgenden werden metrische Zielgrößen betrachtet (z. B. Punktezahlen in Klausuren), die als annähernd normalverteilt angenommen werden. Binäre Zielgrößen (z. B. bestanden (ja/nein)) werden in diesem Artikel nicht behandelt, da sie in der Regel eine deutlich höhere Fallzahl benötigen und damit im Rahmen von Studien in der Lehrforschung kaum anwendbar sind.

4. Cluster-Randomisierung und ihre Anwendung in der Lehrforschung

In den folgenden Abschnitten wird das Prinzip der Cluster-Randomisierung erläutert und auf den Bereich der Lehrforschung angewendet. Weiterhin werden Informationen zu Studiendurchführung, Fallzahlplanung und Auswertung gegeben.

4.1. Definition und Motivation

In einer cluster-randomisierten Studie (engl. cluster randomised trial, Synonyme: „group randomised trial“, „community randomisation trial“, „community intervention trial“) werden komplette soziale Gruppen oder Cluster von Individuen, anstelle von einzelnen Individuen, randomisiert, siehe Abbildung 1.

Studien dieses Typs findet man bisher insbesondere bei der Untersuchung nicht-therapeutischer Interventionen, wie z. B. der Bewertung von Schulungsprogrammen, Präventionsprogrammen oder Maßnahmen zur Gesundheitsförderung. Aus den im letzten Abschnitt genannten Gründen erscheint die Verwendung cluster-randomisierter Studien auch im Bereich der Lehrforschung sinnvoll. Unter publizierten cluster-randomisierten Studien finden sich sowohl Studien mit kleinen als auch Studien mit großen Clustern: Häufig werden Cluster über Haushalte, Familien, Nachbarschaften, Gemeinden, Schulklassen, Arbeitgeber, Krankenhäuser oder Praxen niedergelassener Ärzte defi-

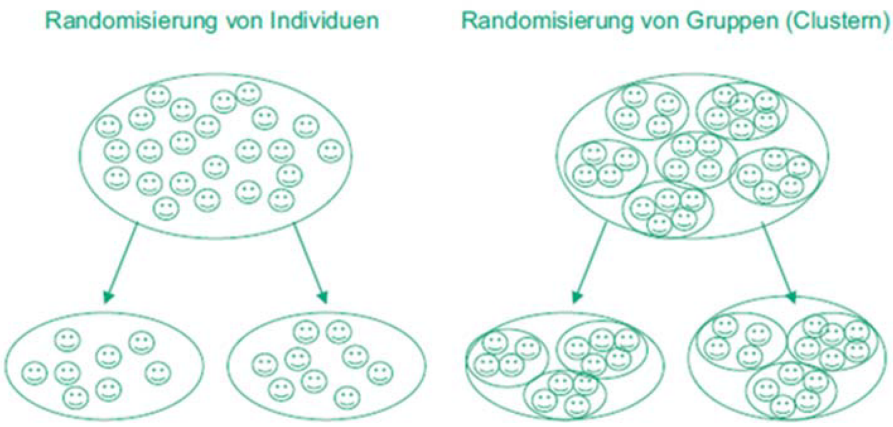

Abbildung 1: Randomisierung von Individuen vs. Randomisierung von Clustern aus [15]

niert. Somit kann die Anzahl von Individuen pro Cluster zwischen 2 und mehreren Tausend liegen. Auch in der Lehrforschung sind verschiedene Clustergrößen denkbar (siehe Abschnitt 3).

Hauptmotivation für die Durchführung einer cluster-randomisierten Studie ist das Bestreben, einen Kontaminationsbias zu vermeiden oder zu verringern. Ein solcher Bias (Verzerrung oder systematischer Fehler) könnte bei Randomisierung von Individuen durch Interaktion zwischen den Individuen aus verschiedenen Studienarmen hervorgerufen werden. Beispielsweise könnten bei individueller Randomisierung Studierende des Kontrollarms leicht durch Studierende des Testarms animiert werden, die speziellen Methoden des Testarms durchzuführen, deren Wirksamkeit geprüft werden soll. Allerdings ist auch bei einer Cluster-Randomisierung diese Möglichkeit nicht ausgeschlossen, sondern nur vermindert. Bei Studien in der Lehrforschung spielt im Zusammenhang mit Kontaminationsbias insbesondere die stark verbreitete Verwendung sozialer Medien wie Facebook eine Rolle. Ein weiterer wesentlicher Grund für die Anwendung der Cluster-Randomisierung in der Lehrforschung ist das Vorliegen natürlicher Cluster, da das Lernen in der Regel in Gruppen erfolgt (siehe Abschnitt 3). Tabelle 1 zeigt wichtige Vor- und Nachteile der Cluster-Randomisierung im Kontext der Lehrforschung.

4.2. Designs

Bei cluster-randomisierten Studien kann zwischen vollständig randomisiertem Design, stratifiziert randomisiertem Design und gematchtem Design unterschieden werden. Bei vollständig randomisiertem Design werden die Cluster zufällig den Gruppen zugeteilt und es wird dabei weder stratifiziert noch gematcht. Ein Beispiel ist eine Studie in der Lehrforschung, bei welcher Seminargruppen entweder in den Testarm oder den Kontrollarm randomisiert werden. Beim stratifiziert randomisierten Design wird geschichtet nach (wenigen) wichtigen bekannten Störgrößen randomisiert, so dass die Verteilung der Störgrößen in Test- und Kontrollarm ähnlich ist. Stratifiziert wird nach Faktoren, die stark mit der Zielgröße assoziiert sind, wie beispielsweise Clustergröße, Geschlecht oder Tag. Ein Beispiel ist eine Studie in der Lehrforschung, bei welcher angenommen wird, dass der Tag, an dem ein Seminar stattfindet, Einfluss auf die Zielgröße hat (d. h. eine Störgröße ist). Hier kann zunächst nach Tag stratifiziert werden (z. B. Montag/Mittwoch/Freitag) und innerhalb jedes Tages werden dann Seminargruppen entweder in den Testarm oder den Kontrollarm randomisiert. Auf diese Art und Weise wird erreicht, dass die Störgröße Tag annähernd gleichmäßig auf beide Arme verteilt wird. Im gematchten Design werden Paare von Clustern gebildet, die so ähnlich wie möglich sind, in Bezug auf wichtige Faktoren, die die Zielgröße beeinflussen. Ein Cluster des Paares wird jeweils in den Testarm und das andere Cluster in den Kontrollarm randomisiert. Dadurch ist eine gute Möglichkeit gegeben, Störgrößen (z. B. Charakteristika aus der Baselineerhebung wie Geschlecht, Fachsemester, Vornote) zwischen beiden Armen zu balancieren, so dass eine Vergleichbarkeit der Arme erreicht wird. Für das Matching sollten nicht zu viele Kriterien herangezogen werden, da dann evtl. kein Cluster mehr gefunden werden kann, welches mit einem anderen ein Paar bilden kann. Unter den in Abschnitt 3 genannten Bedingungen und Anforderungen (Vorgaben für Clustergröße, limitierte Anzahl Studierender und damit limitierte Clusteranzahl, Verfügbarkeit von Ressourcen, individuelle Semesterpläne) ist bei Studien in der Lehrforschung oft von einer vergleichsweise kleinen Clusteranzahl mit einer mehr oder weniger fest vorgegebenen Clustergröße auszugehen. Stratifizierte und gematchte Designs dürften in einem solchen Kontext nur unter besonderen Bedingungen realisierbar sein. Ein Beispiel ist eine multizentrische Studie, die an verschiedenen Einrichtungen durchgeführt wird. Aus diesem Grund wird das vollständig randomisierte Design in der Lehrforschung überwiegen.

4.3. Praktische Durchführung

Studien in der Lehrforschung sind in der Regel bei der Ethikkommission anzuzeigen. Es ist oft allerdings kein schriftliches Einverständnis der beteiligten Studierenden erforderlich, lediglich eine Aufklärung [11].

Ein- und Ausschlusskriterien müssen sowohl auf Individualebene (Studierende) als auch auf Cluster-Ebene (Lehrende) definiert werden. Problematisch ist, dass bei Studien in der Lehrforschung meist keine Verblindung

Tabelle 1: Vor- und Nachteile der Cluster-Randomisierung bei Studien in der Lehrforschung (nach [15])

Vorteile	Nachteile
- Berücksichtigung der natürlichen Lehrsituation (Lehre in Gruppen) - Beachtung der Bedingungen und Anforderungen an wissenschaftliche Untersuchungen in der Lehrforschung - Verminderung des Kontaminationsbias	- Im Vergleich zu individuell randomisierten Studien meist (deutlich) höhere Fallzahlen notwendig - Statistische Abhängigkeit der Studierenden innerhalb von Gruppen: Komplexere statistische Methoden für Fallzahlplanung und Auswertung notwendig

möglich sein wird. Damit besteht die Gefahr eines Bias in der Zielgröße. Dieser Gefahr sollte durch Maßnahmen zur Erreichung der Behandlungs- und Beobachtungsgleichheit entgegengewirkt werden. Beispiele sind eine starke Standardisierung des generellen Vorgehens, ggf. eine verblindete Beurteilung des Erfolgs, z. B. durch einen dritten, nicht in die Studie involvierten Bewerter, der keine Kenntnis über die Zugehörigkeit des jeweiligen Studierenden zu Test- und Kontrollarm hat.

4.4. Fallzahlplanung

4.4.1. Warum eine eigene Fallzahlplanung?

Durch die Cluster-Randomisierung wird eine spezielle Datenstruktur erzeugt, wobei Beobachtungen innerhalb der Cluster meist ähnlicher sind als Beobachtungen aus verschiedenen Clustern (Vorliegen von statistischer Abhängigkeit). Dies bedeutet im Kontext von Studien in der Lehrforschung, dass die Ergebnisse (z. B. Leistungen in der Klausur) von Studierenden innerhalb der gleichen Seminargruppe ähnlicher sind als Ergebnisse von Studierenden verschiedener Seminargruppen. Dadurch kommt es zu einem Effizienz- und Powerverlust, was sich auf die Fallzahlplanung auswirkt: Die effektive Fallzahl einer cluster-randomisierten Studie (d. h. die Anzahl der wirklich statistisch unabhängigen individuellen Beobachtungen) ist niedriger als die tatsächliche Fallzahl (d. h. die Anzahl rekrutierter Studierender). Daher sind Standardverfahren, die von der statistischen Unabhängigkeit aller Beobachtungen ausgehen, für Fallzahlplanung und Auswertung solcher Daten ungeeignet. Die Anwendung von Standardverfahren für die Fallzahlplanung würde zu Studien mit zu geringer Power führen, in denen die Chance einen tatsächlich vorhandenen Unterschied zwischen den Studienarmen nachzuweisen, (deutlich) geringer ist, als in der Planung angenommen. In der Lehrforschung kann dies beispielsweise dazu führen, dass eine neue Lehrmethode, die in der Wirklichkeit besser ist, mit der Studie nicht erkannt wird.

4.4.2. Ähnlichkeitsbestimmung – der Intracluster-Korrelationskoeffizient ICC

Um die Ähnlichkeit der Beobachtungen innerhalb der Cluster im Vergleich zu Beobachtungen aus verschiedenen Clustern zu quantifizieren wird als Maßzahl der Intracluster-Korrelationskoeffizient (synonym: Intraclass-Korrelationskoeffizient; abgekürzt ICC, ρ), verwendet. Der ICC kann auf verschiedene Art definiert werden [12]. Für metrische Zielgrößen wird der ICC oft als Quotient von Varianzen definiert [13], [14]:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} = \frac{\sigma_b^2}{\sigma^2}$$

wobei σ_b^2 die Varianz zwischen den Clustern, σ_w^2 die Varianz innerhalb desselben Clusters und $\sigma^2 = \sigma_b^2 + \sigma_w^2$ die Gesamtvarianz bezeichnet. Mit dieser Definition kann der ICC als Anteil der Varianz zwischen den Clustern an der Gesamtvarianz interpretiert werden, wobei davon ausgegangen wird, dass die Varianz σ_w^2 in jedem Cluster konstant ist. Der ICC kann mit dieser Definition Werte zwischen 0 und 1 annehmen. Seine Größenordnung ist ein Maß für die Stärke der Ähnlichkeit der Beobachtungen innerhalb der Cluster im Vergleich zur Ähnlichkeit der Beobachtungen zwischen den Clustern. Hat der ICC den Wert 1 sind die Beobachtungen innerhalb jedes Clusters gleich. Im Kontext von Studien in der Lehrforschung würde dies beispielsweise bedeuten, dass in jeder Seminargruppe alle Studierenden dieselbe Klausurnote haben (aber nicht notwendigerweise, dass alle Seminargruppen in der Studie dieselbe Note haben). Der ICC hat den Wert 0, wenn alle Beobachtungen statistisch unabhängig sind. Bei Studien in der Lehrforschung würde dies beispielsweise bedeuten, dass die Klausurnoten von Studierenden innerhalb derselben Seminargruppe nicht abhängig sind, d. h. dass kein Einfluss der Seminargruppe auf die Klausurnoten besteht.

Eine Herausforderung ist oft, eine apriori Schätzung des ICC zu erhalten. Der ICC kann z. B. aus Daten einer Pilotstudie berechnet oder der Literatur entnommen werden. Cluster-randomisierte Studien sollten daher ihre post-hoc ermittelten Intracluster-Korrelationskoeffizienten publizieren, damit diese für ähnliche Studien zur Verfügung stehen [15], [16]. Weiterhin ist der ICC nur eine Schätzung aus einer Stichprobe und somit mit Unsicherheit behaftet (Konfidenzintervall [17]). Dies hat besonders für Studien in der Lehrforschung Bedeutung, da hier oft nur kleine Studien mit wenigen Clustern durchgeführt werden können, bei denen der ICC nicht zuverlässig geschätzt werden kann.

Darüber hinaus können unterschiedliche Berechnungsverfahren Einfluss auf den Wert des ICC haben. Eine Übersicht über für metrische Zielgrößen geeignete Berech-

nungsverfahren des ICC gibt [18]. Für binäre Zielgrößen sind entsprechende Methoden in [19] und [20] verfügbar.

4.4.3. Der Designeffekt (DE)

Um in einer cluster-randomisierten Studie die gleiche Power wie in einer individuell randomisierten Studie zu erreichen, müssen in der cluster-randomisierten Studie in der Regel mehr Individuen rekrutiert werden. Die für eine cluster-randomisierte Studie notwendige Fallzahl ergibt sich aus der Fallzahl für die individuell randomisierte Studie durch Multiplikation mit dem Designeffekt (DE), der aus dem ICC ρ und der festen Clustergröße m berechnet wird:

DE = 1+ρ·(m-1)

Das Ergebnis ist eine Gesamtfallzahl und eine sich daraus ergebende Anzahl von Clustern (mit fester Clustergröße) für eine vorgegebene Power. Für Studien in der Lehrforschung bedeutet dies, dass zunächst eine Gesamtanzahl Studierender berechnet wird und daraus anschließend eine Anzahl an Seminargruppen (mit fester Gruppengröße m).

Bei ungleicher Clustergröße kann m ersetzt werden durch das arithmetische Mittel oder durch die maximale Clustergröße. Die Verwendung des arithmetischen Mittels der Clustergröße ist sinnvoll, wenn nur wenig Variabilität in der Clustergröße besteht [12], die Verwendung der maximalen Clustergröße ein konservativer Ansatz. Bei einem Intracluster-Korrelationskoeffizienten von ρ=0 (statistische Unabhängigkeit aller Beobachtungen, siehe oben) ist der Designeffekt DE=1, was bedeutet, dass die cluster-randomisierte Studie dieselbe Fallzahl wie die entsprechende individuell randomisierte Studie hat. Die Bildung von Clustern hat in dem Fall keinen Einfluss auf die Fallzahl. In der Praxis liegt die Größenordnung der meisten ICC zwischen 0.00 und 0.20, wobei eine sehr große Spannweite besteht [21].

4.4.4. Vorgehensweisen bei der Fallzahlplanung

Allgemein können bei der Studienplanung zwei Herangehensweisen betrachtet werden. Zum einen kann im Rahmen eines explorativen Ansatzes für die gegebene maximale Fallzahl bei gegebener Power und Clustergröße ein Mindesteffekt oder bei gegebenem Mindesteffekt und Clustergröße eine Power berechnet werden [22]. Dies ist insbesondere dann sinnvoll, wenn nur eine stark limitierte Anzahl von Beobachtungen zur Verfügung steht. Abbildung 2 zeigt das Schema der Berechnung von Power bzw. Mindesteffekt in Studien in der Lehrforschung bei gegebener Fallzahl.

Zum anderen kann ein konfirmatorischer Ansatz gewählt werden: Für eine vorgegebene Power und einen vorgegebenen Mindesteffekt wird eine Fallzahl (d. h. Anzahl Studierender und eine sich daraus ergebende Clusteranzahl) berechnet. Abbildung 3 zeigt das Schema der Fallzahlberechnung in Studien in der Lehrforschung bei vorgegebener Power und Mindesteffekt.

Wegen der speziellen Bedingungen in der Lehrforschung (limitierte Anzahl Studierender und damit Cluster sowie eine vorgegebene Clustergröße, vergleiche oben und Abschnitt 3) ist die Durchführung konfirmatorischer Studien allerdings limitiert.

Sollen in die Planung einer cluster-randomisierten Studie noch zusätzlich Kovariaten einbezogen werden, ist eine Erweiterung der Definition des ICC nach [14] möglich. Eine weitere Möglichkeit, insbesondere für komplexe Studiendesigns (beispielsweise Berücksichtigung mehrerer Kovariaten im longitudinalen Design, weitere Hierarchieebenen), bietet auch die Simulation (z. B. [23], [24], [25], [26]).

4.5. Auswertung

Um die statistischen Abhängigkeiten innerhalb der Cluster zu berücksichtigen (bei Studien in der Lehrforschung: z. B. Abhängigkeit von Klausurergebnissen von Studierenden in derselben Seminargruppe), muss bei der Auswertung eine Cluster-Adjustierung durchgeführt werden [12]. Eine sogenannte „naive Analyse“ (Cluster-Adjustierung bleibt unberücksichtigt; Anwendung von Standardverfahren wie beispielsweise Zweistichproben-t-Test) kann zur Schätzung von zu kleinen Konfidenzintervallen und p-Werten führen [27], [28]. Für Studien in der Lehrforschung hätte dies zur Konsequenz, dass falsch signifikante Studien berichtet und damit neue Lehrmethoden als vermeintlich besser dargestellt würden.

Die im Rahmen der Studienplanung angewendeten Methoden sollten auch zur Auswertung genutzt werden [23], [28], wobei die Methoden vom Studiendesign (siehe oben) abhängen. Bei der statistischen Analyse kann zwischen der Analyse auf Cluster-Ebene oder auf Individuen-Ebene unterschieden werden [28], [13]. Wegen der sehr komplexen statistischen Methoden ist insbesondere für die Auswertung die Unterstützung durch einen kompetenten Experten (z. B. Statistiker mit entsprechenden Spezialkenntnissen) empfehlenswert. In Bezug auf die human-/zahnmedizinische Lehre sind an fast allen Medizinischen Fakultäten in Deutschland methodisch versierte Institute (z.B. Biometrie-Abteilungen) angebunden, die hierbei entsprechend ihre Expertise einbringen könnten.

Die Analyse auf Cluster-Ebene ist die einfachste Auswertemethode einer cluster-randomisierten Studie und kann als zweistufiger Prozess angesehen werden: Zunächst wird für jedes Cluster ein Summenmaß berechnet (erste Stufe), welches dann mit einem geeigneten statistischen Test verglichen wird (zweite Stufe), siehe z. B. [16]. In Studien in der Lehrforschung können beispielsweise anstelle der individuellen Ergebnisse der Studierenden die Clusterdurchschnittswerte (z. B. die Durchschnittsnote für jede Seminargruppe) in der Analyse (z. B. gewöhnlicher Zweistichproben-t-Test) verwendet werden. Eine vereinfachte Berücksichtigung von Kovariaten ist über Regressionen möglich [13]. Die Analyse auf Cluster-Ebene ist robust insbesondere bei kleiner Clusteranzahl, hat jedoch den Nachteil, dass die Variabilität innerhalb der Cluster unberücksichtigt bleibt. Eine Alternative besteht in der

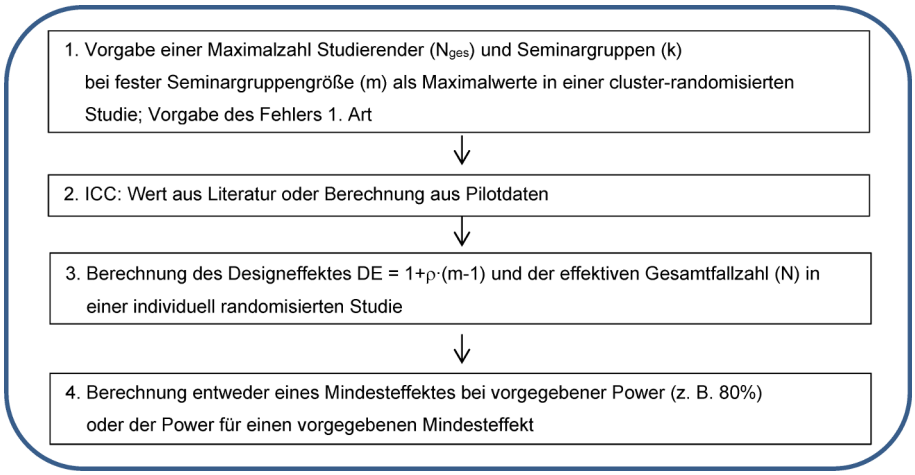

Abbildung 2: Schema für Berechnung von Power bzw. Mindesteffekt bei Studien in der Lehrforschung bei vorgegebener Fallzahl

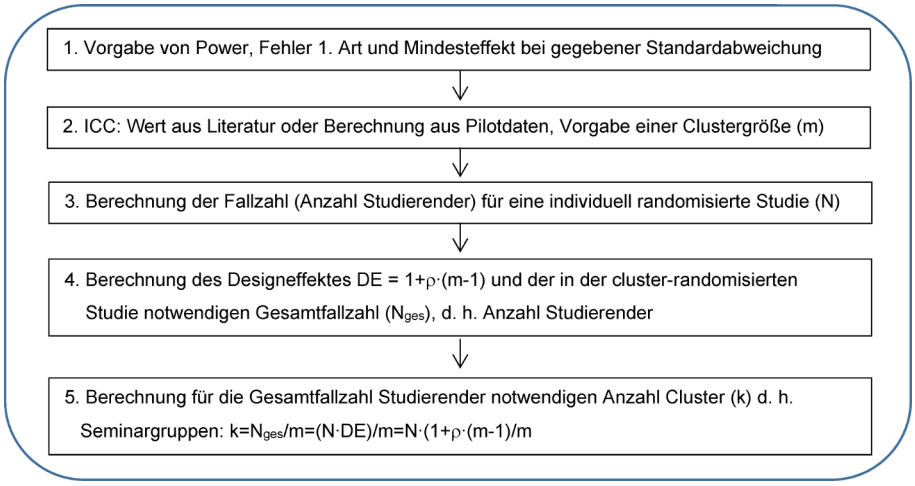

Abbildung 3: Schema für Fallzahlberechnung bei Studien in der Lehrforschung bei vorgegebener Power und Mindesteffekt

Anpassung univariater Teststatistiken (z. B. T-Wert beim T-Test) mit dem Designeffekt, wobei die individuellen Ergebnisse als statistisch unabhängig ausgewertet werden dürfen [15], [29].

Die Analyse auf Individuen-Ebene bietet speziell bei stark variierenden Clustergrößen eine Alternative, da die Analyse auf Cluster-Ebene in dieser Situation nicht so effizient ist. Ein einfaches statistisches Verfahren, welches auch eine Analyse auf Individuen-Ebene bietet, ist der adjustierte Zweistichproben-t-Test [28]. Sollen noch zusätzlich Kovariaten berücksichtigt werden, können Regressionsmodelle mit zufälligen Effekten, gemischte Regressionsmodelle oder verallgemeinerte Schätzgleichungen (GEE Modelle) angewendet werden. Diese Methoden erlauben auch die Berücksichtigung von Faktoren als potentielle Einflussgrößen für den Fall, dass -trotz bekannter prognostischer Faktoren- eine Stratifizierung bei der Cluster-Randomisierung nicht realisiert werden konnte. Im Vergleich zu den Verfahren für die Analyse auf Cluster-Ebene ist dies ein Vorteil, da die Effekte von Kovariaten auf gleicher Ebene wie der Effekt des Studienarms untersucht werden können (als Regressionskoeffizient mit Konfidenzintervall und p-Wert). Die Methoden für die Analyse auf Individuen-Ebene haben den Nachteil, dass sie weniger robust sind, wenn die Clusteranzahl klein ist. Eine Empfehlung ist daher die Verwendung von Methoden der Analyse auf Cluster-Ebene bei weniger als 15 bis 20 Cluster pro Studienarm [13]. Bei Studien mit größerer Clusteranzahl können die Methoden der Analyse auf Individuen-Ebene Vorteile bieten, insbesondere bei stark variabler Clustergröße.

4.6. Berichterstattung

Für die Berichterstattung randomisierter klinischer Studien wurde das CONSORT Statement entwickelt, welches von Campbell et al. für cluster-randomisierte Studien erweitert wurde [30]. Das erweiterte CONSORT-Statement nimmt Bezug auf die Besonderheiten einer cluster-randomisierten Studie und die Publikation einer solchen Studie sollte sich daran orientieren. So wird u.a. gefordert

- die Gründe für die Cluster-Randomisierung zu beschreiben
- die Einheit der Randomisierung und die der Intervention zu nennen
- neben der Anzahl der Individuen auch die Anzahl der Cluster und ihre Größe anzugeben
- die Strukturgleichheit nicht nur auf Individuen-Ebene sondern auch auf Cluster-Ebene zu zeigen
- den ICC (siehe oben) zu berechnen und zu berichten

- die Drop-outs auf Individuen- und auf Cluster-Ebene zu analysieren
- ein Flowchart zur Anzahl der Studienteilnehmer und Cluster im Studienablauf zu zeichnen

5. Anwendungsbeispiel

In diesem Abschnitt werden anhand eines Beispiels Planung, Durchführung und Auswertung einer cluster-randomisierten Studie in der Lehrforschung skizziert. Das Beispiel ist angelehnt an die NANA Studie [31], die zur Illustration von Studien in der klinischen Forschung dient. Die Studie wird als zweiarmige prospektive Beobachtungsstudie durchgeführt. Dabei werden die NAschkatzen (mit Vorliebe für Süßigkeiten) verglichen mit NAgetieren (mit Vorliebe für Knabberei) bezüglich Parametern wie z. B. dem Body Mass Index. Der Name hat aber auch einen Bezug zur Universität Ulm, vor der prominent eine große NANA-Figur steht (siehe Abbildung 4).

Abbildung 4: NANA vor Uni Ulm

Die cluster-randomisierte Studie soll prüfen, ob die Anwendung eines neuen „aktiven Seminarkonzepts“ (anhand der NANA-Studie) in der Biometrieausbildung von Studierenden der Medizin Einfluss auf das Prüfungsergebnis hat. Das „aktive Seminarkonzept“ (Planung, Durchführung und Auswertung einer kleinen empirischen Untersuchung während des Seminars) soll verglichen werden mit dem bisherigen Standardkonzept (Behandlung von Übungsaufgaben in Form eines „klassischen Seminars“). Für die Studie (balanziert, prospektiv, cluster-randomisiert) sollen ganze Seminargruppen entweder in einen Testarm („aktives Seminarkonzept“) oder einen Kontrollarm („klassisches Seminar“) randomisiert werden. Die Studie soll an der Medizinischen Fakultät der Universität Ulm während eines Wintersemesters durchgeführt werden.

Es ist von insgesamt maximal etwa 320 Studierenden auszugehen, die in Seminargruppen von ca. 20 Studierenden von jeweils einem Dozierenden betreut werden. Hieraus ergibt sich eine maximale Anzahl von 16 Clustern (d. h. Seminargruppen) in der Gesamtstudie (d. h. 8 Cluster pro Studienarm), was eher einer kleineren Studie entspricht [32]. Ein mögliches Ergebnis der Cluster-Randomisierung für das Beispiel zeigt Abbildung 5.

Die primäre Zielgröße ist die erreichte Punktezahl in der Klausur, gemessen bei den einzelnen Studierenden (d. h. auf der individuellen Ebene). Da neben dem Einfluss des Dozierenden noch die Gruppenzusammensetzung eine Rolle spielt, ist davon auszugehen, dass die Klausurergebnisse von Studierenden innerhalb der Seminargruppen ähnlicher sind als Ergebnisse von Studierenden verschiedener Seminargruppen. Für die Zielgröße wird angenommen, dass sie metrisch und annähernd normalverteilt ist.

Pilotdaten: Als Pilotdaten stehen die Ergebnisse der Kohorte des Wintersemesters 2015/2016 zur Verfügung: Es wurde ein arithmetisches Mittel von 92,5 Punkten (maximale Punktzahl: 120) bei einer Standardabweichung von 9,16 Punkten ermittelt (siehe Tabelle 2).

Die Ergebnisse der Pilotdaten werden für den Kontrollarm verwendet. Für den Testarm wird angenommen, dass sich die Punktezahl im Mittel um 3 Punkte verbessert (d. h. von 92,5 auf 95,5). Der ICC für die Zielgröße Punktezahl wurde mittels eines linearen gemischten Regressionsmodells geschätzt [12]: Im Ergebnis der Modellanpassung wurde für die Varianzen σ_b^2 = 1,67 und σ_w^2 = 82,36 erhalten, so dass der ICC als ρ=1,67/(1,67+82,63)=0,02 geschätzt wird. Der Designeffekt (DE=1+ρ·(m-1)) ergibt sich damit aus DE=1+0,02·(20-1)=1,38, wobei m die mittlere Clustergröße bezeichnet (hier: Anzahl Studierender pro Seminar; m=20).

Studienplanung: Für die Fallzahlberechnung der Beispielstudie sollen im Folgenden beide der in Abschnitt 4.4 genannten Methoden angewendet werden. Zunächst wird der explorative Ansatz beschrieben.

Explorative Methode: Berechnung von Power bzw. Mindesteffekt bei gegebener Fallzahl

Eine Umsetzung der Schritte 1. bis 4. des Schemas in Abbildung 2 ist im Folgenden für die Beispielstudie beschrieben. Ausgehend von der Maximalzahl von 320 Studierenden pro Semester sind höchstens 16 Cluster (Seminargruppen mit je 20 Studierenden) in der zu planenden Studie möglich. Bei einem Designeffekt von 1,38 (Berechnung: siehe oben) entspricht die Fallzahl von maximal 320 Studierenden in der cluster-randomisierten Studie einer effektiven Fallzahl von maximal etwa 320/1,38~232 Studierenden (gerade Anzahl wegen 1:1 Randomisierung) in einer individuell randomisierten Studie (d. h. 116 Studierende pro Studienarm). Um eine Power von 80% bei einem zweiseitigen Fehler 1. Art von 5% zu erreichen, ist bei dieser Fallzahl mit dem Zweistichproben-t-Test ein Mindestunterschied von 3,4 Punkten notwendig (bei einer Standardabweichung von 9,16), was einer Effektstärke von 0,37 nach Cohen entspricht (kleiner Effekt). Wird von einem Unterschied von 3 Punkten

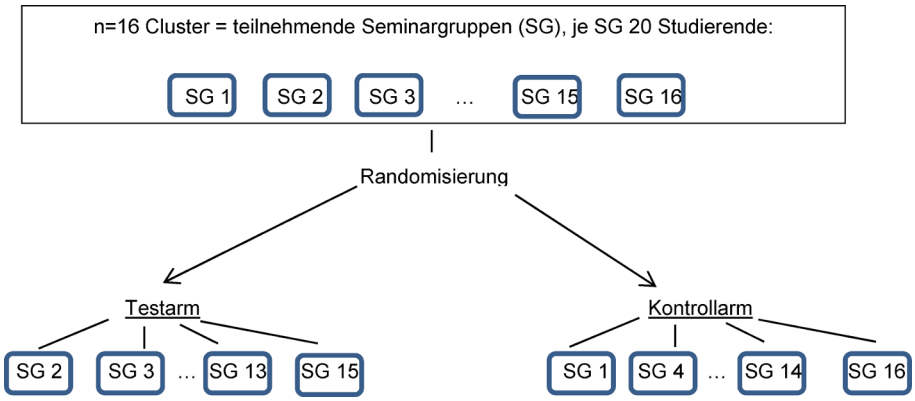

Abbildung 5: Beispielhaftes Ergebnis der Cluster-Randomisierung für die Beispielstudie

Tabelle 2: Ergebnisse der Kohorte des Wintersemesters 2015/2016: Arithmetische Mittel und Standardabweichungen der Punktzahlen in der Gesamtgruppe und in den einzelnen Kursgruppen

Gruppe	Arithmetisches Mittel	Standardabweichung
A1	90,5	6,86
A2	93,1	5,78
A3	94,6	8,97
A5	94,1	8,02
A6	88,7	7,77
A8	95,3	6,38
B1	90,4	16,92
B2	89,7	10,34
B3	96,4	5,38
C1	92,7	9,92
C2	92,3	8,47
C3	93,4	7,50
Gesamt	92,5	9,16

Tabelle 3: Auswirkungen der Größenordnung des ICC auf Mindesteffekt (a) bzw. Power (b) bei einer vorgegebenen Anzahl von 320 Studierenden in 16 Seminargruppen mit je 20 Studierenden. ESS = effektive Fallzahl (kursiv=Studiensituation)

a					b			
ICC	DE	ESS	Mindesteffekt (für Power 80%)		ICC	DE	ESS	Power (für Mindesteffekt 3 Punkte)
0,01	1,19	270	3,2		0,01	1,19	270	77
0,02	*1,38*	*232*	*3,4*		*0,02*	*1,38*	*232*	*70*
0,03	1,57	204	3,7		0,03	1,57	204	64
0,05	1,95	166	4,0		0,05	1,95	166	56

ausgegangen (ursprüngliche Planung), wird bei einem zweiseitigen Fehler 1. Art von 5% eine Power von nur 70% erreicht (bei gegebener Standardabweichung von 9,16). Tabelle 3 zeigt die Auswirkungen der Größenordnungen des ICC auf Mindesteffekt und Power. Die Berechnungen sind mit dem Zweistichproben-t-Test erfolgt, unter der Annahme gleicher Varianzen in beiden Armen.

Die Planung und Durchführung der Beispielstudie nach dieser explorativen Methode ist pragmatisch und erscheint in vielen Fällen realistischer als die Anwendung der konfirmatorischen Methode, da bei dieser oft unrealisierbar hohe Fallzahlen berechnet werden.

Konfirmatorische Methode: Berechnung der Fallzahl bei vorgegebener Power und Mindesteffekt

Eine Umsetzung der Schritte 1. bis 5. des Schemas in Abbildung 3 ist im Folgenden für die Beispielstudie beschrieben. Für die Zielgröße wird angenommen, dass sie metrisch und annähernd normalverteilt ist. Für den Testarm wird angenommen, dass sich die Punktezahl im Mittel um 3 Punkte verändert (d. h. von 92,5 auf 95,5, siehe oben). Die Berechnungen sind mit dem Zweistichproben-t-Test erfolgt, unter der Annahme gleicher Varianzen in beiden Armen (9,16, siehe oben). Für die Fallzahlplanung werden eine Power von 80% und ein Fehler 1. Art von 5% (zweiseitig) angenommen. Aus diesen Angaben ist zunächst die Fallzahl für eine individuell randomisierte Studie zu berechnen (z. B. [12]): Es ergibt sich eine Anzahl von 148 Studierenden pro Studienarm (296 Studierende insgesamt), die in die Studie bei individueller Randomisierung (also ohne Berücksichtigung der Clusterung) eingeschlossen werden müssen. Diese Anzahl muss nun noch um den Designeffekt (DE=1,38, Berechnung siehe

Tabelle 4: Auswirkungen der Größenordnung des ICC und Größe der Seminargruppen auf Gesamtfallzahl und Anzahl der Seminargruppen in der Gesamtstudie (kursiv=Studiensituation)

ICC (ρ)	Seminargruppengröße (m)	Designeffekt (DE)	Gesamtfallzahl (N_{ges})	Anzahl Seminargruppen (k)
0,01	20	1,19	353	18
0,01	15	1,14	338	24
0,01	10	1,09	323	34
0,02	*20*	*1,38*	*409*	*22*
0,02	15	1,28	379	26
0,02	10	1,18	350	36
0,03	20	1,57	465	24
0,03	15	1,42	421	30
0,03	10	1,27	376	38
0,05	20	1,95	578	30
0,05	15	1,70	504	34
0,05	10	1,45	430	44

oben) korrigiert werden: Pro Studienarm müssten 148·1,38~205 Studierende eingeschlossen werden (Gesamtstudie: 409 Studierende was insgesamt ca. k=21 Seminargruppen bedeuten würde). Die Auswirkungen der Größenordnung des ICC und Größe der Seminargruppen auf die Gesamtfallzahl und Anzahl der Seminargruppen ist für die oben beschriebenen Effekte der Beispielstudie in Tabelle 4 enthalten. Die Gesamtfallzahl wurde auf die nächste ganze Zahl gerundet. Die Anzahl der Seminargruppen wurde auf die nächste gerade Zahl gerundet, da in der Beispielstudie eine 1:1 Randomisierung vorgenommen werden soll. Dadurch ist die tatsächliche Gesamtfallzahl höher als die in der Spalte N_{ges} genannte Gesamtfallzahl, womit die Power höhere Werte als 80% erreicht.

Aufgrund der gegebenen Rahmenbedingungen (maximal 320 Studierende, Seminargruppengröße m=20) ist damit die Studie während eines Semesters nicht durchführbar. Eine Durchführung der Studie über mehrere Semester oder als multizentrische Studie erscheint wegen zu starker Unterschiede zwischen verschiedenen Jahrgängen (Studierende und Dozierende, weitere Rahmenbedingungen) oder Universitäten nicht empfehlenswert. Eine konfirmatorische Studie ist in diesem Setting also nicht realistisch. In einer solchen Situation erscheint daher der zuerst genannte explorative Ansatz empfehlenswert, d. h. eine Berechnung von Power bzw. Mindesteffekt bei fester gegebener Fallzahl.

Eine Modifikation des Designs wäre die Anwendung einer stratifizierten Cluster-Randomisierung nach Wochentag (Dienstag, Donnerstag, Freitag).

Auswertung: Aufgrund der eher kleinen Studie mit einer geringen Clusteranzahl und der nahezu konstanten Clustergröße erscheint für die statistische Auswertung eine Analyse auf Cluster-Ebene empfehlenswert (vergleiche Abschnitt 4.5). Dies kann beispielsweise mittels Berechnung von Clusterdurchschnittswerten aus den Ergebnissen der Studie und der Anwendung des Zweistichproben-t-Tests realisiert werden. Weitere Beispiele mit praktischen Darstellungen von Auswertungen cluster-randomisierter Studien sind in [30], [32] und [33] enthalten.

6. Diskussion und Empfehlungen

Neben anderen Studiendesigns (wie beispielsweise Beobachtungsstudien) werden in der Lehrforschung auch häufig prospektive zweiarmige (Interventions)-Studien zum Vergleich verschiedener Lehrmethoden angewendet. Hierbei sollten anerkannte Standards und Methoden wissenschaftlicher Untersuchungen eingehalten werden. Dies sind insbesondere das Vorhandensein eines Kontrollarms und das Erreichen von statistischer Gleichheit (Strukturgleichheit (durch Randomisierung, ggf. Stratifizierung), Behandlungsgleichheit, Beobachtungsgleichheit). Ohne relevante Gründe sollten vergleichende wissenschaftliche Studien nicht mehr ohne Kontrollarm durchgeführt werden. Aber auch quasi-experimentelle Studien mit Kontrollarm, jedoch ohne Randomisierung sollten vermieden werden. Ein wesentlicher Kritikpunkt an den Ergebnissen solcher Studien ist der Mangel an Strukturgleichheit verbunden mit der Gefahr vermengter Effekte. Theoretisch kann man sich dagegen schützen, indem man die Studienteilnehmer den Studienarmen streng zufällig zuweist, entweder durch individuelle Randomisierung oder Cluster-Randomisierung. Wegen der Vorteile der Randomisierung sollte der zusätzliche Aufwand, wenn irgend möglich, in Kauf genommen werden, vor allem, da dieser im Vergleich zum Aufwand der gesamten Studie gering ist: Die Durchführung einer Studie erfordert meist viele Ressourcen, die Randomisierung dagegen vergleichsweise wenig. Der Gewinn an Interpretierbarkeit und Aussagekraft der Studienergebnisse ist aber enorm. Im Vergleich zu Studien aus anderen Bereichen gibt es in der Lehrforschung jedoch einige besondere Bedingungen und Anforderungen, welche die Planung, Durchführung und Auswertung von Studien beeinflussen. Wegen

des Vorliegens natürlicher Cluster kann eine Randomisierung in diesem Bereich meist nur als Cluster-Randomisierung realisiert werden, bei einer limitierten Anzahl Studierender und einer vorgegebenen annähernd konstanten Clustergröße. Weiterhin muss die räumliche und zeitliche Verfügbarkeit verschiedener Ressourcen wie Dozierende, Seminarräume, Labore, Hörsäle, Computerpools beachtet werden. Bei der Fallzahlplanung ist die Clusterstruktur zu berücksichtigen, da die Ergebnisse Studierender innerhalb der Cluster (z. B. innerhalb von Seminargruppen) ähnlicher sind als Ergebnisse Studierender aus verschiedenen Seminargruppen. Je nach Stärke dieser Ähnlichkeit (gemessen über den ICC) kann die zum Erreichen einer bestimmten Power notwendige Fallzahl bei cluster-randomisierten Studien deutlich über der Fallzahl einer entsprechenden individuell randomisierten Studie liegen. Daher werden viele Studien in der Lehrforschung aufgrund ihrer limitierten maximal möglichen Fallzahl (aus Gründen der Machbarkeit) lediglich explorativen Charakter besitzen. Speziell hier ist die Strukturgleichheit wichtig, damit gefundene Unterschiede mit den in der Studie untersuchten Methoden erklärt werden können. Auch bei der statistischen Auswertung cluster-randomisierter Studien ist auf eine adäquate statistische Methodik zu achten, die die aus der Clusterstruktur sich ergebenden Abhängigkeiten angemessen berücksichtigt. Wegen der komplexen statistischen Methoden, die in allen Phasen einer cluster-randomisierten Studie notwendig sind, ist bei der praktischen Durchführung solcher Studien Unterstützung durch einen kompetenten Experten mit entsprechenden Spezialkenntnissen empfehlenswert. Dies können beispielweise wissenschaftliche Mitarbeiter von biometrischen Institutionen sein, welche es an den meisten Universitäten mit einer Medizinischen Fakultät gibt.

Neben den in Tabelle 1 genannten Nachteilen besteht in cluster-randomisierten Studien -im Vergleich zu konventionell randomisierten Studien- eine höhere Gefahr, dass die Strukturgleichheit auf Individualebene nicht erreicht wird. Dies kann die interne Validität gefährden, welche auch wegen der meist fehlenden Verblindung bei cluster-randomisierten Studien kritisch zu hinterfragen ist [12]. Hier muss im Rahmen der statistischen Auswertung eine Adjustierung für die ungleich verteilten Merkmale erfolgen, z. B. durch ein geeignetes Regressionsverfahren [12], [13]. Wie bei allen klinischen Studien kann auch bei cluster-randomisierten Studien bei erfüllter interner Validität die externe Validität nur heuristisch begründet werden. Dies ist in der Lehrforschung vermutlich schwieriger als in klinischen Studien, da die Bedingungen an den verschiedenen Lehreinrichtungen zu verschieden sind. Wegen der höheren Fallzahlen und der komplexeren Methodik sollte deshalb gerade bei Studien in der Lehrforschung in der Planungsphase überlegt werden, ob eine Cluster-Randomisierung gerechtfertigt und notwendig ist [34].

Abschließend fassen die folgenden Empfehlungen wesentliche Maßnahmen zur Qualitätssicherung von prospektiven zweiarmigen Studien in der Lehrforschung unter Berücksichtigung von Clustern zusammen.

1. Lehre wird meist in Gruppen von Studierenden durchgeführt, so dass eine natürliche Cluster-Struktur gegeben ist, was zu einer Cluster-Randomisierung führt.
2. Die Cluster-Randomisierung muss bei Studiendesign, Fallzahlplanung, Auswertung und Berichterstattung berücksichtigt werden.
3. In eine cluster-randomisierte Studie sollten nicht zu wenige Cluster eingeschlossen werden: Weniger als 8-10 Cluster sollten nicht eingeschlossen werden [32].
4. Bei sehr wenigen oder stark unterschiedlichen Clustern kann ein Matching von Clustern sinnvoll sein.
5. Verblindung ist meist nicht möglich. Die Verwendung möglichst objektiver Zielgrößen und eine verblindete Bewertung, wie beispielsweise die Bewertung von Ergebnissen im PBL durch unabhängige und nicht an der Studie beteiligte Personen, ist daher empfehlenswert und dient der Verbesserung der internen Validität.
6. Möglichst Aufrechterhaltung der Strukturgleichheit: Schaffung gleicher Bedingungen wie z. B. Uhrzeiten, Seminarräume für die zu vergleichenden Studienarme.

Auf Grund unserer Erfahrungen und der hier genannten Argumente empfehlen wir bei prospektiven zweiarmigen vergleichenden Studien die Nutzung von Kontrollarmen und eine adäquate Randomisierung, um auch in der Lehrforschung gute und überzeugende Ergebnisse zu erreichen.

Insbesondere die Cluster-Randomisierung kann hierbei ein entscheidender Baustein sein, der daher bei Studien im Bereich der Lehrforschung verstärkt genutzt werden sollte.

Danksagung

Wir bedanken uns bei Jacquie Klesing, Board-certified Editor in the Life Sciences (ELS) und Übersetzerin, für ihre Unterstützung mit dem Manuskript.

Interessenkonflikt

Die Autoren erklären, dass sie keine Interessenkonflikte im Zusammenhang mit diesem Artikel haben.

Literatur

1. Buss B, Wagner R, Bauder M, Fenik Y, Riessen R, Lammerding-Köppel M, Gawaz M, Fateh-Moghadam S, Weyrich P, Celebi N. Student tutors for hands-on training in focused emergency echocardiography – a randomized controlled trial. BMC Med Educ. 2012;12:101. DOI: 10.1186/1472-6920-12-101

2. Herter DA, Wagner R, Holderried F, Fenik Y, Riessen R, Weyrich P, Celebi N. Effect of supervised students' involvement on diagnostic accuracy in hospitalized medical patients–a prospective controlled study. PLoS One. 2012;7(9):e44866. DOI: 10.1371/journal.pone.0044866
3. Werner A, Holderried F, Schäffeler N, Weyrich P, Riessen R, Zipfel S, Celebi N. Communication training for advanced medical students improves information recall of medical laypersons in simulated informed consent talks - a randomized controlled trial. BMC Med Educ. 2013;1:13-15. DOI: 10.1186/1472-6920-13-15
4. Herrmann-Werner A, Nikendei C, Keifenheim K, Bosse HM, Lund F, Wagner R, Celebi N, Zipfel S, Weyrich P. Best practice" skills lab training vs. a "see one, do one" approach in undergraduate medical education: an RCT on students' long-term ability to perform procedural clinical skills. PLoS One. 2013;8(9):e76354. DOI: 10.1371/journal.pone.0076354
5. Ackel-Eisnach K, Raes P, Hönikl L, Bauer D, Wagener S, Möltner A, Jünger J, Fischer MR. Is German Medical Education Research on the rise? An analysis of publications from the years 2004 to 2013. GMS Z Med Ausbild. 2015;32(3):Doc30. DOI: 10.3205/zma000972
6. Schumacher M, Schulgen G. Methodik Klinischer Studien, Methodische Grundlagen der Planung, Durchführung und Auswertung. 3. Auflage. Heidelberg: Springer Verlag; 2008.
7. Armitage P. The role of randomization in clinical trials. Stat Med. 1982;1:345-352. DOI: 10.1002/sim.4780010412
8. Boet S, Sharma S, Goldman J, Reeves S. Review article: medical education research: an overview of methods. Can J Anaesth. 2012;59(2):159-170. DOI: 10.1007/s12630-011-9635-y
9. Fisher LD. Ethics of Randomized Trials. In: Armitage P, Colton T (Hrsg). Encyclopedia of Biostatistics. Chichester: Wiley & Sons Ltd; 1998. P.1394-1398.
10. Gaus, W, Muche, R. Medizinische Statistik. Stuttgart: Schattauer Verlag; 2013.
11. Korzilius H. EU-Verordnung über klinische Prüfungen: Kompromiss verabschiedet. Dtsch Ärztebl. 2014;5.
12. Eldridge SM, Kerry S. A Practical Guide to Cluster Randomised Trials in Health Services Research. Weinheim: Wiley; 2012. DOI: 10.1002/9781119966241
13. Hayes RJ, Moulton LH. Cluster Randomised Trials. Oxford: Oxford University Press; 2009. DOI: 10.1201/9781584888178
14. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. Int Stat Rev. 2009;77:378-394. DOI: 10.1111/j.1751-5823.2009.00092.x
15. Chenot JF. Cluster-randomisierte Studien: eine wichtige Methode in der allgemeinmedizinischen Forschung. Z Evid Fortbild Qual Gesundheitswes. 2009;103(7):475-480. DOI: 10.1016/j.zefq.2009.07.004
16. Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. BMJ. 1998;316(7142):1455. DOI: 10.1136/bmj.316.7142.1455
17. Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. Stat Med. 2002;21:3757-3774. DOI: 10.1002/sim.1330
18. Donner A. A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model. Int Stat Rev. 1986;54(1):67-82. DOI: 10.2307/1403259
19. Ridout MS, Demétrio CG, Firth D. Estimating intraclass correlation for binary data. Biometrics. 1999;55(1):137-148. DOI: 10.1111/j.0006-341X.1999.00137.x
20. Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp Clin Trials. 2012;33(5):869-880. DOI: 10.1016/j.cct.2012.05.004
21. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. J Clin Epidemiol. 2004;57(8):785-794. DOI: 10.1016/j.jclinepi.2003.12.013
22. Hemming K, Girling AJ, Sitch AJ, Marsh J, Lilford RJ. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. BMC Med Res Methodol. 2011;11:102. DOI: 10.1186/1471-2288-11-102
23. Dang Q, Mazumdar S, Houck PR. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. Comput Methods Programs Biomed. 200;91(2):122-127.
24. Dreyhaupt J. Instrumente für Power- und Fallzahlberechnungen bei komplexen hierarchischen Studiendesigns in der Versorgungsforschung. Monit Versorgungsforsch. 2015;6:49-54.
25. Dreyhaupt J. Generelle Fallzahl- und Powerabschätzung über Simulation bei Studien mit komplexen hierarchischen Daten als Unterstützung der Studienplanung in der Versorgungsforschung. Ulm: Universität Ulm; 2015. Zugänglich unter/available from: URL: http://vts.uni-ulm.de/query/longview.meta.asp?document_id=9509
26. Landau S, Stahl D. Sample size and power calculations for medical studies by simulation when closed form expressions are not available. Stat Methods Med Res. 2013;22(3):324-345. DOI: 10.1177/0962280212439578
27. Bland JM, Kerry SM. Trials randomised in clusters. BMJ. 1997;315(7108):600. DOI: 10.1136/bmj.315.7108.600
28. Donner A, Klar N. Design and Analysis of Cluster Randomization trials in Health Research. Weinheim: John Wiley & Sons, Ltd; 2010.
29. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. Fam Pract. 2000;17(2):192-196. DOI: 10.1093/fampra/17.2.192
30. Campbell MK, Piaggio G, Elbourne DR, Altman DG; CONSORT Group (2012). Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012. DOI: 10.1136/bmj.e5661
31. Mayer B, Danner B. Von Naschkatzen und Nagetieren – Eine interaktive Einführung in die Medizinische Biometrie mit der NANA-Studie. In: Rauch G, Muche R, Vonthein R (Hrsg). Zeig mir Biostatistik! Ideen und Material für einen guten Biometrie-Unterricht. Heidelberg: Springer Verlag; 2014. S.3-14. DOI: 10.1007/978-3-642-54336-4_1
32. Eldridge SM, Costeloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? Stat. Methods Med Res. 2016:1039-1056. DOI: 10.1177/0962280215588242
33. Campbell MK Analysis of cluster randomized trials in primary care: a practical approach. BMJ. 1998;316:1455.
34. Kuß O, Jahn P, Renz P, Landenberger M. Cluster-randomisierte Studien in der Pflegewissenschaft. Halle Beitr Gesundheit Pflegewissenschaft. 2009;8(1):302-310.

Korrespondenzadresse:
Dr. Jens Dreyhaupt
Universität Ulm, Institut für Epidemiologie und Medizinische Biometrie, Schwabstr. 13, 89075 Ulm, Deutschland, Telefon: +49(0)731/50-26895, Fax: +49(0)731/50-26902
jens.dreyhaupt@uni-ulm.de

Bitte zitieren als
Dreyhaupt J, Mayer B, Keis O, Öchsner W, Muche R. Cluster-randomized Studies in Educational Research: Principles and Methodological Aspects. GMS J Med Educ. 2017;34(2):Doc26.
DOI: 10.3205/zma001103, URN: urn:nbn:de:0183-zma0011038

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2017-34/zma001103.shtml

Eingereicht: *16.08.2016*
Überarbeitet: *17.11.2016*
Angenommen: *29.12.2016*
Veröffentlicht: *15.05.2017*