HUMBOLDT-UNIVERSITÄT ZU BERLIN

Scalar-on-composition regression to evaluate the impact of class composition on educational achievement

Master's Thesis

for acquiring the degree of
Master of Science (M.Sc.)

in Statistics

at the School of Business and Economics
of Humboldt-Universität zu Berlin

submitted by

Gesa Julia Kröger

Examiner: Prof. Dr. Sonja Greven
Dr. Felix Weinhardt

Berlin, 06.11.2021

Abstract

In recent years the influence of group behaviour, the so-called peer effect, on individuals has been of interest. Therefore, a new form of peer effect which is of compositional character is considered. Using a composition as peer effects would offer a possibility to consider how the distribution of the peers educational achievements based on their test scores influences an individual of the cohort while they attend school. To analyse the effect of compositional peer effects, the methods of compositional data analysis are used. These methods are applied on the data set of Project STAR containing information about students throughout their whole school career. The compositional term is based on the distribution of test scores at the beginning of the project, when the students were in kindergarten, within each student's class. To analyse the influence of such terms, zero imputation methods and ilr-transformation are used to apply classical statistical models. In the first step, the impact of the used zero imputation methods and interval selection of the continuous variable are studied. Then the influence of the compositional peer effect, containing the information of the distribution of the students inside the same class, is analysed. These analyses show that there are indeed significant impacts on an individual's subsequent test scores based on the distribution of test scores in their class. The higher the ratio of students with higher scores was in kindergarten, the more the individual test score in the following years decreased and vice versa. However, using the distribution of the students in the first, second or third grade as covariate changes this effect. In this case, the higher the ratio of students scoring high is, the more the individual score increased in the subsequent years.

Contents

List of Tables

List of Figures

1 Motivation

In recent years, social scientists have raised the question how the composition of groups can affect the individual and overall performance. To understand the determinants of a groups performance is important in a lot of modern work contexts. This can include interactions while working on complex tasks as a group or the interaction of students studying within the same class. How social interactions influence the performance of individuals is the subject of various theories which are not just considering group performance to be the average individual performance, but rather are considering the median or geometric mean instead for the average performance of an individual. The essential problem is that the group performance is a endogenous effect and the data contains a non-observed selection bias which is challenging in studies containing group behaviour.

The analysis of the impact of group compositions was mainly focused on the educational sector, due to the general setting that students are put into classes (cohorts) and stayed within these as long as they attend the same school. The understanding if and how students learn from each other within the group is needed to understand how the individual and overall performance can be changed though class reorganization. The social interactions between students throughout their education are the so-called ”peer effects” and the students which are part of the same group at some point in time are called ”peers”.

The analysis of peer effects is complicated, because of endogenous group formations and selection bias. However, experiments where students and teachers were randomly assigned into different classes and are followed through their educational careers exist and avoid the problem of endogenous variables. The general model used for an estimation of the impact of peer effects on an individuals performance, assuming random assignment into groups, is expressed by

$$Y_{ict} = \alpha + \beta_1 Y_{-i}^{-t} + X_{ic}'\delta + \varepsilon_{ict}, \tag{1.1}$$

where Y_{ict} is the outcome of individual i at time t belonging to the group c. Y_{-i}^{-t} is the peer variable which is the average test score of all other individuals (except i) of the group and based on a test score of time $-t$, time $-t$ being a point in time before the formation of group c. X_{ic} denotes the matrix of individual or group-level variables affecting the outcome. Lastly ε_{ict} is the error term which is assumed to be random. For a group structure, i.e. the separation of individuals into cohorts, a within-group correlation can be assumed, so that only the across-group variation is used for the estimation of standard errors in the estimators α, β_1 and δ.

Using the previous setting expressed in equation 1.1, it is possible to utilize outcome variables of the individual which are in the future, while the affiliation to a groups is based on a past group of the individual. For example, considering middle or high school scores of a nationally comparable exam while the group structure is based on classes in elementary school. If the data on both points in time are known for a set of students, the data sets can be related and analysed based on the group structure.

In the case of equation 1.1 the estimation is giving the influence of average peer performance on the individual outcomes. This kind of analysis can be extended to estimate these effects based on ability grouping to get an understanding whether high an low performing peers are affected in the

same way by the average peer abilities.

The literature covered a big range of treatment variables, e.g. the share of nationally very good and bad peers (Lavy et al., 2012) or the rank of a student in the ability distribution (Murphy and Weinhardt, 2020). The share of females, racial composition (Whitmore, 2005) and peer personalities (Golsteyn et al., 2017) have also been considered as non-cognitive measures.

In this thesis the grouped structure of a data set is used to analyse the impact of a compositional peer effect. Therefore, the peer effect is considered to be compositional, e.g. percentages or proportions, and the influence of each composition on the individual outcome is analysed. The analysis requires the techniques of the compositional data analysis which will be introduced in section 2 and the limitation and problems in the application are discussed. Then the analysis is performed on a data set from the project called student-teacher-achievement-ratio (STAR) which is well-known and was studied a lot due to its random assignment of students and teachers to classrooms of small and regular size. Project STAR has been used in several studies on peer effects controlling for the class size (Whitmore and Krueger, 2001) and non-cognitive skills (Bietenbeck, 2019). Other studies considered peer effects on girls achievements (Whitmore, 2005) and long-term effects on the earnings (Chetty et al., 2011). The data set and execution of the project STAR is going to be elaborated in section 3.

A more flexible regression technique such as compositional data analysis to this data set from Project STAR can be applied for several reasons. As peers do only occur in groups, it is of advantage to consider a distribution of peer abilities instead of measures as the mean or standard deviation. Further, the peer effects can be different depending on the cognitive and non-cognitive skills, gender and race of the individuals belonging to a group. All these variables were covered during the long time span of the project and are available for estimation. Therefore, a large number of variables are available as dependent and independent variables leaving room for multiple possible approaches based on compositional data analysis. The applied model utilizing compositional peer effects is introduced in section 4 and the results are presented in section 5.

To measure the sensitivity of the model a simulation is performed in section 6 and section 7 is summarising the results collected in the previous sections and gives the concluding remarks.

2 Compositional Data Analysis

This section is discussing the definition of composition data and the related geometrical properties and principles of compositional data which have to be abided by are introduced. Afterwards, the fitting vector space satisfying the principles of compositional data as well as basic operations for this vector space are defined. On the basis of the vector space, the so-called Aitchison geometry, a set of representations of coordinates based on log-ratios and the transformations of compositional data into the Euclidean space are introduced. Furthermore, the density functions used in the context of compositional data as well as their application to linear models using compositional covariates are discussed.

2.1 Introduction to Compositional Data

Compositional data are multivariate observations which contain relative instead of absolute information. They are strictly positive and, in general, sum up to a constant. Typical examples are percentages and proportions. The difference between using absolute and relative information of the data, is that any rescaling of the data from the given units can affect the information carried for absolute data. However, relative data is not affected by any rescaling, since the ratios between the different components stay the same. It should be considered beforehand what the goal of the analysis is. If the interest is lying in absolute information, a non-compositional regression is favourable to a compositional regression (Filzmoser et al., 2018, pp. 1–2). The components of the compositional vector $x = [x_1, x_2, \cdots, x_D]$ are called parts, due to their compositional characteristics and x is called D-part composition.

Definition: A vector, $x = [x_1, x_2, \cdots, x_D]$, is a **D-part composition** when all its components are strictly positive real numbers and carry only relative information.

As explained above, the relative information lies in the ratio between the different components of the composition. In general, the information belongs to a whole, so that all parts are summing up to a constant κ. In this case the data is called closed. A typical example of closed data are percentages which sum up to $\kappa = 100$ (Pawlowsky-Glahn et al., 2015, pp. 8–9).

Nevertheless, it is also possible that a data set contains compositions for which only a part of the whole composition is known. This applies for example when parts per million or units like molarities, were no constant sum is feasible, are used. Even in this case it is possible to get proportions by considering only a subcomposition (Pawlowsky-Glahn et al., 2015, p. 11) and using transformation, introduced in section 2.3.2, thus the operations of compositional data can be applied. Due to this characteristic of compositions, Pawlowsky-Glahn et al. (2015) defined them as equivalence classes of proportional vectors. This leads to the possibility to select the constant sum vector as a representative of a composition using a fitting scaling factor on the proportions. The operation to assign a constant sum representative is called closure and results in a rescaled vector of the initial vector so that the sum over all components is κ (Pawlowsky-Glahn et al., 2015, pp. 8–9).

Definition: For any vector x of D strictly positive real components

$$x = [x_1, x_2, \cdots, x_D] \in \mathbb{R}_+^D, \quad x_i > 0, \quad \forall i = 1, 2, ..., D,$$

the **closure of x** to $\kappa > 0$ is defined as

$$\mathcal{C}(x) = \left(\frac{\kappa x_1}{\sum_{i=1}^D x_i}, \frac{\kappa x_2}{\sum_{i=1}^D x_i}, \cdots, \frac{\kappa x_D}{\sum_{i=1}^D x_i} \right).$$

Using the definition of closure, the two vectors x and y can be called compositionally equivalent for any constant κ, if $\mathcal{C}(x) = \mathcal{C}(y)$. This can be pictured by a ray through the origin on which all compositionally equivalent vectors are located (Filzmoser et al., 2018, p. 39). If only some parts of the composition are of interest, a composition does not have to be closed. In this case, there is the possibility to consider subcompositions, vectors containing only some parts of the whole composition, as defined in Pawlowsky-Glahn et al. (2015, p. 10).

Definition: Given a composition x and a selection of indices $S = i_1, ..., i_s$, a **subcomposition** x_s with s parts, is obtained by applying the closure operation to the subvector $[x_{i_1}, x_{i_2}, ..., x_{i_s}]$ of x. The set of subscripts S indicate which parts are selected in the subcomposition, not necessarily the first s ones.

2.2 Principles of Compositional Data

Since compositional data contain relative information and the standard statistical methods were implemented for absolute information in the Euclidean geometry, it can lead to problems or misleading results, if they are applied without considering the difference of the data format. To keep the properties of the data when applying methods, three main principles were formally introduced by Aitchison (1994). Any operation for compositions has to adhere these principles to not destroy the properties and information carried by the data.

Definition (Scale Invariance): The information carried by compositions is independent of their units. Even if there is no information about the total, the analysis should yield the same results for compositionally equivalent vectors or precisely, the composition should be invariant under any change of scale.

Definition (Permutation Invariance): Any change of order within the composition does not alter the information conveyed by the vector of the composition.

Definition (Subcompositional Coherence): Subcompositional coherence can be summarized as subcompositional dominance and ratio preserving. This means the results yielded by a composition should not be contradicted by the results of a subcomposition. The **subcompositional dominance** is restricting the distances between parts of a subcomposition to always be smaller or equal to the distances between the parts of the composition. **Ratio preserving** means that the ratio between any parts of the subcomposition should be the same as the analogue ratio in the composition.

While principles like permutation invariance and subcompositional dominance should apply to any statistical method, scale invariance is directly derived from the definition of compositional data, as the information conveyed by relative data stays the same, even if it is rescaled (Pawlowsky-Glahn et al., 2015, pp. 13–16 and Filzmoser et al., 2018, pp. 11–12). For example the vectors $x_1 = [1, 5, 2]$ and $x_2 = [1000, 5000, 2000]$ are the same composition, just using different scales and any statistical method applied to both separately has to lead to the same result, since the ratio between the parts is the same.

2.3 Geometrical Properties

Taking the standard methodologies of the Euclidean geometry to analyse a set of compositional data can yield misleading results. Due to this fact, a set of functions and techniques analogous to the ones in the Euclidean space are needed, including scalar product, Euclidean norm, Euclidean distance and an application of the basic mathematical operations in the Euclidean space. For this reason, Aitchison (1986) proposed the logarithm of a ratio (log-ratio) methodology, those idea is that compositional data are defined in another geometry in their sample space. This geometry was later on named after him and called Aitchison geometry.

2.3.1 Aitchison Geometry

For compositional data the sample space, called simplex, is characterised by a set of closed compositions, for which the defined sum constraint is not relevant in further considerations, due to the scale invariance (Filzmoser et al., 2018, pp. 37–38).

Definition: The **sample space** of compositional data is the **simplex**,

$$S^D = \left\{ x = [x_1, x_2, \cdots, x_D] \in \mathbb{R}^D \,\middle|\, x_i > 0, i = 1, 2, ..., D; \sum_{i=1}^{D} x_i = \kappa \right\}, \tag{2.1}$$

where κ is any arbitrary positive real number.

As already mentioned the methods applied to compositions and thus also the simplex, have to follow the three principles of compositional data. The Euclidean space does not fulfil these constraints in general. So that an appropriate set of methods and techniques using the log-ratio approach have been proposed by Aitchison (1982), helping to define a vector space for the simplex, known as Aitchison geometry, shown independently by Pawlowsky-Glahn and Egozcue (2001) and Billheimer et al. (1997, 2001).

Analogously to the operations of addition of two vectors and multiplying a vector by a constant in the Euclidean space, the operation perturbation and powering are defined on the simplex. Both operations fulfil the described requirements of section 2.2, the principles of compositional data and the requirements for operations of a vector space (Egozcue and Pawlowsky-Glahn, 2011, p. 17).

Definition: **Perturbation** of the compositions $x, y \in S^D$ is defined as composition

$$x \oplus y = \mathcal{C}(x_1 y_1, x_2 y_2, ..., x_D y_D). \tag{2.2}$$

Definition: **Powering** of a composition $x \in S^D$ by a real number α is defined as the composition

$$\alpha \odot x = \mathcal{C}(x_1^\alpha, x_2^\alpha, ..., x_D^\alpha). \tag{2.3}$$

Combining both operations one can also define a difference perturbation in the following way.

Definition: **Difference Perturbation** of two compositions $x, y \in S^D$ is defined as the composition

$$x \ominus y = x \oplus [(-1) \odot y] = \left(\frac{x_1}{y_1}, \frac{x_2}{y_2}, ..., \frac{x_D}{y_D} \right). \tag{2.4}$$

It can be easily shown that the neutral element n of perturbation is the composition containing all equal parts, i.e. in a D-part composition the neutral element is $n = (\frac{1}{D}, \cdots, \frac{1}{D})$. Then if the difference perturbation is applied to the element itself, it yields the neutral element. Up to this point the basic operations needed to obtain a vector space structure on the simplex have been defined. It is further possible to obtain an Euclidean vector space structure, if definitions for the inner product, norm and distance are defined in an Aitchison sense (Pawlowsky-Glahn et al., 2015, p. 24).

Definition: **Aitchison inner product** of two compositions $x = (x_1, x_2, ..., x_D)' \in S^D$ and $y = (y_1, y_2, ..., y_D)' \in S^D$ is defined as

$$\langle x, y \rangle_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \left(\frac{x_i}{x_j} \right) \ln \left(\frac{y_i}{y_j} \right). \tag{2.5}$$

Definition: **Aitchison norm** of a composition $x = (x_1, x_2, ..., x_D)' \in S^D$ is defined as

$$\|x\|_A = \sqrt{\langle x, x \rangle_A} = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \left(\frac{x_i}{x_j} \right)^2}. \tag{2.6}$$

Definition: **Aitchison distance** between two compositions $x = (x_1, x_2, ..., x_D)' \in S^D$ and $y = (y_1, y_2, ..., y_D)' \in S^D$ is defined as

$$d_A(x, y) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left[\ln \left(\frac{x_i}{x_j} \right) - \ln \left(\frac{y_i}{y_j} \right) \right]^2}. \tag{2.7}$$

The new Euclidean vector space structure on the simplex is called the Aitchison geometry. The name was chosen in regards to John Aitchison who pioneered in the analysis of compositional data. As mentioned above, all definitions are based on the log-ratio approach proposed by Aitchison (1982). For the log-ratios a total of $D(D-1)$ non-zero combinations exist for a D-part composition. However, as it was shown by Filzmoser et al. (2018) the total number is reduced to $D(D-1)/2$,

due to the symmetry of the log-ratios, since it holds that $\ln \frac{x_i}{x_j} = -\ln \frac{x_i}{x_j}$. It can further be shown, that with $D-1$ log-ratios it is possible to construct any other log-ratio with a linear combination of these log-ratios. This leads to the conclusion that a D-part composition can be represented by a $(D-1)$-dimensional subspace and will not lose any of the information (Filzmoser et al., 2018, pp. 41–43). Based on this idea coordinate representations are defined which help to express the D-part composition by $(D-1)$-dimensional coordinates in a Euclidean real space.

2.3.2 Coordinate Representations

As mentioned in the previous section, log-ratios can be a very convenient transformation for compositional data, since they have beneficial properties and are easy to handle - for instance, the inverse log-ratio produces the same results with the opposite sign and the results will lie in the positive real space as it is needed for compositions. If a log-ratio is also scale invariant, it is called logcontrast (Aitchison, 1986).

Definition: Consider a composition $x = (x_1, x_2, ..., x_D) \in S^D$ and some coefficients $\alpha_i \in \mathbb{R}, \forall i = 1, 2, ..., D$. A **logcontrast** is a function

$$f(x) = \sum_{i=1}^{D} \alpha_i \ln x_i, \quad \sum_{i=1}^{D} \alpha_i = 0. \tag{2.8}$$

Functions of logcontrasts build the basis to define different types of transformations. In a practical sense to analyse compositional data, a transformation to the Euclidean vector space will be applied to them and then standard statistical methods are employed over the transformed data. Transformations can be viewed as an expression of the original compositions onto the Euclidean geometry in the real space, while keeping the principles of compositional data (Filzmoser et al., 2018, pp. 43–44). This point of view can help to overcome the obstacle of interpreting the results of a compositional analysis.

The first transformations based on the log-ratio approach introduced by Aitchison (1986) were the **additive log-ratio transformation (alr)** and the **centered log-ratio transformation (clr)**. Historically the alr transformation was the first transformation used for modelling a classical statistical model with compositional data. The transformation is mapping the data from the D-dimensional simplex S^D to the $D-1$-dimensional real space $\mathbb{R}^{D-1}$.

Definition: For x a composition in the D-part simplex, the **alr transformation** is calculated by applying the natural logarithm componentwise

$$\mathrm{alr}(x) = \ln \left(\frac{x_1}{x_D}, \frac{x_2}{x_D}, ..., \frac{x_{D-1}}{x_D} \right). \tag{2.9}$$

The advantages of the alr transformation are that it is easy to invert the transformation and that operation like perturbation and powering can be reduced to simple operations in the real space. The downside of the alr transformation is that it is not invariant under permutation of components. Due to this fact, applying standard statistical methods can lead to problems or could not even be applicable

(Egozcue and Pawlowsky-Glahn, 2011, p. 20). Another disadvantage lies in the subjectivity of the choice of ratio-variable x_D and that the underlying coordinate system will be non-orthogonal. Additionally, operations like Aitchison inner product, norm and distance are in general not corresponding to the operations in the Euclidean space, unlike the perturbation and powering operations (Filzmoser et al., 2018, p. 45). Therefore, the alr transformation is only an isomorphism between S^D and $\mathbb{R}^{D-1}$ and not an isometry. Due to the multiple disadvantages the alr transformation is generally not used anymore.

In response to the problems of the alr transformation, Aitchison (1986) introduced the **centred log-ratio transformation (clr)**. Instead of using a specific component in the denominator, like the alr transformation, it uses the geometric mean and avoids the subjectivity of the choice.

Definition: For x a composition in the D-part simplex, the **clr transformation** is calculated by

$$x^* = \mathrm{clr}(x) = \ln\left(\frac{x_1}{g_m(x)}, \frac{x_2}{g_m(x)}, ..., \frac{x_{D-1}}{g_m(x)}\right), \quad g_m(x) = \sqrt[D]{\prod_{i=1}^{D} x_i}\,. \tag{2.10}$$

Exactly like the alr transformation, the operation perturbation and powering correspond to the sum and product in the real space of the Euclidean geometry and the inverse transformation can be easily calculated by

$$x = \mathrm{clr}^{-1}(x^*) = \mathcal{C}(\exp(x^*)). \tag{2.11}$$

In contrast to the alr transformation, the clr transformation maps the data from S^D to the D-dimensional real space $\mathbb{R}^D$ and the components of the clr transformation always sum up to zero. However, this means our composition has D parts, but only $D-1$ parts were needed to build a orthonormal basis in the Aitchison geometry as mentioned in the section 2.2. In this case our composition x^* can be seen as coefficients with respect to a generating system. A huge advantage of clr to alr is that it can be used to define a metric structure in the simplex. This means we can show that the operations of the Aitchison geometry hold (Filzmoser et al., 2018, p. 46 and Egozcue and Pawlowsky-Glahn, 2011, pp. 20–21).

Consider two compositions $x \in S^D$ and $y \in S^D$, then it holds that

$$\langle x, y\rangle_A = \langle \mathrm{clr}(x), \mathrm{clr}(y)\rangle, \quad \|x\|_A = \|\mathrm{clr}(x)\|, \quad d_A(x, y) = d(\mathrm{clr}(x), \mathrm{clr}(y)). \tag{2.12}$$

Thereby, the clr coefficients are representing an isometry. This means all concepts of the simplex are maintained after mapping it to the real space using clr transformation (Filzmoser et al., 2018, p. 48). This led to a new idea, with the goal to build an orthogonal basis and represent the composition by this form.

To take the step from clr, were the composition x from S^D was mapped to a $(D-1)$-dimensional hyperplane in $\mathbb{R}^D$, the orthonormal basis and corresponding coordinates should be build to transform the composition onto the $(D-1)$-dimensional real space. This approach of constructing orthonormal coordinates was introduced by Egozcue et al. (2003) and is called **isometric log-ratio transformation (ilr)**, due to the fact that these coordinates are logcontrasts and are an isometry. More precisely,

a transformation in the way of $\mathrm{irl}(\vec{e}_i) = e_i$, $\forall\, i = 1, 2, ..., D-1$ has to be constructed, with $\vec{e}_i$ being the i-th vector of the canonical basis in the $(D-1)$-dimensional real space (Egozcue et al., 2003, p. 294).

Definition: For any composition $x \in S^D$, the **ilr transformation** associated to an Aitchison-orthonormal basis in S^D, e_i, $i = 1, 2, ..., D-1$, is the transformation from S^D to $\mathbb{R}^{D-1}$ given by

$$x^* = \mathrm{ilr}(x) = \left(\langle x, e_1\rangle_A, \langle x, e_2\rangle_A, ..., \langle x, e_{D-1}\rangle_A\right), \quad \langle e_i, e_j\rangle_A = 0,\ i \neq j,\ \|e_i\|_A = 1. \tag{2.13}$$

This formula shows that the ilr coordinates can be identified with the coordinates of $x \in S^D$ with respect to either the orthonormal basis $e_1, e_2, ..., e_{D-1}$ or the canonical basis in $\mathbb{R}^{D-1}$, leading to $\mathrm{ilr}(x) = \sum \langle x, e_i\rangle_A \vec{e}_i$ (Egozcue et al., 2003, p. 295). The inverse expression then correnspponds to the following expression

$$x = \mathrm{ilr}^{-1}(x^*) = \bigoplus_{i=1}^{D-1} \left(\langle x^*, \vec{e}_i\rangle \odot e_i\right) = \bigoplus_{j=1}^{D-1} x_j^* \odot e_j. \tag{2.14}$$

2.3.3 Relationship between Coordinates

For later computations it can be of advantage to consider the relationship between the transformations, especially the relationship between the clr and ilr transformations. The ilr coordinates are, as mentioned above, constructed via an orthonormal basis. The corresponding orthonormal basis of ilr coordinates generally are of the form

$$u_i = \sqrt{\frac{i}{i+1}} \left(\underbrace{\frac{1}{i}, ..., \frac{1}{i}}_{i \text{ elements}}, -1, 0, ..., 0 \right), \qquad u_i \in \mathbb{R}^D,\ i = 1, 2, ..., D-1, \tag{2.15}$$

as it was shown by Egozcue et al. (2003) and it holds that $\mathrm{clr}(e_i) = u_i$, as a results of expressing the Aitchison orthonormal basis $e_i \in S^D$ as

$$e_i = C(\exp(u_i)) = C\left[\exp\left(\sqrt{\frac{1}{i(i+1)}}, ..., \sqrt{\frac{1}{i(i+1)}}, -\sqrt{\frac{i}{i+1}}, 0, ..., 0\right)\right], \quad i = 1, .., D-1, \tag{2.16}$$

which is orthonormal with respect to the Aitchison inner product. Another property shown by Aitchison (1986) is, that for any $x_1, x_2 \in S^D$ and $\alpha_1, \alpha_2 \in \mathbb{R}$ it holds that

$$\mathrm{clr}\left((\alpha_1 \odot x_1) \oplus (\alpha_2 \odot x_2)\right) = \alpha_1 \mathrm{clr}(x_1) + \alpha_2 \mathrm{clr}(x_2).$$

Considering the composition $\mathrm{ilr}(x) = (y_1, y_2, ..., y_{D-1})$ and using the inverse transformation (Eq. 2.14), the clr transformation of x can be rewritten as

$$\mathrm{clr}(x) = \mathrm{clr}\left(\bigoplus_{j=1}^{D-1} y_j \odot e_j\right) = \sum_{j=1}^{D-1} y_j \mathrm{clr}(e_j) = \sum_{j=1}^{D-1} y_j u_j = \mathrm{ilr}(x)U, \tag{2.17}$$

where U is the $(D-1) \times D$ matrix of the orthonormal basis vectors $\mathrm{clr}(e_j) = u_j$ as rows (Egozcue et al., 2003, pp. 296–297). The equation (2.17) shows how ilr coordinates can be rewritten as clr coordinates. The inverse relationship as shown by Egozcue et al. (2003, p. 290) is obtained by

$$\mathrm{ilr}(x) = [\langle x, e_1\rangle_A, \langle x, e_2\rangle_A, ..., \langle x, e_{D-1}\rangle_A] \overset{(1)}{=} \frac{1}{D}\mathrm{clr}(x)MU' = \mathrm{clr}(x)U', \tag{2.18}$$

where the step at (1) the Aitchison inner product expressed in the clr-transformed space is used and M is an $D \times D$ matrix having $D-1$ as diagonal values and -1 everywhere else. Thereby, having one transformation the compositions can be directly reexpressed in another transformation.

Egozcue et al. (2003) also presented the relationship between the alr transformation and ilr and clr transformation respectively, but as the alr transformation is not used widely anymore, the details are not further elaborated here.

2.4 Handling of Zeros and Missing Values

Since the log-ratio methodology is used in the context of compositional data, the parts of these compositions are required to be non-zero to get feasible results. Because it is possible for zeros to occur in the compositions, procedures are necessary to deal with them. Before a certain method is used, it has to be considered what kind of zeros or missing values are present. The literature distinguishes three types of zeros (Martín-Fernández et al., 2011):

Rounded Zeros are values which lie below a certain detection limit and thus are rounded to zero. They often appear in continuous variables and their true value is not zero and unknown, but the information about a possible maximum value - the detection limit - is known.

Count Zeros are values that occur due to the closure of a vector of counts. They appear in categorical variables and can occur in the case that the vector of counts may not be scale invariant or due to the limited size of a sample.

Structural Zeros are values which were actually recorded as zero and were not influenced by the experimental settings, e.g. a detection limit or imprecise measurements.

To analyse compositional data containing zero values, different methodologies to impute zeros are available. Imputation in this case means to replace the zero value by a small quantity. Most of the time, the classification of methods dealing with missing values can be applied to the techniques dealing with zero values, because the concepts are strongly connected to another.

Missing data was classified into three types by Little and Rubin (2002): Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR). MCAR is the

case were the missing values are independent of all other parameters and variables, MAR are missing values depending on observed variables and independent on unobserved information and NMAR are the missing values depending on the observed and unobserved information. If the three types of zeros are then classified in the sense of missing data, they generally are a case of NMAR, however sometimes rounded and count zeros are assumed to be a case of MAR for some methods.

Similarly to the principle of missing values where the treatment is different depending on the type, the different natures of zeros have to be considered to decide on their treatment.

2.4.1 Non-parametric Imputation

For rounded zeros the techniques most often applied are parameteric and non-parametric imputation. Using non-parametric imputation the values causing problems are taken and replaced by a non-problematic value. Typical examples include replacing the value by a fraction of the detection limit itself (Palarea-Albaladejo and Martín-Fernández, 2008) or an estimated mean, e.g. the arithmetic or geometric mean, of the observed values (Martín-Fernández et al., 2003a). As a non-parametric measure Aitchison (1986, p. 269) proposed the additive replacement. This method had the issue of distorting the covariance structure, due to the fact that the additive replacement is not subcompositionally coherent which was shown by Fry et al. (2000). This was resolved by the method given by Martín-Fernández et al. (2003a), the multiplicative replacement.

The multiplicative method from Martín-Fernández et al. (2003a) replaces every rounded zero by a small value δ_{ij}, large enough for computation, and afterwards modifies the other non-zero values of the data set accordingly to keep the closure. The modification is based on the sum-constraint of compositional data and thus the data is changed in a multiplicative way, but still fulfils the constraint laid upon the composition. Considering the composition x_i and the sum-constraint c_i, the j-th row of the replaced vector r_i is computed by

$$r_{ij} = \begin{cases} \delta_{ij} & \text{if } x_{ij} = 0 \\ x_{ij}\left(1 - \frac{\sum_{k|x_{ik}=0}\delta_{ik}}{c_i}\right) & \text{if } x_{ij} > 0 \end{cases}. \tag{2.19}$$

After imputation the multivariate analysis can be performed and the results can be interpreted with regard to some sensitivity to the replacement of the zero values. However, this non-parametric replacement strategy might not be sufficient as it could introduce artificial correlation, because it imputes exactly the same value for all compositions (Hijazi, 2011, p. 2) and thus if the proportion of zeros is high, especially if it is above 10% of the data, it can lead to an underestimation of the compositional variability (Filzmoser et al., 2018, p. 255). It is recommended to use parametric approaches with multivariate imputation in the case of high proportions of zeros.

Further imputation strategies include the random imputation (Boogaart and Tolosana-Delgado, 2013, pp. 216–218) which uses the conditional distribution of the zero values. It computes a random realization of the distribution and replaces the zero value by it. Parametric approaches using multivariate imputation can be used as well, more details follow in section 2.4.4 and section 2.4.5.

2.4.2 Projection

Structural zeros cannot be replaced, as it matters that they were actually recorded as zero. In general, there is no standard methodology defined to be used in this context. However, a technique considering only the subcompositions might be a solution depending on the underlying problem.

One possibility is to use the projection approach developed by Boogaart et al. (2006). This approach uses the fact that no information about any ratio with a part containing zeros is known. Thus only information of a subcomposition is known and is used to project the clr-transformed subcomposition into the null space of the vectors containing zeros. In other words the information of a composition with zeros is represented by the projected values and the projection mapping.

Therefore, Boogaart et al. (2006) proposed a way to compute the projected values $P_M \mathrm{clr}(x)$. The hyperplane of the clr-coordinates is spanned by non-orthogonal and non-basis vectors $w_i = e_i - \frac{1}{D}1$ for $i = 1, ...D$ and then the subcomposition of x_i of the non-zero values, denoted as x_{M^C}, is represented by the projection P_M of the clr-transformed vector $\mathrm{clr}(x)$ onto the null space $\{w_i | i \in M\}$ and M is the set of indices of the zero values. Boogaart et al. (2006) showed that the projection can be calculated by

$$(P_M \mathrm{clr}(x))_i = \begin{cases} \mathrm{clr}(x_{M^C})_i & \text{, if } i \in M^C \\ 0 & \text{, if } i \in M \end{cases} \tag{2.20}$$

where M^C is the complement of M and thus the set of indices of the non-zero values and x_{M^C} is the subcomposition of the composition x. The projected clr vector can then be used for further analyses.

2.4.3 Indicator of Subcomponents

An alternative approach to handle zero values, particularly structural zeros, in continuous or discrete data is to use an indicator for every occurring zero pattern and use the associated subcomponents to each pattern (Martín-Fernández et al., 2011, p. 53–54). In this case observations containing a zero at the same component are grouped together and compared to compositions with a positive value at the component. If M is the set of indices of zero values in the composition x and $\mathcal{M}$ is the set of all possible zero patterns of M, then M^c denotes the complementary set of each zero pattern $M \in \mathcal{M}$. Then d_M is a indicator variable to show, if the zero pattern M is present in composition x

$$d_M(x) = \begin{cases} 1 & \text{if the structural zero pattern of x is equal to } M, \\ 0 & \text{otherwise} \end{cases}$$

and b_M is the effect for each zero pattern in the compositional predictor η^{comp} of the form

$$\eta^{\mathrm{comp}} = \sum_{M \in \mathcal{M}} d_M(x) \langle b_{M^c}, x_{M^c} \rangle_A = \sum_{M \in \mathcal{M}} d_M(x) \langle \mathrm{ilr}(b_{M^c}), \mathrm{ilr}(x_{M^c}) \rangle. \tag{2.21}$$

To compute the predictor, it is required to fit for each subcomposition of non-zero components $\mathrm{ilr}(x_{M^c})$ the compositional coefficient separately for each zero pattern. In the case of only one

non-zero component in the composition the Aitchison inner product is zero and no contribution to the predictor is given (Verbelen et al., 2018, p. 1291). This setting is thus only successful, if simple zero structures are present which is rarely the case. Since zero structures are usually more complex and lead to a larger set of subgroups, due to the D-part composition consisting of over $2D-1$ possible zero patterns, the data set would be split into smaller data sets for each group and consequently lead to an insufficient sample size as the result (Filzmoser et al., 2018, p. 266).

2.4.4 Parametric Imputation with EM Algorithms

As an alternative to simple imputation a series of Expectation-Maximisation(EM)-based algorithms were developed to overcome the issue of zero values in compositional data. These methods take the covariance structure into account and consequently lead to different imputed values across the rows. The design of these methods in the context of compositional data is to keep the imputation below a certain detection limit. In general, the EM-based methods were designed for the treatment of rounded zeros.

Hron et al. (2010) first approached the implementation of methods not designed for detection limit problems. The designed algorithms are using a k-nearest neighbour imputation and an EM-based regression imputation which uses pivot coordinates to handle the zero or missing value problem within compositional data. These approaches were taking the multivariate data information into account for the imputation, whereas the earlier non-parametric approaches, for example by Martín-Fernández et al. (2003a) did not consider the multivariate structure of the data.

Following these ideas, algorithms considering the detection limits in compositional data were developed. Recently Martín-Fernández et al. (2012) introduced an new approach by combining the idea of the modified EM algorithm using alr coordinates by Palarea-Albaladejo et al. (2007) and Palarea-Albaladejo and Martín-Fernández (2008) and the ilr-based technique for compositional data by Hron et al. (2010). The resulting approach is estimating the rounded zeros using an iterative regression and ensures that the estimated values lie below the detection limit using censored regression, following the approach by Palarea-Albaladejo and Martín-Fernández (2008). For high-dimensional data Templ et al. (2016) implemented an EM-based algorithm to impute rounded zeros in all variables sequentially using partial least squares regressions to deal with the high-dimensional covariates.

The main idea throughout all zero treatment methods using EM-based algorithms is to use an multiple regression analysis for compositional data on the covariates to retrieve data without zero values. Therefore, a censored regression (Palarea-Albaladejo and Martín-Fernández, 2008) is used as the main focus lies on rounded zeros. The algorithm is also considering the threshold in log-ratio coordinates which are corresponding to the log-ratio coordinates applied to the composition. The transformation and inverse transformation needs to be applied in every step leading to a complex algorithm.

The detection limits are in general not the same for each part of the composition and can even differ among the observation, e.g. if the measurements were taken in different laboratories. It can be expressed by $d_{ik}, i = 1, ..., n, k = 1, ..., D$ with $d_{ik} \equiv d_{i1}^{(k)}$ representing the detection limit for the k-th part of the composition $X = (x_1, ..., x_D)$. The ilr-coordinates $Z^{(k)}$ of the zeros, which occur when $x_{i1}^{(k)} < d_{i1}^{(k)}$, are unknown with the restriction $z_{i1}^{(k)} < \psi_{i1}^{(k)}$ where $\psi_{i1}^{(k)}$ is the ilr-transformed

threshold

$$\psi_{i1}^{(k)} = \sqrt{\frac{D-1}{D}} \ln \left(\frac{d_{i1}^{(k)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_{ij}^{(k)}}} \right). \tag{2.22}$$

The steps of the algorithm of Martín-Fernández et al. (2012) are then as follows:

1. Initialisation of rounded zeros: Imputation by 65% of the detection limit or any other univariate method (see section 2.4.1).
2. Sorting parts of the composition decreasingly by the number of zero values.
3. For each part of the composition $k = 1, ..., D$ repeat steps 4 to 8:
4. Express the compositional part $x^{(k)}$ in ilr-coordinates using Eq. 2.13.
5. Express the threshold $d^{(k)}$ in ilr-coordinates using Eq. 2.22.
6. Denote $\mathcal{M}_k \subset \{1, ..., n\}$ as the set of indices of the zero values in part $x^{(k)}$ of the composition and $\mathcal{M}_k^C$ as the complementary set of non-zero indices. Then $z_1^{\mathcal{M}_k}$ and $z_1^{\mathcal{M}_k^C}$ denote the first coordinate and matrices $Z_{-1}^{\mathcal{M}_k}$ and $Z_{-1}^{\mathcal{M}_k^C}$ contain the remaining columns without the first coordinate and add as the first column a vector consisting of only ones. Based on these specifications the following regression model with an intercept is estimated

$$z_1^{\mathcal{M}_k^C} = Z_{-1}^{\mathcal{M}_k^C} b + e, \tag{2.23}$$

   and the regression coefficients b and the error term e are unknown.
7. Estimate the coefficients of Eq. 2.23, denoted as $\hat{b}^{(k)}$. Then denote the i-th row of $Z_{-1}^{\mathcal{M}_k}$ as $z_{i,-1}^{\mathcal{M}_k}$ and replace each unknown value by the conditional expected value

$$\hat{z}_{i1}^{\mathcal{M}_k} = (z_{i,-1}^{\mathcal{M}_k})' \hat{b}^{(k)} - \hat{\sigma}^{(k)} \frac{\phi\left(s_i^{(k)}\right)}{\Phi\left(s_i^{(k)}\right)}, \quad s_i^{(k)} = \frac{\psi_{i1}^{(k)} - (z_{i,-1}^{\mathcal{M}_k})' \hat{b}^{(k)}}{\hat{\sigma}^{(k)}}, \qquad \forall i \in \mathcal{M}_k, \tag{2.24}$$

   where ϕ and Φ are the density and distribution function of the standard normal distribution and $\hat{\sigma}^{(k)}$ is the estimated conditional standard deviation of $z_1^{\mathcal{M}_k}$.
8. Express the updated values in the original space using the inverse ilr-transformation (Eq. 2.14). The non-zero values are changed as well while keeping the ratios between them.
9. If the Frobenius norm of the difference between the covariance matrices of the present and previous iteration are below a certain boundary, stop the iteration which started in step 3.
10. Rearrange parts in the original order.

In the case of high-dimensional data this algorithm cannot be used, due to the regression in Eq. 2.24 where the classical or robust regressions cannot deal with high-dimensional data. Therefore, a partial least squares regression is required. The algorithm of Templ et al. (2016) is then based on bootstrapping to replace the zero values. In this case n samples are bootstrapped and split into pairs, to each pair a PLS regression is applied and using a tenfold cross-validation the predicted error

sum of squares (PRESS) is computed. Based on the PRESS a PLS model with a smaller amount of components is selected. For a detailed description check Filzmoser et al. (2018, p. 260–262) or Templ et al. (2016).

Additionally, to the above EM-Algorithms to impute rounded zeros, Hijazi (2011) proposes an algorithm which can be applied, if the compositional data arise from a Dirichlet distribution. In the algorithm of Palarea-Albaladejo et al. (2007) and Palarea-Albaladejo and Martín-Fernández (2008) this case of compositional data was not covered, as their algorithm assume compositional data of a normal distribution. However, in this thesis only the case of data arising from a Normal distribution will be covered.

2.4.5 Further Methods and Missing Data

In the last years new imputation approaches using deep learning methods were introduced and developed in the context of rounded and count zeros which are available in the software R. These imputation methods include, besides the mentioned EM algorithms, algorithms based on k-nearest neighbours methods, Kaplan-Meier smoothing splines (Lubbe et al., 2021, p. 2, Templ, 2021, p. 166) and Bayesian posterior estimates (Martín-Fernández et al., 2015) to impute rounded zeros and are included into the R packages `robCompositions` (Templ et al., 2011, version 2.3.0) and `zCompositions` (Palarea-Albaladejo and Martín-Fernández, 2015, version 1.3.4). Templ (2021, p. 166) gives a detailed overview.

As discussed in the previous section, EM algorithm were developed to treat rounded zeros. However, the most recent idea in the context to treat rounded zeros is based on artificial neural networks and was introduced by Templ (2021). To impute the zeros one artificial network is used per variable to be imputed and an EM-based algorithm is applied.

In the case of count zeros in compositional data Martín-Fernández et al. (2015) developed a bayesian-multiplicative method which was based on the general idea of the multiplicative replacement by Martín-Fernández et al. (2011) and replacing a zero value by an expected value using the posterior Bayesian estimate. This approach unfortunately does not fully account for the scale invariance. As a consequence, model-based replacement algorithms, as described in section 2.4.4 can be utilized in the context of count data as well, considering 1 as the detection limit (Martín-Fernández et al., 2015, p. 154–155).

Missing data in the context of compositional data can not be treated as in the classical multivariate case, because of the data differences and the theoretical complexity. The k-nearest neighbour approach and the iterative model-based imputation by Hron et al. (2010) were developed to tend to this problem of missing data within compositional data sets. For missing data the same approaches as for zeros apply, as their concepts are very similar in the context of compositional data. Therefore, to treat missing values one can also use the introduced imputation methods of section 2.4.1 or the projection approach in section 2.4.2 as well as the approaches mentioned in this section.

2.5 Descriptive Statistics for Compositional Data

Since the data is compositional and as mentioned standard methods cannot be applied, a compositional alternative has to be considered for functions of descriptive statistics applied to the data. Thus the standard descriptive functions including arithmetic mean, variance and covariance should not be used. New methods in the framework of the Aitchison geometry are used.

2.5.1 Center

The arithmetic mean of the Euclidean geometry is replaced by the sample center (Aitchison, 1997) with respect to the Aitchison geometry. It can be represented by the component-wise geometric mean. Like the name suggests it characterises the center of the sample data.

Definition: A value of central tendency of a compositional sample is the closed geometric mean. It is called **center**. For a $n \times D$ compositional data matrix X, it is defined as

$$\operatorname{cen}(X) = \hat{g} = \mathcal{C}\left(\hat{g}_1, \hat{g}_2, ..., \hat{g}_D\right), \quad \hat{g}_j = \left(\prod_{i=1}^{n} x_{ij}\right)^{1/n}, \; j = 1, 2, ..., D. \tag{2.25}$$

The center does satisfy the principles of compositional data and respects the Aitchison geometry. Additionally, it can be shown that the center of centered data is the neutral element n and consequently the effect of a relative scale can be suppressed. This characteristic is often used for graphical purposes like tenary diagrams (Filzmoser et al., 2018, p. 70).

2.5.2 Variation Matrix and Total Variance

As a compositional alternative to the variance and standard deviation, the variation matrix and total variance defined by Aitchison (1986) can be used.

Definition: Considering a D-part composition x. Dispersion in a compositional data set can be described either by the **variation matrix**, defined as

$$T = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1D} \\ t_{21} & t_{22} & \cdots & t_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t_{D1} & t_{D2} & \cdots & t_{DD} \end{pmatrix}, \quad t_{ij} = \operatorname{var}\left(\ln \frac{x_i}{x_j}\right), \tag{2.26}$$

or by the **normalized variation matrix** defined as

$$T^* = \begin{pmatrix} t^*_{11} & t^*_{12} & \cdots & t^*_{1D} \\ t^*_{21} & t^*_{22} & \cdots & t^*_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D1} & t^*_{D2} & \cdots & t^*_{DD} \end{pmatrix}, \quad t^*_{ij} = \operatorname{var}\left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\right). \tag{2.27}$$

The so-called variation matrix is constructed by the variances of the pairwise log-ratios. This means, the usual variance is applied to the log-ratio of the parts i and j and then the matrix of all variances of pairwise log-ratios is build. If each pairwise log-ratios is multiplied by the factor $\frac{1}{\sqrt{2}}$, they are called balances and the normalized variation matrix is constructed by taking the variances over all balances (Pawlowsky-Glahn et al., 2015, pp. 66–67). Both matrices are symmetric and the diagonal elements contain only zeros by definition. The variances can not be directly calculated in praxis, but can be estimated via unbiased maximum likelihood estimators.

Definition: Consider a $n \times D$ compositional data matrix X. A measure of global dispersion of a compositional data set is the **total variance**, defined by

$$\mathrm{totvar}(X) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\mathrm{var}\left(\ln\frac{x_i}{x_j}\right) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}t_{ij} = \frac{1}{n}\sum_{k=1}^{n}d_A^2(x_k,\hat{g}). \tag{2.28}$$

As in Eq. 2.28, the total variance as it was defined first by Aitchison (1997) summarizes the variation matrix into one quantity. It does make sense to calculate a total variance, since all parts of the composition share common scale (Pawlowsky-Glahn et al., 2015, p. 68). Furthermore, it can be shown with the last expression, that the total variance is equivalent to the metric variance defined by Pawlowsky-Glahn and Egozcue (2001).

2.5.3 Centering and Scaling of Data

Centering and scaling of compositional data has a particular role in applications for visualization purposes. Using centering makes it possible to improve the data representation in an easy way. Analogously to moving real data in the real space towards the origin, the compositional data is moved to the barycenter of the underlying simplex which is represented by the neutral element n. The general strategy introduced by Pawlowsky-Glahn et al. (2015, pp. 68–69) to center the data is row-wise perturbing of the compositional data matrix X by the inverse center $\mathrm{cen}(X)^{-1}$.

As an compositional alternative to scaling using the classical approach in the real space, i.e. dividing a centered variable by the standard deviation, the centered composition data can be scaled with the factor $\mathrm{totvar}(X)^{-1/2}$ which preserves the relative contribution of each log-ratio in the variation. The important difference in the compositional scaling to standardization in the real space is that it can not be applied part-by-part, but to the data set as a whole (Pawlowsky-Glahn et al., 2015, p. 69). For a standardized composition the center is the neutral element and the total variance is 1 (Pawlowsky-Glahn et al., 2015, p. 112).

2.6 Normal Distribution on the Simplex

To take the next step to statistical inference of the data a probability model has to be constructed. Therefore, a probability distribution on the simplex is needed. There are two distributions mainly used in the context of compositional data, the additive logistic-normal distribution introduced by Aitchison and Shen (1980) which is called the normal on the simplex, due the definition introduced by Mateu-Figueras et al. (2013). The other mainly used distribution is the Dirichlet distribution

(Narayanan, 1991).

Definition: Given a random composition X with the sample space S^D. Then it is said to have a **normal on the simplex**, with parameters μ and Σ and random orthonormal coordinates of X, $x^* = h(x)$, if its density is

$$f_X^S(x^*) = \frac{1}{(2\pi)^{(D-1)/2}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^*-\mu)'\Sigma^{-1}(x^*-\mu)\right). \tag{2.29}$$

If X follows a normal on the simplex the notation $X \sim \mathcal{N}_S^D(\mu, \Sigma, \alpha)$ is used, indicating a model on the simplex with the subscript S and a D-part composition with the superscript D (Mateu-Figueras et al., 2013, pp. 11–12). The function $h(\cdot)$ to generate orthonormal coordinates of X, can be either one of the introduced coordinate representations, i.e. alr, clr or ilr. Through this definition it is shown if X follows a normal distribution on the simplex, then X^* has to be multivariate normal distributed on $\mathbb{R}^{D-1}$ (Pawlowsky-Glahn et al., 2015, p. 114).

Mateu-Figueras et al. (2013) showed that the normal on the simplex is closed under perturbation and powering and is invariant under perturbation. This is providing some good tools for handling compositional data. It was further shown that the family of normal on the simplex is closed under permutation and subcompositions (Mateu-Figueras et al., 2013, p. 12).

Pawlowsky-Glahn et al. (2015) defined equivalently to the Lebesgue measure in the $\mathbb{R}^D$ space an Aitchison measure on the S^D. The Aitchison measure is constructed from the Lebesgue measure on the space of coordinates. In this case the ilr coordinates are used. The probability density function (pdf) (Eq. 2.29) in the definition of the normal on the simplex of Mateu-Figueras et al. (2013) is defined with respect to the Aitchison measure on the S^D.

Based on the definition of the pdf of the normal on the simplex and the Aitchison measure it is shown in Pawlowsky-Glahn et al. (2015, pp.115–119) that many characteristics and theorems based on the normal distribution in the real space can be transferred onto the normal on the simplex in S^D. In particular, the central limit theorem can be defined in the simplex analogously to the central limit theorem in real space.

The second distribution is the Dirichlet distribution which can be obtained as a closure over independent, equally scaled, gamma-distributed and positive random variables (Aitchison, 1986, p. 58 recited in Pawlowsky-Glahn et al., 2015, p. 121). This distribution was mainly used before the log-ratio approach was formally introduced (Filzmoser et al., 2018, p. 85).

Definition: Given a random composition X with the sample space S^D and the positive parameters $\alpha = (\alpha_1, \alpha_2, ..., \alpha_D)'$.Then X is said to follow a **Dirichlet distribution**, if its density is

$$f(x; \alpha_1, \alpha_2, ..., \alpha_D) = \frac{\Gamma(\alpha_1 + ... + \alpha_D)}{\Gamma(\alpha_1) \cdot ... \cdot \Gamma(\alpha_D)}\, x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} ... x_D^{\alpha_D - 1}, \tag{2.30}$$

where $\Gamma(\cdot)$ denotes the Euler Gamma function.

The Dirichlet distribution has many advantageous characteristics and is widely applied in the Bayesian statistics, but the approach is not scale invariant. This problem can also not be overcome by redefining the distribution with respect to the Aitchison measure like in the case of the normal distribution

(Filzmoser et al., 2018, pp. 85–86; Pawlowsky-Glahn et al., 2015, pp. 121–124). So that the focus will be on the normal on the simplex in the following chapters.

2.7 Regression Models

As mentioned before, it was possible to define the compositional version of a normal distribution. In the same manner, it is possible to use the standard regression methods in the compositional context, if the compositional data are mapped into the Euclidean space.

As linear models relate response variables to covariates there are multiple cases to consider for compositional data. The cases can be divided into the three main groups: linear models with compositions as response, linear models with compositions as covariates and linear models with compositions as responses and covariates. In the latter part of this thesis a regression using compositional data as covariates is performed, so that only this case is discussed in detail in the following section.

2.7.1 Linear Model with Compositional Covariates

The aim of the regression is to relate a D-part composition $x_i = (x_1, x_2, ...x_D)'$ of the $n \times D$ matrix X to the real response $y = (y_1, y_2, ..., y_n) \in \mathbb{R}^D$. This leads to the follwing expression of the model

$$Y = \beta_0 + \langle b, X \rangle_A, \quad b \in S^D,\ \beta_0 \in \mathbb{R}^D. \tag{2.31}$$

Pawlowsky-Glahn et al. (2015) suggests that the linear regression is fit to the response y using a linear function of $\mathrm{ilr}(x)$, so that the classical least squares criterion (SSE) is applied to the target function,

$$SSE = \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \mathrm{ilr}(b), \mathrm{ilr}_i(x) \rangle\right)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \langle \beta, z_i \rangle\right)^2. \tag{2.32}$$

The regression then yields the coefficients $\beta = \mathrm{ilr}(b)$ of y with respect to the transformed coordinates $z_i = \mathrm{ilr}_i(x)$ and these can be computed into the actual b values using the function ilr^{-1} as it was defined in Eq. 2.14. Therefore, the regression can be done without further changes after the data was transformed into the ilr coordinates and no more methods are needed. The clr coordinates could also be a possible choice of transformation, but due to numerical reasons Pawlowsky-Glahn et al. (2015, p. 158) recommends to avoid them.

On the regression coefficients the standard tests can be used as usual and for the ilr coordinate coefficients it has to be kept in mind that they are related to a basis and if the model is being reduced this basis can also change (Pawlowsky-Glahn et al., 2015, p. 158).

2.7.2 Interpretation of predictors

Using compositions only the the ratios between the parts and thus their relative information is considered as covariate, therefore the information given by the estimates are also based on the relative information. So that the main issues of using compositional data within a regression is the diffi-

culty of interpretation. To include compositional data in a classic regression within the Euclidean space they need to be transferred into the Euclidean space using transformations, e.g. by isometric log-ratio (ilr) transformation and then they can be included in the regression model as explanatory variables. Therefore, a compositional predictor term of the composition $x \in S^D$ is included into the model for $\beta \in \mathbb{R}^{D-1}$ and $z = \text{ilr}(x) \in \mathbb{R}^{D-1}$ in the following way

$$\eta = \beta_1 z_1 + \beta_2 z_2 + ... + \beta_{D-1} z_{D-1} = \sum_{i=1}^{D-1} \beta_i z_i. \tag{2.33}$$

After the model fit, the estimated non-compositional coefficients can be interpreted as usual. However, for compositional terms only the first coefficient corresponding to x_1, the first part of the composition, has a comprehensible interpretation. Since z_1, the ilr-transformed first part of the composition, explains the relevant information for x_1, but every other transformed model parameter includes information of multiple parts of the composition and thus can not be interpreted in a straightforward fashion. A general approach for interpretation could be to refit the model for every part of the composition. If a composition has more than two parts, this could already be computational intensive. Hence, some method for interpretation without having to refit the model is needed.

Verbelen et al. (2018) have developed a new idea for this problem. This approach fits the model once and then uses transformations and inverse transformations to get interpretable values for the model parameters. First, the inverse ilr-transformation is applied to the model coefficients β as it was proposed by Boogaart and Tolosana-Delgado (2013) and Pawlowsky-Glahn et al. (2015), i.e. setting $b = \text{ilr}^{-1}(\beta) \in S^D$ and rewrite the compositional predictor into the following expression

$$\eta = \sum_{i=1}^{D-1} \beta_i z_i = \sum_{i=1}^{D-1} \text{ilr}(b)_i \text{ilr}(x)_i = \langle b, x \rangle_A. \tag{2.34}$$

For the new interpretation approach Verbelen et al. (2018, pp. 1289–1290) is using the facts that the composition $b \in S^D$ can be interpreted as simplicial gradient with respect to x (Barceló-Vidal et al., 2011) and it is the compositional direction along which the predictor is increasing the fastest. So that with a fitting perturbation a new interpretation for the effects of the composition parts is possible. They proposed a perturbation of the composition in the direction of each component. For this the composition x was perturbed by the closure $C(\alpha, 1, ..., 1)'$, i.e. $\tilde{x} = x \oplus C(\alpha, 1..., 1)' = C(\alpha x_1, x_2, ..., x_D)'$, using $\alpha > 0, \alpha \in \mathbb{R}$ as first element and holding the remaining parts constant, to get the relative ratio change of α in the first part. In this sense $\alpha < 1$ would signify a decrease and $\alpha > 1$ an increase in the relative ratio. The same can be achieved for the relative ratio change in the other parts, if α is placed at another position. The change of the predictor is then given by

$$\begin{aligned} \langle b, C(\alpha, 1, ..., 1) \rangle_A &= \langle b, (\alpha(D-1+\alpha)^{-1}, (D-1+\alpha)^{-1}, ..., (D-1+\alpha)^{-1}) \rangle_A \\ &= \frac{1}{D} \ln(\alpha) \sum_{j=1}^{D} \ln\left(\frac{b_1}{b_j}\right) = \ln(\alpha) \left(\ln(b_1) - \frac{1}{D} \sum_{j=1}^{D} \ln(b_j) \right) = \ln(\alpha) \text{clr}(b)_1 \end{aligned} \tag{2.35}$$

where the relationship derived in equation 2.18 is used in the last step of the equation, as it can

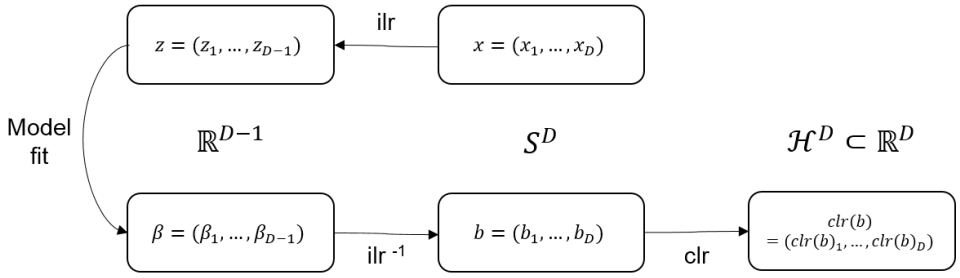

Figure 2.1: Process to Calculate Interpretable Regression Parameters via representation in the ilr-, simplex and clr-space based on Verbelen et al. (2018, p. 1301)

be shown that $UU' = I_D - \frac{1}{D}1_D 1'_D$. Using this relationship Verbelen et al. (2018) showed that the clr-transformation of b is the best way to understand the effect of a ratio change.

Using a linear model the effect of the ratio change in part $i = 1, ..., D$ of the composition on the response would be interpreted as: **When the ratio in part i changes by α the response y changes by $\ln(\alpha)\mathrm{clr}(b)_i$.**

This leads to the interpretation, if the clr transformation is applied, that a positive and negative effect of each part is indicated by the sign of the clr-transformed coefficients. Moreover, a direct interpretation of the compositional term is given. For example, if the ratio in the first term doubles, i.e. $\alpha = 2$, and if the estimated coefficients in the simplex are $b = (0.25, 0.35, 0.4)$. Then the clr-transformed estimate in the first part is $\mathrm{clr}(b)_1 = -0.2688$ and the impact on the response is a decrease of $\ln(2)\mathrm{clr}(b)_1 = -0.186$. In comparison, if the ratio in the third part doubles, then the clr-transformed estimate is $\mathrm{clr}(b)_3 = 0.2012$ and thus the impact on the response increases by $\ln(2)\mathrm{clr}(b)_3 = 0.139$. If the ratio α decreases then the impact on the response reverses and a negative estimate leads to a positive impact and vice versa.

2.7.3 Variable Selection

Variable selection is important to find the subset of covariates which best represent the response. The aim is to construct a simple model to explain the response, in other words the smallest model that fits the data and at the same time does not add any noise in the model, because unnecessary covariates were considered. Other effects would be collinearity between covariates and the cost for prediction. Thus a good model selection method has to be employed (Filzmoser et al., 2018, pp. 192–193).

A common approach for model selection is the stepwise variable selection using the Akaike Information Criterion (AIC). In this approach either the Backward Stepwise or Forward Stepwise procedures are used. The Backward Stepwise algorithm starts with a model considering many covariates, e.g. a full model, and then removes a covariate in each step until the AIC is not decreasing any more or the smallest model, e.g. the model only considering a constant, is reached. The Forward Stepwise algorithm starts with the smallest possible model, e.g. model with only a constant, and adds a covariate in each step as long as the AIC is decreasing. In each step the variable leading to the lowest AIC value is added. This is repeated until the AIC is no longer decreasing or the full model is reached. The AIC value itself will be explained in the subsequent chapter.

Due to the fact that the AIC is rotation-invariant, any selection of orthonormal coordinates for the compositions will lead to the same results and therefore the AIC can be applied in stepwise procedures handling compositions. But in the context of compositional data, the covariates being removed in the backward stepwise procedures can contain the coordinates of the composition. By removing one of the coordinates the left coordinates will form a subcomposition to represent the composition (Filzmoser et al., 2018, p. 193). In the case of forward stepwise procedures for D-part compositions a lot of computational effort is needed, especially if D is large. Thus Filzmoser et al. (2018) did not recommend the Forward Stepwise procedure for the cases including compositional covariates.

2.7.4 Robustness

In section 2.7.1 the common least squares criterion - SSE - was minimized to find the optimal estimator for the covariates. The results of this minimization can be heavily influenced by outliers, because they can yield large residuals and thus dominate the value of the SSE.

Two types of outliers are of importance, the outliers in the response and the outliers in the explanatory variables. To achieve a robust regression the procedure needs to handle both types in order to achieve the best estimators. The first approach replaced the least squares criterion by the M-estimator which is expressed by

$$\hat{b}_M = \underset{b}{\operatorname{argmin}} \sum_{I=1}^{n} \rho \left(\frac{r_i(b)}{\hat{\sigma}(b)} \right), \tag{2.36}$$

where $\rho(\cdot)$ is a function which can be squared around zero, but is bounded for large values, $\hat{\sigma}(b)$ is the estimated residual scale dependent on the unknown regression coefficient. This leads to the challenge that the estimation of the residual scale is dependent on the regression coefficients and vice versa. A resolution to this regression task leads to the MM-estimator which is using an iterative algorithm and some tuning parameters and provides robust estimation and is efficient (Filzmoser et al., 2018, pp. 194–195). Further approaches with least trimmed sum-of-squares and robust bootstrap where developed as well and implemented into the R package `robustbase`.

2.8 Application in R

To handle compositional data in applications, different methods were developed. One of those methods within `R` is the package `compositions` developed by Boogaart et al. (2020). This package with the addition of packages including standard multivariate analysis tools are making it possible to apply the theory presented in this section to the actual data sets. Another very useful package to handle compositional data with robust methods is the package `robCompositions` developed by Templ et al. (2011).

Before starting the analysis in the later sections both packages and their dependent packages, e.g. `tensorA`, `colorspace` and `robustbase`, are installed on a local machine. Throughout all applications presented, the methods included in these packages are used.

3 Project STAR

The data set of this project is an open source collected by Achilles et al. (2008). In this section the focus will be on the setting of data collection as well as the descriptives of the data set.

3.1 Data Setting and Experiment Execution

The Tennessee Student Teacher Achievement Ratio (STAR) project was a large scale randomized experiment conducted in the time span of 1985-1989 in the state Tennessee. It covered the first four years of school, from kindergarten until the third grade, of students from the participating schools. A total of 79 public schools within Tennessee took part in the project which lead to a total of 11,600 students. The schools of rural, urban, suburban and inner city area were all included due to a specification of the legislative, so that the schools were not clustered, but spread across the state. The goal of this study was to investigate the effect of the class size on the students achievements.

For this purpose all students entering a participating school in 1985 were randomly assigned to a classroom of one of the following three types: a small class consisting of 13-17 student, a regular class having 22-25 students or a regular-sized class with an additional teacher aide. The teachers were as well randomly assigned to their classrooms. In the case that a student entered one of the cohorts of the participating schools in one of the later years, they were added to the experiment and as well randomly assigned to a class type. All randomizations were done within each school. Other than the class size no major changes, e.g. in process and organization were undertaken.

This includes the fact that there were no teacher trainings offered for the experiment excluding 15 schools which were offered a three-day training before the third and fourth year of the experiment. The training was offered independently from the class type. However, in the STAR technical report of Word et al. (1990, pp. 116–127) it is shown that the teacher training had no significant influence on the performance of a teacher compared to teachers without the training.

Whether the randomization was really maintained in the execution of the experiment was discussed in multiple studies (Whitmore Schanzenbach, 2006/2007 and Sojourner, 2013). There it was extensively checked, if the class assignment was not done in a compensatory manner and that the class assignment of new participants was random. No systematic differences could be found and evidence supporting the randomization could be provided.

So that a school could participate in the project some constraints were set which limited the sample. One important figure which was required from each school was the minimum cohort size. The participating schools needed a cohort of at least 57 students, to have all three classroom types with class size within the defined range available. As a result, the schools in the project were in general larger than non-participating schools and because of geological consideration in the selection of project school, also the demographics of the students were influenced and do differ from the state average. One specific difference to the state-wide schools was the higher proportion of inner-city schools participating in the project, as there were more schools satisfying the restraints of the project within the inner-city regions. This had a direct impact on the project as Word et al. (1990, pp. 7–8) showed that on average the projects schools scored lower than state average and than the comparison schools. The differences were mostly between the inner-city and suburban areas and

Table 3.1: Description of variables within the student data set of project STAR

Variable	*Description*
Student Demographics	
`stdntid`	Unique student id which is assigned to every student
`schid`	School id of the school which the student attended
`gender`	Students sex (*male, female*)
`race`	Students race or ethnicity (*white, black, asian, hispanic, native american, other*)
`birthyear, birthmonth, birthday`	Students date of birth
Grade Dependent Information	
`g<n>classtype`	The class type which a student was randomly assigned to in grade <n>$\in \{k, 1, 2, 3\}$ (*small class, regular class, regular + aide class*)
`g<n>freelunch`	Students free-lunch status as indicator for the socio-economic status in grade <n>$\in \{k, 1, 2, 3\}$ (*free lunch, non-free lunch*)
`g<n>tmathss`	Students SAT total math score of grade <n>$\in \{k, 1, 2, 3\}$
`g<n>treadss`	Students SAT total reading score of grade <n>$\in \{k, 1, 2, 3\}$
`g<n>tmathss`	Students CTBS total math score of grade <n>$\in \{4, 5, 6, 7, 8\}$
`g<n>treadss`	Students CTBS total reading score of grade <n>$\in \{4, 5, 6, 7, 8\}$

Note: This table represents only a fraction of the variables included in the data set. In brackets are the categories available for the categorical variables. The whole data set covers 379 different variables, including all test results from kindergarten to 8th grade as well as the non-cognitive skills documented in 4th grade and 8th grade and the selected courses in high-school, their GPA and graduation status. Furthermore, it was documented whether a student took the ACT or SAT in the year of graduation.

less among suburban, rural and urban area.

Furthermore, it should be taken into account that a relatively large fraction of students exited the experiment within the four years. This could happen due to school moves or grade skipping for example. This was one of the sources of deviation from the perfect setting possible which would have been that all students remain in the same class type throughout the whole experiment.

As the project started the students each got an individual identification number and information about their demographics, including their race or ethnicity, sex and age, as well as whether they received a free or reduced prize lunch and their test scores were collected. The attended class type of each students was recorded as well. In Table 3.1 an overview of the variables of interest is presented. The teacher demographics, education levels and experience and information about the schools, among others average daily attendance, school enrolment and percentage of students receiving free lunch, were also collected as these might have affected the experiment results, but won't be included in later analysis, due to the focus being the direct peer effects.

The recorded test scores within the four years of the execution of the project all belonged to the Stanford Achievement Test (SAT). These tests are norm-referenced and were administered to the appropriate level of the students for each grade. The subjects on which the students were tested in each grade included math, reading, spelling and listening and all their respective test scores were included into the data set.

From the fourth grade on the students were put back into regular-sized classes. However, due to follow up studies to measure middle and long-term impacts, the participating students were followed up until high school and additional information of the following years were recorded as well. So that

the data of the test scores between kindergarten and third grade, while the students were part of the project, are available as well as the test scores from fourth to eighth grade. When the students entered their fourth grade the Tennessee Comprehensive Assessment Program (TCAP) was started, so that starting then up to eighth grade the scores were determined using the Comprehensive Tests of Basic Skills (CTBS) instead of the SAT and later collected and added to the data set for the students participating in the project STAR.

Furthermore, while students were in fourth and eighth grade a random subset of the students were part of a test for non-cognitive skills based on the Student Participation Questionnaire (Finn et al., 1991). The high school courses taken, the graduation status and whether a student took the ACT or SAT college-admission test and the scores of these are collected to each participating students as well. The college-admission scores were linked to the STAR information in cooperation with the ATC Inc., the College Board and the Educational Testing Service (ETS) by Whitmore and Krueger (2001).

Thus a complex set of data of the project execution, the middle school test scores, two non-cognitive tests on a subset of students executed in fourth and eighth grade and high school scores and graduation as well as college admission was collected for each student and linked together by Achilles et al. (2008).

3.2 Sample Selection and Descriptive Statistics

The kindergarten sample includes a total of 6325 students of 79 schools and divided into 325 classrooms. I exclude 26 students with missing demographic characteristics. Then the data set is separated into two parts, one for the math scores and one for the reading scores as both scores are available throughout kindergarten to eighth grade and thus are suitable for regression analysis based on test scores. The first analysis is studying the impact of peer effects from the time the cohorts were formed in kindergarten on the students right after the project ended. Therefore, in each of those two data sets the students who are missing test scores in the respective subject in either kindergarten or grade four are dropped. Not every student which was recorded in kindergarten, while the project was going on, also has information of them in fourth grade documented, so that this condition is reducing both data sets to 3250 students for the math test score and 3217 students for the reading test score. Table 3.3 reports descriptive statistics for the students demographic characteristics and kindergarten and fourth grade scores.

Overall 6299 students were recorded in kindergarten without missing demographics. For those students slightly more students were male than female and almost half of the students received

Table 3.2: Percentages of participation in the project per grade over all students recorded

Grade	*Yes, participated*	*No, did not participate*
kindergarten	54.5	45.5
grade 1	58.9	41.1
grade 2	59.0	41.0
grade 3	58.6	41.4

Note: Percentages taken over all 11.601 available student data sets.

Table 3.3: Descriptive statistics over all relevant variables in the sample of Project STAR

	N	*Min*	*Mean*	*Max*	*SD*
Overvall Demographics Kindergarten	6299				
female		0	0.5140	1	0.4998
non free lunch		0	0.4844	1	0.4998
small class		0	0.3004	1	0.4585
race		1	-	6	-
Test Scores					
kindergarten math score	5852	288	485.4	626	47.7335
kindergarten reading score	5770	315	436.7	627	31.7341
4th-grade math score*	2385	492	715.5	840	42.3032
4th-grade reading score*	3011	517	624.3	775	37.5038
Math Data Set	2259				
female		0	0.512174	1	0.49996
non free lunch		0	0.698097	1	0.45919
small class		0	0.306773	1	0.46126
race		1	-	6	-
kindergarten math score		375	500.0	626	44.4350
4th-grade math score		492	715.5	840	42.6251
Reading Data Set	2822				
female		0	0.512403	1	0.49993
non free lunch		0	0.620482	1	0.48535
small class		0	0.310064	1	0.46260
race		1	-	6	-
kindergarten reading score		358	444.8	627	31.8621
4th-grade reading score		517	624.4	775	37.5309

Note: This table reports the descriptive statistics separately for the overall demographics of students entering in kindergarten, the students considered in the math score data set and the reading score data set. In the latter two, is was required that the scores of the student in kindergarten and fourth grade are available. Gender is measured by an indicator taking the value 1 if the student is male and 0 otherwise. Free lunch takes the value 1, if a student is receiving a free or reduced-price lunch and 0 otherwise, small class takes the value 1, if the student was assigned to a small-sized class and 0 otherwise and race is a factor variable with 6 levels and thus the mean and standard deviation are not reported as they do not carry any meaningful information. The marked test scores (*) are only considering scores of students who took part in the experiment while in kindergarten.

a free or reduced-price lunch. Around 30% of the students were assigned to a small-sized class. Over 95% of the students were either black or white leaving only a minority of students having other ethnicities or race.

For the participation in the experiment it can be shown that a larger fraction of students entered in grade 1. Table 3.2 shows the distribution of all participating students at some time in the study per grade. The results show that in first grade the rate of participating students were 5% higher than in the kindergarten and thus implying that a larger group must have entered at that time, which then stayed almost constant until third grade.

For the students which entered the experiment in kindergarten 5852 data sets were collected for the math test scores and 5770 for the reading test scores without any missing values for the student demographics. For these students around half of them also had their fourth grade CTBS score collected, further reducing the data set as mentioned above. A main challenge is that different test

scores were used in different grades, SAT in kindergarten and the CTBS in fourth grade, thus the scores cannot be compared directly.

In the data set for the math test score and reading test score the student demographics differ slightly. In both sets more females than males and up to 40% of students receiving a free lunch in kindergarten are present. The assignment to a small class in kindergarten still are 30% of the students which corresponds with the overall value. Furthermore, is the mean for math scores in the reduced set higher for the kindergarten and fourth grade scores as well as the mean for the reading scores.

More students scoring in the lower range seem to not have been recorded in the higher grade and thus increasing the mean. A possible reason could be that they repeated a year and thus the scores of the higher grade could not be collected, when the data sets were connected to the other data sets of the TCAP in the year they would have been, considering the students were on pace.

As the data of project STAR was collected by various research teams during and after the experiment, the following section is giving a brief overview of the main independent and dependent variables of the following empirical analysis.

Students Demographics. Within project STAR information of the student containing their gender, race, date of birth and status for free or reduced-price lunch per grade were collected. The variable gender was separated in male and female, the lunch status was either yes (they received a free lunch) or no and race had six variables: White, Black, Asian, Native American, Hispanic and others. Because the students generally entered the school at the same age, the variable 'age' does not convey a lot of information and is not included. Due to the fact that minorities were only represented by a small part of the data, their estimates will be sensitive to outliers and therefore are aggregated. Thus the variable 'race' is transformed into an indicator variable having the value 1, if a student is white and 0 otherwise.

Class Size. The classes were separated into the three types: small class, regular class and regular + aide class and the data recorded per student was indication which class type was visited in which grade of the experiment. For the empirical analysis this variable is only considered as impact of the class size. Hence, it is converted to an indicator variable whether a student visited a small class. This variable has the value 1, if the student was part of a small class and 0 otherwise.

Grade K to Grade 3 Test Scores. The test scores were recorded at the end of each grade with an grade-appropriate version of the SAT. The test scores recorded in kindergarten are part of the independent variables of the empirical analysis.

Outcomes. All attendees of public schools took the CTBS in the spring of the grades 4 to 8. The same as the SAT, the CTBS test is a standardized multiple-choice exam which is used to test students in the subjects mathematics, reading, language studies and sciences in each grade. The influence of the distribution of the student performance in elementary school on the later test scores are studied in the empirical analysis.

Class Score Distribution in Grade K. In the next step a data extension of converting the scores of the students into intervals and computing the percentage of students within a class, scoring in the

interval for each class, is performed. This new variable is adding relative information of the class score distribution of the class in which each student was part of to the existing data. Hence, the influence of the relative distribution on the outcome can be studied.

3.3 Data Extension

Peer effects within this data set were already researched in multiple studies (Boozer and Cacciola, 2001 and Krueger, 1999), however compositional peer effects were not a part of any study up to now. Since the data offers the possibility to consider some variables as composition, if they are transformed using the given information, I am analysing the impact by these compositions and will later compare them to the known results of the typical used peer effects.

As an extension a variable considering the composition of grades within each classroom is added to each student. This classroom test score composition contains the distribution of a class' test score in a certain subject in kindergarten. For this fitting intervals for the distribution of the scores in the subject and grade are selected, considering the scores over all students and minimizing the zeros produced in each interval. Then the percentage of students scoring within each interval is calculated for each class recorded during the project in kindergarten. The respective class distribution is added to the information of each student in the class.

The main challenge to select fitting intervals is to look for good borders, so that as few zeros as possible are introduced to the data set, as they will have to be imputed later on and could lead to an significant change in the model estimation. Keeping this in mind the intervals presented in Table 3.4 were chosen.

Due to the relative information added to the data set a compositional approach is needed in the empirical study. For a first analysis of the model a set of four equidistant intervals was selected. Table 3.4 is reporting the descriptive statistics of these compositional variables. It is clearly visible in the minimum column that this selection of intervals leads to a lot of zeros for whom a fitting approach for their handling will be needed, if a regression with an coordinate transformation of compositional terms is supposed to be used. The exact number of zeros produced across each interval are presented in Table 3.5. The high number of zeros in each interval show how widely the distribution of SAT score in the subjects can differ across different classes and schools. For example

Table 3.4: Descriptive Statistics of the Compositional Terms computed in both Data Sets

Math				Reading			
Intervals	*Center*	*Min*	*Max*	*Intervals*	*Center*	*Min*	*Max*
$<$450	0.2129	0	95.7	$<$425	0.3362	0	100.0
450–480	0.2564	0	71.4	425–445	0.2839	0	78.6
480–510	0.2409	0	53.8	445–465	0.2055	0	68.8
$>$510	0.2898	0	86.7	$>$465	0.1744	0	100.0
Total Variance		1.3224		*Total Variance*		1.7779	

Note: The formula for the center and total variance are as they were defined in section 2.5. The test scores are those of the kindergarten and only for the students for whom the kindergarten test score and the 8th grade test score is available.

Table 3.5: Number of zero produced across each interval in the math and reading data set

Math		Reading	
Intervals	*Number Zeros*	*Intervals*	*Number Zeros*
$<$450	330	$<$425	197
450–480	77	425–445	52
480–510	67	445–465	233
$>$510	130	$>$465	692
Total	604	*Total*	1174

the score distribution variables within some classes are only non-zero in the highest two intervals and other classes only had scores within the lower ranges. Another issue which could lead to this amount of observed zeros is that the classes of small size have between 13 to 17 students, so that for this low number of students the scores maybe are not spread completely across all four intervals. Therefore, a number different from zero could most probably be observed in those intervals, if the sample set, in this case the classroom, had a larger amount of students. Hence, the zeros in the added class composition variables are categorized as count zeros, as defined in section 2.4.

Table 3.4 further presents the center values of each composition which add up to one and show where on average the higher or lower ratios of the composition are. The ratio of students within a classroom in kindergarten for the math score is on average the highest in the range over 510 and for the reading score below 425, so that most of the students of a classroom score in these ranges for the respective subjects.

In a later part of this thesis, in section 5.2, it will be checked, if the width of the selected composition intervals does affect the regression results and lead to different results or if they only change the accuracy of the regression. However, in the next step the fitting zero imputation method is selected.

4 Model Building and Zero Handling

4.1 Methodology GAM

The fourth grade scores are modeled using Normal regression models. Therefore, the fourth grade score is denoted by Y_i for student i and the model is denoted by $Y_i \sim \mathcal{N}(\mu_i, \sigma_i)$ with $\mu_i = E(Y_i)$ representing the expected score in grade four by the student i and σ_i is the variance for student i. An identity relation is used as the link between the mean and the predictor η_i, i.e. it is set to $\mu_i = \eta_i$. Due to the normal model, the probability density function for the response is assumed to be

$$P(Y_i = y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma_i}\right)\right) \tag{4.1}$$

Then the predictor including continuous, categorical and compositional data can generally be expressed by

$$\eta_i = \beta_0 + \eta_i^{\text{categorical}} + \eta_i^{\text{continuous}} + \eta_i^{\text{composition}} + \eta_i^{\text{re}}, \tag{4.2}$$

where β_0 represents the intercept, all categorical covariates are represented by $\eta_i^{\text{categorical}}$, the continuous covariates are contained in $\eta_i^{\text{continuous}}$, the term $\eta_i^{\text{composition}}$ represents the compositional predictor and the classroom random effects are specified in η_i^{re}.

As introduced in section 2.7.1 the compositional covariates are ilr-transformed before any method is applied and are then included into the compositional predictor. For the class score distribution variables denoted by $x_{im} \in S^D$ for student i and in the intervals $m = 1, ..., n$, selected in the data extension described in section 3.3, the predictor term has the form

$$\eta_i^{\text{composition}} = \sum_{m=1}^{n-1} \beta_m^{\text{comp}} \text{ilr}(x)_{im} = \sum_{m=1}^{n-1} \beta_m^{\text{comp}} z_{im}. \tag{4.3}$$

As the identity link function is used, the interpretation of the compositional data based on the results of section 2.7.2 is: The relative ratio change of α in component $m = 1, ..., n$, affects the response scale by the factor $\ln(\alpha)\text{clr}(b)_m$ with $b = \text{ilr}^{-1}(\beta^{\text{comp}})$.

The model framework of choice is the generalized additive model (GAM), which has first been introduced by Hastie and Tibshirani (1986). This flexible statistical method supports the incorporation of non-linearities as the regression effects are often non-linear. Therefore, the continuous covariates, in our data set, the test scores from kindergarten, are expressed in the following way

$$\eta_i^{\text{continuous}} = \sum_{j=1}^{J} f_j(u_{ji}) \tag{4.4}$$

where u_{ji} are the continuous explanatory variables and f_j represents a smooth function for the j-th continuous explanatory variable. As Hastie et al. (2009, pp. 295–296) explains, these smooth - alternatively called nonparametric - functions are fit by a scatterplot smoother. Typical examples of these smoothers are the cubic smoothing spline, kernel smoother or P-splines. For fitting an algorithm estimating all J functions simultaneously is needed. The first algorithm to do this was

proposed by Hastie and Tibshirani (1986) and it is the basis of the application methods for GAMs until this day. As these models keep the additive form as it is known from the classical linear regression models, the interpretation does also not change and can be kept the same as before.

A link function is used to link the response to the predictor. Depending on the exponential family and character of the response variable, different sets of link functions are available. As described above, the response is fitted using a normal model with an identity link. The general form of the GAMs is similar to the generalized linear models (GLM) and due to this, one can easily find an additive model for every GLM model by changing the linear term to the smoothing function (Hastie et al., 2009, p. 296).

The application of GAM models in R is supported by the package `mgcv` (Wood, 2017), which provides automatic smoothing parameter selection for multiple penalized spline components. In addition, a set of different smoothing functions and exponential families with multiple link functions is provided.

The approach of the GLM is employed to include categorical and compositional terms into the model. So that the terms are added linearly, i.e. the predictor has the form

$$\eta_i^{\text{categorical}} + \eta_i^{\text{composition}} = V_i\beta^{\text{cat}} + Z_i\beta^{\text{comp}} \tag{4.5}$$

where V_i is the row vector of the matrix of only categorical covariates, which contains all categorical covariates for student i in each row, β^{cat} represents the parameter vector corresponding to these covariates and Z_i is the i-th row vector of the matrix of the ilr-transformed compositional covariates and β^{comp} the corresponding parameter vector of the compositional terms. The linear terms are then linked to the response via the same link function used for the continuous covariates while putting them in the GLM model setting. The mixed model using linear and non-linear effects can also be fitted by the function `gam` of the R package `mgcv`.

In the last step, the predictor of the random effect η^{re} is added to the model. The resulting random effects are assumed to be independently and identically distributed (i.i.d.) normal with an unknown variance which is estimated. Due to this assumption, the random effects can be treated like a smooth term where the coefficients can be subject to a ridge penalty, i.e. an identity matrix as a penalty, as it is of full rank and no further centering constraints are required (Wood, 2008).

To apply the forward or backward stepwise variable selection to a fitted model, the AIC has to be computed for the estimated model parameters. If GAM is used for the model fit, the AIC formula changes to the following expression

$$AIC = -2\hat{\ell} + 2\text{EDF} \tag{4.6}$$

where $\hat{\ell}$ is the log-likelihood evaluated on the model parameter estimated by penalized maximum likelihood and EDF denotes the effective degrees of freedom. The effective degrees of freedom are a transformation of the smoothing parameter λ since it does not represent a meaningful measurement due to its' dependency on the units of the data and thus to obtain a scale-free measure of complexity (Harezlak et al., 2018, p. 28). The exact calculation of the EDF is described by Wood et al. (2016, p. 1556) and was applied to the AIC function in the R package `mgcv` (version 1.8-33). After applying the adapted AIC for GAM, the variable selection methods can be performed in the

same way as explained in section 2.7.3.

4.2 Dealing with Zeros in Compositions

Previously, in section 2.4, the possible methods for handling zeros in compositional data were introduced. However, in the data of project STAR, the class composition depends on the intervals which were chosen and this selection of intervals introduces zeros to the composition, as it was shown in section 3.3.

To treat these zeros, deciding which of the introduced methods fits the current data is crucial. The first instinct would be to say that it is a case of structural zeros, because no students scored within the specified interval in a particular classroom. However, the nature of the project was to assign students to classrooms of different sizes and research the effect on their educational achievements. Due to this, around a third of the classes participating in the project consisted of only 13 to 17 students. Therefore, it can be argued, see section 3.3, that the zeros can be assumed to be count zeros.

This distinction of the zeros is essential, because the process to impute zeros values would be different in both cases and there are limitations on which methods can be used for each type of zero. The most advances in zero handling approaches were made for rounded zeros in the last two decades. Several EM-based algorithms were developed and are now considered efficient methods to make a broader range of possible functions available. The research in the area of count zeros did also take some steps forward and even showed that the methods found in the area of rounded zeros can be applied as well (Martín-Fernández et al., 2015). However, there is still no general method for structural zeros (Martín-Fernández et al., 2011, p. 53–54). Different approaches, including conditioning by zero patterns, projection or imputation by constants, were previously discussed in section 2.4 and all included certain disadvantages. For example, the conditioning approach is only feasible on a data set, if the composition does not contain many different parts. This problem arises because for every zero pattern a data subset is used to compute the regressor per indicator and the number of required indicator variables is increasing exponentially with each part (Filzmoser et al., 2018). On the other hand, imputation of the zeros by a constant would not take into account that the true value of the part is zero and replace it by a small value, as this goes against the nature of a structural zero this approach is in general not supported (Boogaart and Tolosana-Delgado, 2013, Filzmoser et al., 2018).

Since the nature of the zeros is assumed to be count zeros, the methods for rounded zeros like the multiplicative replacement, implemented within the R package `compositions` as well as the EM-based methods by Martín-Fernández et al. (2012) and Templ et al. (2016) are applied and compared to the projection approach also used for structural zeros and the k-nearest neighbour approach by Hron et al. (2010) developed for missing values. Even though the last approach using the k-nearest neighbours was generally developed for the imputation of missing data in compositions, due to the similar concepts of zeros and missing values in compositional data analysis, it is included in the first analysis to compare the approaches on the data. This method has the limitation of not considering a detection limit and thus, the imputed values are not limited by a threshold and possibly contain larger values.

Other limitations for the approaches of imputation by a constant and projection, which corresponds to an imputation by the mean, include the fact that they are not taking the properties of the data set below them into consideration. This results in them being inflexible and can introduce an artificial correlation. Furthermore, none of the developed imputations include a multiple imputation approach to keep the risk of underestimation low. The multiple imputation approach was first looked at by Martín-Fernández et al. (2003b), however, in the most recent application methods, it was not implemented and most approaches use an EM-based method. Moreover, currently, none of the EM-based methods include the knowledge of the response into their imputation approach and thus can only use the information of the compositional data, even though a more efficient estimation of imputed values could be achieved using the complete information at hand.

Therefore, new methods were continuously developed in the context of zero imputation of compositional data sets, but there are still many challenges surrounding this topic. Especially for count and structural zeros, a fitting approach that can be applied to any data set is still missing.

4.3 Comparison of the Zero Handling Methods

In the first analysis the impact of the chosen zero imputation method on the coefficient estimates of the composition is observed. These results serve as a decision basis on which imputation method is selected later on. Therefore, the data of the compositions will first be imputed by each method mentioned above: multiplicative replacement, EM-based methods by Martín-Fernández et al. (2012) (`impRZilr`) and Templ et al. (2016) (`imputeBDLs`), projection and k-nearest-neighbours (KNN) method by Hron et al. (2010). Afterwards the compositions are ilr-transformed, $\mathrm{ilr}(x_{is}^{\mathrm{comp}}) = z_{is}^{\mathrm{comp}}$, and then included into the specified GAM model of the expression

$$\begin{aligned}\eta_{is} =& \beta_0 + \langle \beta^{\mathrm{comp}}, z_{is}^{\mathrm{comp}} \rangle_A + \sum_{j=1}^{J} f_j\left(x_{jis}^{\mathrm{score\_k}}\right) + \beta_1 x_i^{\mathrm{sex}} \\ &+ \beta_2 x_i^{\mathrm{race}} + \beta_3 x_i^{\mathrm{clsize}} + \beta_4 x_i^{\mathrm{freelunch}} + \beta_5 x_i^{\mathrm{schid}} + \delta^{\mathrm{clid}},\end{aligned} \tag{4.7}$$

with the covariates, *comp* as the class composition in the subject s in kindergarten, for which the mathematics and reading scores are used and *score_k* is the individual score in the subject s in kindergarten. Further covariates include *sex* as the students gender, *race* the students ethnicity and race, *clsize* the class size the student belonged to in kindergarten, which is either a small class or a regular class and *freelunch* the free lunch status in kindergarten. The school id denoted as *schid* is added as school-fixed-effect and the class id denoted as *clid* is added for the class-random-effect of the students in kindergarten.

The observed results for both the model with the math test scores and the model with the reading scores show similar behaviours based on the zero imputations used. Therefore, only the results for the model using math scores are included in Figure 4.1 and the results using the reading score are included in the appendix. The multiplicative imputation is called imputation by a constant in Figure 4.1, as it is an imputation by a constant to which then a closure is applied.

Except for the compositional terms, which show more diverse parameter estimations, all other non-compositional terms, including the intercept, are in a close range between the model estimations as

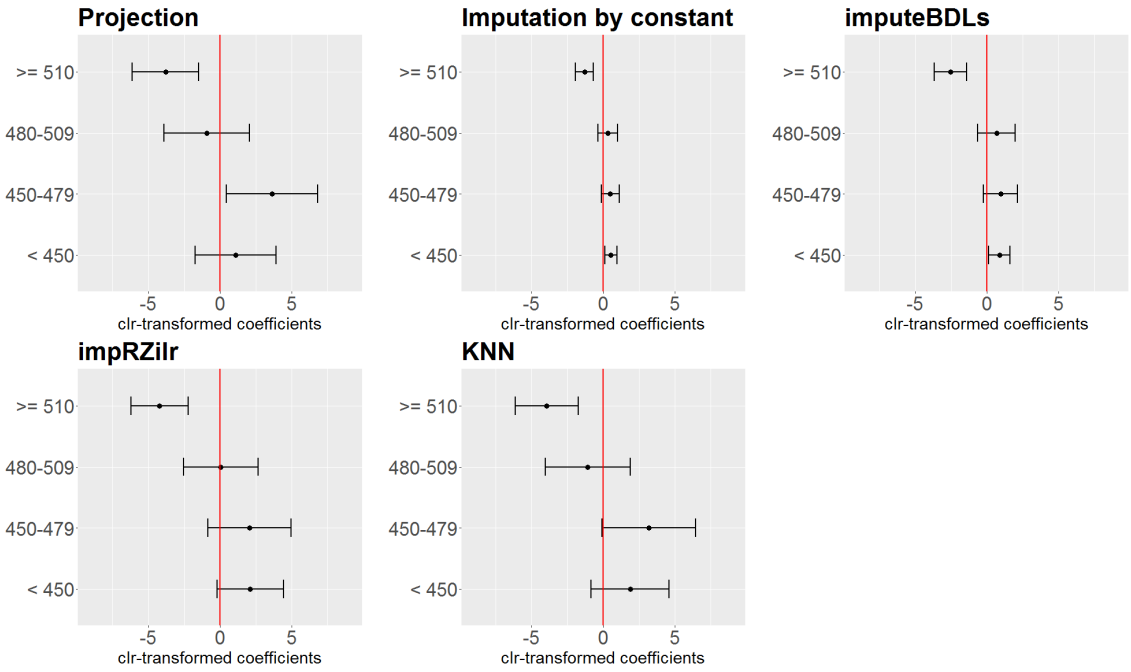

Figure 4.1: clr-transformed compositional estimates and their confidence bounds in the math score model for the composition of the kindergarten math scores (devided into scores below 450, 450 to 479, 480 to 509 and above 510)

expected since the underlying model and data are the same. Moreover, the tests regarding model assumptions of homoscedasticity and normality correspond in all models of both data sets. This includes that all models support the homoscedasticity assumption and that across the reading models, the normality assumption holds, but in all math models, the same kind of normality issues show. Thus the zero imputation method does mainly affect the compositional estimates. So that for the evaluation of which method is appropriate, the focus will only be on the results of the compositional model parameters.

For the imputation by a constant value, zeros are replaced by two-thirds of 10^{-6}, which was set as the detection limit. For the algorithm `imputeBDLs` a robust linear regression model is used and the predictors with the lowest variation are not chosen in every step. In the case of the `impRZilr` algorithm, a robust regression using an MM estimator is solved and for the k-nearest-neighbours function, the approach using the mean is applied.

In Figure 4.1 the points represent the estimated coefficients for the composition after they got transformed into the Aitchison geometry using the inverse ilr transformation and then clr-transformation to get an interpretable result, as explained in section 2.7.2. The red line shows the zero point as a reference which means the compositional part does not impact the response if it lies on the line. The positive and negative impacts on the response can directly be understood by the sign of the coefficient. Since β can be rewritten as $\beta = \mathrm{ilr}(b) = U^T \mathrm{clr}(b)$, as shown in section 2.3.3, and it holds that $UU^T = I_D - \frac{1}{D} 1_D 1_D^T$, the clr transformed coefficients can be reexpressed as $\mathrm{clr}(b) = U\beta$. Using this knowledge, the covariance matrix for the clr transformed covariates is computed by $U\hat{\Sigma}U^T$ with the estimated covariance matrix $\hat{\Sigma}$ of the coefficient β. Based on this, in addition to the coefficients, their clr-transformed 95% confidence intervals are visualized in Figure 4.1.

The differences in the compositional estimates of the model are highly significant for choosing a method later on, as the data set introduces a high amount of zeros in the compositional terms. The most noticeable result in Figure 4.1 is that the imputation by a small constant leads to coefficients very close to the neutral element, which is the composition of all zeros after a clr-transformation. Thus the impact in the interpretation is large and in general, no impact on the response is observed. As this estimation does not coincide with the estimators of other imputation results, the interpretation of these coefficients can be misleading due to the imputation.

The EM-based algorithm `imputeBDLs` is also resulting in coefficients closer to the neutral element, but is not as strongly affecting the estimators as the imputation by a constant. Even though the interpretation is possible using this imputation method, each part of the composition will be similar. Moreover, depending on the choice of intervals, this function does not always lead to feasible results in this context. Therefore, it may not be the best fitting method for the data set at hand.

The other approaches, including projection, `impRZilr` algorithm and k-nearest-neighbour algorithm, are producing very similar estimates and only have some minor differences, which are mainly in the two inner intervals. The outer borders of the intervals, which were selected, show only minor differences as well in the estimation of the three approaches. In both inner intervals, the estimates across the three approaches show a consistent positive and negative impact on the response, but the actual estimate is shifting its value. The two approaches that were not explicitly designed for rounded zeros, the projection and KNN, are similar because the projection can be interpreted as the geometrical mean of the clr-hyperplane and in the KNN method, the mean approach was used. Therefore, both methods are based on a mean value of the composition. Since the results of the two approaches do not differ a lot from the `impRZilr` algorithm by Martín-Fernández et al. (2012) using a robust regression, the usage of these methods seems to fit the current data set as well and would not skew the results.

In contrast to the results of the coefficients, the confidence intervals of the imputation by a constant and `imputeBDLs` algorithm are a lot smaller than the relatively large confidence intervals of the other approaches. This observation leaves a higher degree of uncertainty in the approaches with large confidence intervals and a less accurate predictor.

All in all, due to the large effect of the constant imputation and `imputeBDLs` algorithm on the estimated coefficients and the hardly interpretable results, as well as feasibility issues, they are not the best choice for the underlying data. Therefore, the imputation method of choice should be either projection, `impRZilr` or k-nearest-neighbour algorithm as they all show estimation results that are close together with no more considerable differences and which are also interpretable.

Using the AIC value to measure the goodness-of-fit for all models shows that in both subjects, the best model is the one using the `impRZilr` algorithm. It is closely followed by `imputeBDLs` in the math model. In the reading model, it is leading in front of the k-nearest-neighbour and projection approach.

So that the `impRZilr` algorithm was chosen for the zero imputation throughout all further models because it offers the possibility the get meaningful estimates which are interpretable. Furthermore, it belongs to the three methods with the least impact on the estimated parameters. Moreover, an advantage to the other methods is that this approach offers a robust imputation method where the

Table 4.1: AIC value of the model of each zero imputation method

	Projection	constant Imp.	`imputeBDLs`	`impRZilr`	KNN
Math Model	22461.58	22457.29	22454.64	22454.29	22458.81
Reading Model	27162.62	27173.45	27172.3	27156.96	27162.37

MM estimator is used in the regression so that this imputation is less sensitive to outliers and is performing the best prediction based on the AIC. The `imputeBDLs` algorithm has a robust regression as well, but it showed more computation issues depending on the selected intervals chosen and thus was not selected. Furthermore, was the `impRZilr` algorithm developed for rounded zeros and is thus imputing the zeros by values below a threshold which is also needed for the assumption of count zeros.

However, if computational very demanding calculations are performed, it can be considered to use the mean-based approaches like projection and KNN, since they are computationally faster than the EM-based method and lead to estimations almost the same as in the `impRZilr`.

5 Regression Analysis

After the differences in the zero imputation approaches are analysed and a suitable method is selected, another consideration in this context has to be: How do the selected intervals influence the estimation? Since the variable which is transformed into a compositional variable was originally continuous and not categorical, the parts of the composition are not fixed. Thus, the parts are chosen as intervals over the continuous variable and therefore, histograms over the covariate are considered as dependent variables. Hence, the next step is to refit the estimation for multiple sets of intervals to investigate the influence of the selection. Furthermore, the impact on the response will be estimated separately for the two class sizes to analyse whether the influence of the compositional variable changes. Finally, as the last step of the analysis, the long-term impact of the compositional covariate on the response is estimated until the students reach grade eight.

5.1 Baseline-Model without Composition

Before discussing the impacts of compositional group distribution, the impact of the standard peer effects as introduced in equation 1.1 are estimated. Each classroom is considered a peer group, so that the sample mean in each group, net of the individuals outcome, is the variable of interest. This quantity is also often referred to as the 'leave-out mean' denoted as $\bar{y}_{-i,cs}$ for individual i in subject s and in class c which has a total of N students with

$$\bar{y}_{-i,cs} = \frac{1}{N-1}(N\bar{y}_{cs} - y_{ics}), \tag{5.1}$$

where $\bar{y}_{cs}$ is the mean across each class and y_{ics} is the individual score. In observational data sets, sometimes the inclusive mean, the mean including the score of individual i, is used because only a fraction of peers of a cohort are available and thus, the inclusive mean is more representative of the cohort. As all individuals belonging to a cohort are known in Project STAR, there is no need to use an inclusive mean to study the impact on the response. Comparable studies of peer effects in the same setting have been studied (Boozer and Cacciola, 2001) and pointed out further group problems, including that each group is not an exogenous variable and even if the groups are formed in an exogenous way, the individual and group outcomes were simultaneously collected, the so-called 'reflection problem' by Manski (1993).

The concern regarding endogenous variables will be ignored in this context, as it was shown in previous studies that the assignment to each class is indeed random (Whitmore Schanzenbach, 2006/2007 and Sojourner, 2013). As expressed in equation 4.7 variables of interest are then given by the students demographics, including gender (`sex`), race (`race`) and free lunch status (`freelunch`) and their assignment to a class size, either small or regular class (`clsize`). This way the effect of the treatment of class size assignment on the subsequent test scores is captured as well. Additionally, the between-school effect is included to estimate the fixed effects across each school (`schid`), because the randomization in project STAR was executed within each school. As the last variable a class-random-effect δ^{clid} is added for the effect across each class. Then the predictor is expressed

Table 5.1: Regression estimates in the non-compositional peer effect model using the average class score

	Math Fourth Grade	Reading Fourth Grade
Intercept	657.99*** (24.788)	493.10*** (28.477)
Peers' Mean Test Score	0.048 (0.051)	0.269*** (0.067)
Female	6.989*** (1.662)	6.737*** (1.308)
White	7.270 (3.800)	10.139*** (3.071)
Non Free Lunch	17.438*** (2.077)	13.404*** (1.650)
Small Class	3.221 (1.887)	4.598** (1.562)
Class Random Effects	Yes	Yes
School Fixed Effects	Yes	Yes
AIC	23034.41	28043.93
R^2	0.1653	0.1747

Note: The standard errors are added in parenthesis. The stars show the significance based on the p-value from very significant '***' with values below 0.001 to '**' with values below 0.01 and '*' with a p-value below 0.05.

for individual i in class c and subject s by

$$\eta_{is} = \beta_0 + \beta_1 \bar{y}_{-i,cs} + \beta_2 x_i^{\text{sex}} + \beta_3 x_i^{\text{race}} + \beta_4 x_{ic}^{\text{clsize}} + \beta_5 x_i^{\text{freelunch}} + \beta_6 x_{ic}^{\text{schid}} + \delta^{\text{clid}}. \tag{5.2}$$

A normal regression using GAM is estimating the variables of interest and is presented in Table 5.1. Because almost all variables are categorical, the GAM corresponds to a general linear model (GLM) to which the class-random-effect was added. Summarising the results presented in Table 5.1, the peer effect in the model using the math test scores is small with a value of 0.05 and thus, the group behaviour had a considerably small impact on the test scores after the individual reached the fourth grade. However, considering the reading test scores, the peer effect is capturing a higher impact on the response with 0.27, so that for this subject, the peers had a long-lasting positive influence on the individual's performance, which could be captured. The estimates are generally in accordance with the previous papers by Boozer and Cacciola (2001) and Krueger (1999). The estimates for small classes are significant and higher than the regular classes by 3.2 points in the math scores and 4.6 points higher in the reading scores.

Furthermore, a large discrepancy between different socio-economic classes can be shown by the estimates of the free lunch variable. The students who are not receiving a free lunch have 17.4 points more in math and 13.3 points more in reading than those who receive a free lunch. So that there seems to be a big gap between students of better-earning families and students from low-income families. In accordance with other studies, it can be observed that girls and students who are white generally scored higher than the other groups.

After applying the backward stepwise algorithm to select the best model, the math score model without the class-random-effect is the best fit based on the AIC. For the model considering the reading score based on the AIC, the model without any school-fixed-effect and class-random-effect is the best fit. For the math score, the peer effect is not only small and has almost no impact on the fourth grade score, it is also not significant. Whereas in the reading model, peers have a significant effect and an impact on the fourth grade score of the individual. If the average peer score

in kindergarten increased by a unit, the individual score in the fourth grade increased by 0.2 points.

In the case of using the peer group mean to study the influence on the performance of an individual, the interpretation is straightforward. Suppose the mean increases, the individual's score later in time increases as well. However, if the class composition is used, the results are not as straightforwardly interpretable. It cannot be said that if the peer group scores higher, the individual scores higher as well. The correct interpretation would be: If the relative ratio within a peer group increases in a specific score range, the estimate shows how the individual's score is influenced. Thus, not only a general estimate of the cohort is used to see the influence, but the whole distribution of the variable is considered. It is shown how the combination of students in a cohort changes the individual scores.

5.2 Influence of Interval Selection

As a starting point to analyse the influence of the size of the selected intervals on the regression results, a set of three intervals is chosen for each of the kindergarten test scores in the two data sets. For selecting the first intervals, a set minimizing the amount of zeros and thus the potential effect of the zero imputation is searched and selected. After the first three intervals are set, sets of finer ranges dividing the initial interval ranges in half of the previous size while keeping the outer intervals constant are used. This separation leads to the selection of a set of three, four, six and ten equidistant intervals as specified in Table 5.2.

This separation of intervals keeps them comparable since the borders are related. The intervals were not split any further as some classes contain only 13 students. A set of 10 intervals is already bound to introduce zeros into the data set and even more intervals with certainty will introduce more zeros which can lead to skewed results through the imputation process.

For each set of intervals, the compositional term based on the kindergarten scores is computed as a proportion of students in each class who scored within those ranges. The zeros are imputed by the method `impRZilr` afterwards. For the imputation, a robust regression method with the MM-estimator is used so that the impact of outliers does not skew the estimation results in any other

Table 5.2: Overview of the selected intervals for the kindergarten scores per data set

Math				Reading			
3 Intv.	*4 Intv.*	*6 Intv.*	*10 Intv.*	*3 Intv.*	*4 Intv.*	*6 Intv.*	*10 Intv.*
$<$450	$<$450	$<$450	$<$450	$<$425	$<$425	$<$425	$<$425
450–510	450–480	450–465	450–458	425–465	425–445	425–435	425–430
$>$510	480–510	465–480	458–465	$>$465	445–465	435–445	430–435
	$>$510	480–495	465–473		$>$465	445–455	435–440
		495–510	473–480			455–465	440–445
		$>$510	480–488			$>$465	445–450
			488–495				450–455
			495–503				455–460
			503–510				460–465
			$>$510				$>$465
Total Number of Zeros							
468	604	1702	6983	911	1176	3025	9708

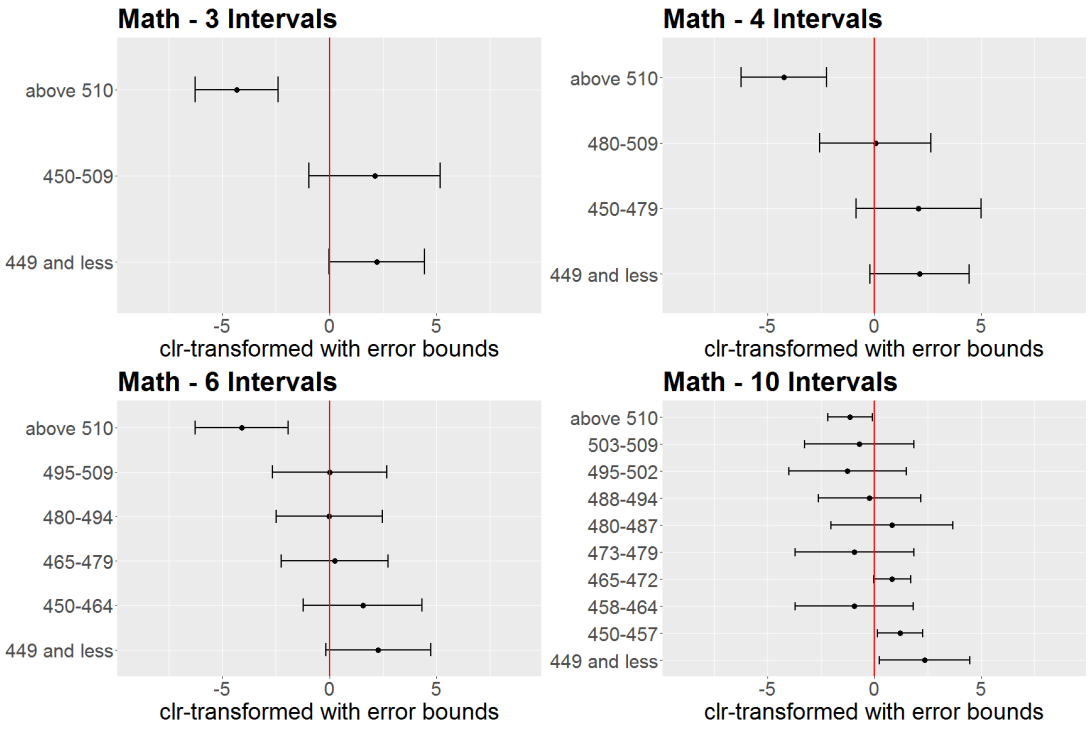

Figure 5.1: clr-transformed compositional coefficients with their confidence bounds for each interval selection of the math score model

direction.

In the previous model, it was unnecessary to include the individual's score into the peer group mean because enough observations were available to represent each cohort well. However, if the compositional terms are used as a group distribution, the individual's score is included to have enough observations to represent the group. A problem arising in this context is that students who scored in a specific interval are counted. Due to this, some intervals have a low number of students more often. Excluding each individual increases the number of zeros, although it is a problematic value in the log-ratio methods needed for the estimation.

The model stays constant while the compositional terms are replaced in each model fit. Hence, the model is refitted for each set of intervals in the composition. The results are shown in Figure 5.1 as the clr-transformed estimates and their transformed 95% confidence intervals, as it was done in section 4.2 considering the zero imputation methods. The direction of the estimators should not be contradicting in the corresponding intervals; otherwise, the regression is not robust to the selection of intervals for continuous variables. Therefore, it needs to be considered how the estimation can be adjusted to be robust to the selection.

The compositional estimates of the model using math scores and the one using reading scores show analogous results. Hence, only the estimators of the math score are visualized in Figure 5.1. The estimation results of the reading scores are added to the appendix. The coefficients of the outer intervals are generally constant throughout the models and correspond to the general trend of the parameters: The interval on the lower scores has a positive impact while the interval of the higher scores has a negative impact. In detail, this means, if the ratio of students within a class in kindergarten having higher scores increased, the score in fourth grade of an individual decreased.

On the other hand, if the ratio of students within a class in kindergarten scoring in the lower range increased, the individual score in fourth grade increased as well. This result is counter-intuitive; one would assume that an individual's score would increase if more of their peers performed better, as it was in the baseline model of the previous section.

For the other intervals between the outer ranges, it can be observed that the more intervals are added, the more their model parameters are close to the zero line. This means the parameter of each interval has a smaller impact on the response if the ratio within their class in this compositional part increases. This effect can even lead to the result that some parts of the composition have no impact on the response anymore, e.g. because none of the students have scored in this interval or only very few, which did not have a large enough influence on their peers.

Especially the clr-transformed estimates for ten intervals seem to be getting closer to the neutral element as more compositional parts are included. The main factor contributing to this observation is most probably the increasing number of zeros occurring in the model with the increasing number of intervals. Depending on the imputation method, this can have multiple effects. For example, in the applied EM algorithm to impute the zeros by a very small value below a detection limit, the general ratio of the composition is kept. However, the zeros occur so often because the number of intervals is too big for this data set that the actual data points cannot show their influence after the computation. Therefore, the results are normalized closer to the neutral element.

In detail, it can be observed in Figure 5.1 that for the intervals of the upper half, the response would decrease or not change if the ratio in these intervals increases. The predictors estimated for six intervals show that the fourth grade math test score decreases if a higher ratio of students scored in the highest interval, above 510 points, in their kindergarten classroom. No influence is shown on the fourth grade scores if a higher relative ratio of students with kindergarten scores in the intervals 480 to 494 and 495 to 509 is present in a class. If a higher relative ratio of students in the intervals below 479 is present in a class, the score in fourth grade increases. The interpretation for the other interval widths corresponds to this finding.

Quantitatively the estimates of the model using three intervals for the math test score lead to the interpretation that if the ratio of the highest interval doubles while keeping all others constant, the response decreases by $\ln(\alpha)\mathrm{clr}(b_1) = \ln(2) * (-4.329) = -3.001$ points. On the other hand, if the ratio in the lowest interval doubles while the others are constant, the score in fourth grade increased by $\ln(2) * 2.210 = 1.532$ points and if the ratio in the third intervals doubles, the response increases by $\ln(2) * 2.119 = 1.469$ points.

Considering the reading score, the direction of the coefficient corresponds to the observation of the math test scores. In that setting, a higher relative ratio of students scoring low in the kindergarten class increases the score in the fourth grade and a larger relative ratio of students scoring high decreases the fourth grade score. In the case of three, four and ten intervals, the results are coinciding as well.

This effect could be explained by teachers responding to the class level so that their way of teaching is adapting to whether more students are getting higher scores and can follow the class faster or more students have problems following the class contents. If a higher ratio of students is in a classroom, it could be that more of the students scoring low are left behind. In contrast, if more

Table 5.3: Estimates of the non-compositional terms in each of the GAM models

	Math Fourth Grade				Reading Fourth Grade			
	3 Intervals	*4 Intervals*	*6 Intervals*	*10 Intervals*	*3 Intervals*	*4 Intervals*	*6 Intervals*	*10 Intervals*
Intercept	693.88***	694.43***	696.34***	701.28***	611.00***	614.48***	614.00***	622.56***
	(9.312)	(9.260)	(9.333)	(9.297)	(9.045)	(8.910)	(9.022)	(8.741)
Female	6.005***	5.985***	6.024***	5.953***	4.975***	4.986***	4.970***	5.016***
	(1.460)	(1.459)	(1.463)	(1.465)	(1.121)	(1.121)	(1.121)	(1.122)
White	-2.308	-2.245	-2.240	-2.292	5.485*	5.367*	5.382*	5.377*
	(3.354)	(3.354)	(3.358)	(3.361)	(2.628)	(2.628)	(2.627)	(2.632)
Non Free Lunch	10.703***	10.670***	10.722***	10.832***	4.489**	4.401**	4.360**	4.403**
	(1.842)	(1.842)	(1.844)	(1.846)	(1.441)	(1.442)	(1.442)	(1.443)
Small Class	0.300	0.453	0.089	-1.092	4.581**	4.543**	4.776**	4.799**
	(1.766)	(1.775)	(1.780)	(1.849)	(1.522)	(1.502)	(1.509)	(1.553)
Class Random Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
AIC	22453.74	22454.26	22457.9	22462.9	27158.44	27157.77	27158.74	27164.21
R^2	0.3649	0.3650	0.3649	0.3649	0.4054	0.4053	0.4051	0.4054

Note: The standard errors are added in parenthesis. The stars show the significance based on the p-value from very significant '***' with values below 0.001 to '**' with values below 0.01 and '*' with a p-value below 0.05. The R^2 value denotes the adjusted R^2 of the model.

students in a class are scoring low, then the teacher is trying more to teach in a way that everyone can follow more quickly and thus, the scores, later on, can also be influenced positively.

Since the minimum class in the data set is 13 students, the maximum of intervals should not exceed 13 as well, but to get interpretable results, it should be smaller, e.g. six intervals maximum. This way, there are enough students distributed over these intervals to receive meaningful results.

The other variables, including females, white students, receiving a non-free lunch and assignment to a small class correspond to the previous estimates in section 5.1. Table 5.3 shows that females score across all interval selections in the model with the reading scores around 4.9 to 5 points higher, white students scored 5.4 to 5.5 points higher and students with non-free lunches have 4.3 to 4.5 points more. The influence of a small class on the individual's performance is also across all models using the reading score constant around 4.5 to 4.8 points. Considering the models with math test scores, females scored on average 6 points higher and students with non-free lunch had 10.7 points higher. Non-significant variables are the class size variable indicating small classes for which students received from -1.0 points less to 0.4 more and the variable indicating white students which scored on average 2.2 points less.

In Table 5.3 the AIC value shows that with an increasing amount of intervals, the value of the AIC is increasing as well. Thus the models' predictive performance is decreasing. Using the math scores, the model with three intervals has the best fit based on the AIC. Whereas in the model including the reading score, four intervals lead to the best AIC value. Based on the model, it can be observed that if the number of intervals is too small, an underestimation of the predictor can happen. If too many intervals are separated, the model overestimates the predictor. Thus using the AIC as a measure, the best prediction based on the choice of intervals can be selected for each model. Using the adjusted R^2, one can then further measure how well the model explains the data. Comparing these values with the results retrieved in Table 5.1 shows an increase of the adjusted R^2 value in the models of each subject. Hence the data is explained more by the models considering a compositional term.

Table 5.4: Variables included in each model after variable selection using the AIC values

	Math Intervals				*Reading Intervals*			
	3	4	6	10	3	4	6	10
Intercept	X	X	X	X	X	X	X	X
Female	X	X	X	X	X	X	X	X
White					X	X	X	X
Non Free Lunch	X	X	X	X	X	X	X	X
Small Class					X	X	X	X
Class Random Eff.					X	X	X	X
School Fixed Effects	X	X	X	X	X	X	X	X

Applying the backward stepwise variable selection with the AIC value on all models in both subjects leads to the variable selection displayed in Table 5.4. Considering the math test scores, the variables class composition, gender, free lunch status, individual score and school-fixed-effects are significant. In the case of the reading score, no variables get omitted from the model because the AIC value does not improve with any omitted variable. This shows that the significant variables do not change

with the change of the selected intervals. Nevertheless, more variables are omitted from the model compared to the baseline model for the math scores. The peer effect, which is non-significant in the baseline model, is significant as the compositional peer effect.

5.3 Influence of Interval Selection per Class Size

As discussed previously, some classes can have only 13 to 17 students and are considered small classes. Other classes in the data set contain 22 to 27 students and are considered regular-sized. Therefore, it is possible that a different distribution of test scores is present, based on the class size or that the impact of the covariates changes based on the size of the class an individual attended. The compositional term is computed in different intervals for each class size to cover these cases. Then one model per subject is fitted to analyse the impact of the covariates on the response separately for each class size.

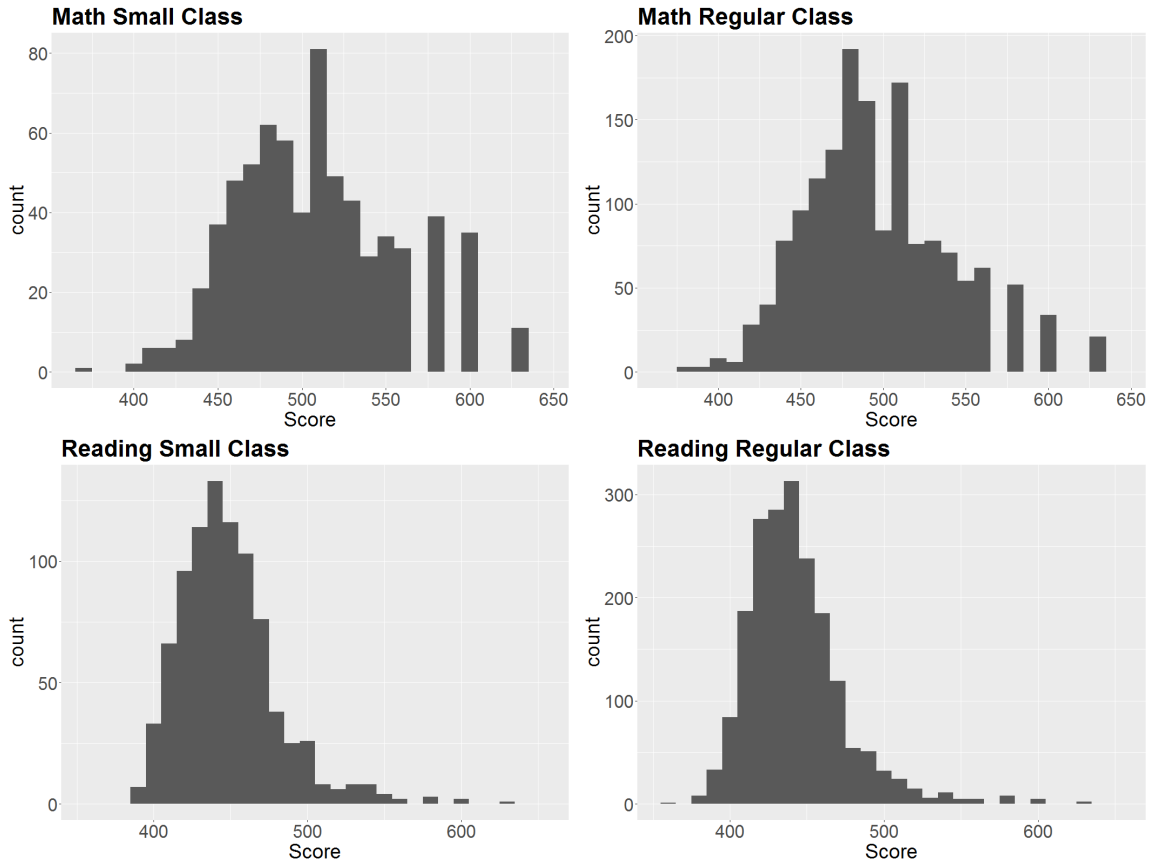

Figure 5.2: Distribution of kindergarten test scores for math and reading separated by class size which was either small or regular over all students participating in project STAR while in kindergarten

Therefore, the two data sets, based on the two subjects, math and reading, are further separated for the two class sizes. For the math test score and small classes, the remaining set consists of 693 students and for regular classes of 1566 students. In the data set of the reading score, 875 students attended a small-sized class and 1947 students a regular class. The distribution of the kindergarten test score in each of the described data sets is shown in Figure 5.2.

If the random assignment to the classes was carried out as planned, the student demographics in

Table 5.5: Overview of student demographics per class size in the math and reading data set

	Math Data Set			*Reading Data Set*		
	N	*Mean*	*SD*	*N*	*Mean*	*SD*
Small Classes	693			1566		
female		0.5007	0.5004		0.5086	0.5002
non free lunch		0.6768	0.4680		0.6068	0.4887
white		0.8398	0.3670		0.7326	0.4429
Regular Classes	875			1947		
female		0.5172	0.4999		0.5141	0.4999
not free lunch		0.7075	0.4550		0.6266	0.4838
white		0.8563	0.3509		0.7442	0.4364

Note: This table reports the descriptive statistics separately for the class size of the students considered in the math score data set and in the reading score data set. Gender is measured by an indicator taking the value 1 if the student is female and 0 otherwise. Non free lunch takes the value 1, if a student is not receiving a free or reduced-price lunch and 0 otherwise. Race takes the value 1, if the student is white and 0 otherwise.

small and regular classes should be the same as well as they should correspond to the distribution across the whole data set. Table 5.5 reports the mean and standard deviation of the student's gender, race and free lunch status. Across each data set belonging to either the math or reading data, the proportion of females, white students and the students who are not receiving a free meal in kindergarten are not showing big differences. Furthermore, the proportion of the students corresponds to the distribution of the whole data set, which is shown in Table 3.3. Considering these observations and the results found for tests of the randomization by Whitmore Schanzenbach (2006/2007) and Chetty et al. (2011), it can be assumed that the random assignment across the class types was executed correctly. Therefore, the regression results are not influenced by changes in the student's demographics in the class.

The previous section showed that it does not make sense to separate the composition into too many intervals due to the limited amount of students a class can have. To compare the two groups of small-sized and regular-sized classes, four sets of intervals are selected and compared similarly to the previous models. A set of three, four, five and six intervals is formed for each subject and class size. The selected intervals are chosen in a manner that the amount of occurring zeros is minimized. The selection is presented in Table 5.6.

Due to the separation into the class size, the general distribution of scores is different in both groups. In Figure 5.2 it is shown that the general distribution is similar between the two groups of each subject. Nevertheless, since the total number of observations is different, the counted students across each score are lower in the group of students in a small class. Thus the curve is flatter than the curve of regular-sized classes and the borders of the intervals are selected differently. Otherwise, it would lead to a higher amount of zeros within the data set of the small classes. Across both subjects, the main proportion of regular-sized classes is a little higher around the median of the score than in the smaller classes. Thus the range of intervals on the lower test scores is selected up to a higher value than the regular class intervals. The intervals in the upper range are also selected lower than the intervals of the regular classes. All remaining intervals are selected equidistant to each other.

Figure 5.2 above shows how the scores are distributed in the four cases. The math score distribution

Table 5.6: Overview of selected intervals for kindergarten scores per class size used to calculate the compositional terms

Math Regular Class				***Math Small Class***			
3 Intv.	*4 Intv.*	*5 Intv.*	*6 Intv.*	*3 Intv.*	*4 Intv.*	*5 Intv.*	*6 Intv.*
<460	<460	<460	<460	<470	<470	<470	<470
460–509	460–484	460–476	460–471	470–509	470–489	470–482	470–479
≥ 510	485–509	477–493	472–484	≥ 510	490–509	483–495	480–489
	≥ 510	494–509	485–496		≥ 510	496–509	490–499
		≥ 510	497–509			≥ 510	500–509
			≥ 510				≥ 510
Zeros				Zeros			
90	263	634	1098	181	253	489	995
Reading Regular Class				***Reading Small Class***			
3 Intv.	*4 Intv.*	*5 Intv.*	*6 Intv.*	*3 Intv.*	*4 Intv.*	*5 Intv.*	*6 Intv.*
<425	<425	<425	<425	<435	<435	<435	<435
425–439	425–431	425–429	425–428	435–454	435–444	435–441	435–439
≥ 440	432–439	430–434	429–431	≥ 455	445–454	442–448	440–444
	≥ 440	435–439	432–435		≥ 455	449–454	445–449
		≥ 440	436–439			≥ 455	450–454
			≥ 440				≥ 455
Zeros				Zeros			
187	515	1105	1754	239	665	1571	2851

in the subsets does not show the expected normal distribution, while the reading score shows a normally distributed variable. Thus the model, including the math score, can lead to issues with the normality, which will be discussed later.

After the computation and imputation of the new compositional terms the model of equation 4.7, exchanging the compositional terms only and using the separated data sets, is fitted. As before the compositional variables are included in their ilr-transformed form, $\mathrm{ilr}(x_{is}^{\mathrm{comp}}) = z_{is}^{\mathrm{comp}}$,

$$\begin{aligned}\eta_{is} =& \beta_0 + \langle \beta^{\mathrm{comp}}, z_{is}^{\mathrm{comp}} \rangle_A + \sum_{j=1}^{J} f_j\left(x_{jis}^{\mathrm{score\_k}}\right) + \beta_1 x_i^{\mathrm{sex}} \\ &+ \beta_2 x_i^{\mathrm{race}} + \beta_3 x_i^{\mathrm{freelunch}} + \beta_4 x_i^{\mathrm{schid}} + \delta^{\mathrm{clid}}.\end{aligned} \tag{5.3}$$

The terms for variable clsize are omitted from the model because the data was separated based on this variable and it is a constant in the models. All other variables for scores, gender, race and free lunch, as well as school-fixed-effects and class-random-effects, are kept in the models.

Table 5.7 reports the estimates of the regressions relating the mid-term educational attainments of the students to their class composition separately for the class size. Panel A and Panel B separate the results for the small-sized class and regular-sized class, respectively. The separation of the data set leads to similar results in the two subsets, which resemble the model considering the whole data set. Due to the small subset of small-sized classes, the estimators do not all show significance. The significant variables in both models correspond in the direction of the estimates but have minor

Table 5.7: Estimates of the non-compositional terms in the GAM models for math and reading scores

	Math Fourth Grade				*Reading Fourth Grade*			
	3 Intervals	*4 Intervals*	*5 Intervals*	*6 Intervals*	*3 Intervals*	*4 Intervals*	*5 Intervals*	*6 Intervals*
Panel A: Small Classes								
Intercept	723.70***	724.995***	694.45***	715.565***	625.50***	628.137***	630.75***	634.47***
	(20.450)	(21.307)	(24.809)	(24.934)	(18.846)	(18.865)	(19.883)	(20.416)
Female	4.526 (2.704)	4.403 (2.707)	4.498 (2.698)	4.445 (2.708)	2.741 (2.087)	2.598 (2.088)	2.814 (2.092)	2.761 (2.092)
White	-8.285	-8.346	-8.487	-8.766	-0.374	-0.507	-0.551	-0.548
	(6.402)	(6.402)	(6.396)	(6.416)	(4.819)	(4.821)	(4.821)	(4.822)
Non Free Lunch	11.794***	11.813***	12.008***	11.703***	4.072 (2.562)	3.977 (2.560)	4.041 (2.561)	4.027 (2.561)
	(3.274)	(3.274)	(3.264)	(3.276)				
Class Random Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
AIC	6933.66	6932.7	6931.09	6934.33	8480.06	8481.02	8481.29	8482.58
R^2	0.3471	0.3484	0.3450	0.3482	0.4151	0.4142	0.4148	0.4140
Panel B: Regular Classes								
Intercept	683.35***	682.72***	682.21***	683.03***	619.22***	619.80***	620.75***	620.12***
	(10.685)	(10.737)	(10.871)	(10.963)	(9.675)	(9.946)	(10.259)	(10.706)
Female	6.448***	6.420***	6.389***	6.404***	5.979***	5.994***	5.969***	5.918***
	(1.767)	(1.768)	(1.770)	(1.771)	(1.342)	(1.343)	(1.344)	(1.344)
White	-0.088	-0.055	0.042 (3.990)	0.144 (4.006)	7.083*	7.076*	7.150*	7.118*
	(3.988)	(3.989)			(3.146)	(3.149)	(3.150)	(3.153)
Non Free Lunch	9.600***	9.593***	9.610***	9.604***	4.893**	4.941**	4.917**	4.987**
	(2.284)	(2.285)	(2.285)	(2.287)	(1.768)	(1.770)	(1.771)	(1.771)
Class Random Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
AIC	15581.87	15582.89	15584.12	15585.53	18724.8	18725.38	18726.69	18729.03
R^2	0.3769	0.3769	0.3768	0.3767	0.3962	0.3965	0.3966	0.3964

Note: The standard errors are added in parenthesis. The stars show the significance based on the p-value from very significant '***' with values below 0.001 to '**' with values below 0.01 and '*' with a p-value below 0.05. The R^2 value denotes the adjusted R^2 for the model.

differences between both models. For example, in the model using math test scores, the intercept with a value from 724 down to 694 for small classes is higher than the intercept for regular-sized classes with a value around 683. The same holds for the students not receiving a free lunch in small classes. They have around 11.7 points more in the fourth grade, while they score around 9.6 points higher in regular classes. The other demographics show differences depending on the class size as well. The females have 4.4 points more in fourth grade if they attended a small class and 6.4 points more if they were assigned to a regular-sized class. The biggest difference is that white students assigned to a small class scored over 8 points less than the other students, but if they were assigned to a regular class, they scored only 0.1 points less than the other students. Therefore, the ethnicity and race of a student in regular classes did not impact the later score, but in the small classes, they did impact the scores.

In the case of the reading score, the intercept is similar for both class sizes. The small classes have a value of around 630 and regular classes of 620. However, except for the non-free lunch status covariate, which is 4.0 points in small classes and 4.9 points in regular classes, all other variables show some differences. This can be shown, for example, with the estimator for the coefficient if the student is white. In a small-sized class, white students score 0.5 points lower than the rest in the fourth grade, but white students get 7.1 points more than the other students in regular-sized classes. Thus, in this subject, the effect of ethnicity and race of students seems to have the opposite effect as in math. On the other hand, the influence on females corresponds to the results in the math score. In smaller classes, females have higher scores in fourth grade, around 2.7 points, but if they were in regular classes, they score even higher in fourth grade with around 5.9 points.

After the variable selection using the AIC as a measure on the best models, the model with three intervals of the reading score was selected in both class sizes. For the math models, the model with five intervals for the small classes and the model with three intervals for regular-sized classes are the best fit. All models are reduced to their significant variables. In the model with the reading score in small classes, the variables of the class composition, free lunch status and the school-fixed and class-random-effects are contained, while in regular classes, the full model without the class-random-effect is selected. Using the math scores, the models for both class sizes yield the best fit for the full model without the race and class-random-effect.

Thus, considering math scores, the significance of the variables does not change with respect to the class sizes. Nevertheless, the influence of each variable on the response changes depending on the class size. If the reading score is used, the significance in the variables with respect to the class size is changing. Hence, leading to different variables affecting the response. However, the variables contained in both class sizes for the reading score also have a similar impact on the response.

The compositional predictors, which are shown in Figure 5.3, separate the effects of the intervals belonging to the compositional peer effect on the individual scores. This allows to qualitatively interpret and visualize the effect of each interval section on the fourth grade score. Because the results for all sets of intervals in both subjects are showing the same behaviour, only a subset of figures is included in Figure 5.3 and the other results are added in the appendix.

The clr-transformed predictors of the math score show in Figure 5.3 that the students in classes with relatively more students who scored high in kindergarten have a lower score in the fourth grade. On the other hand, if students attended a class with relatively more students scoring in the lower part of

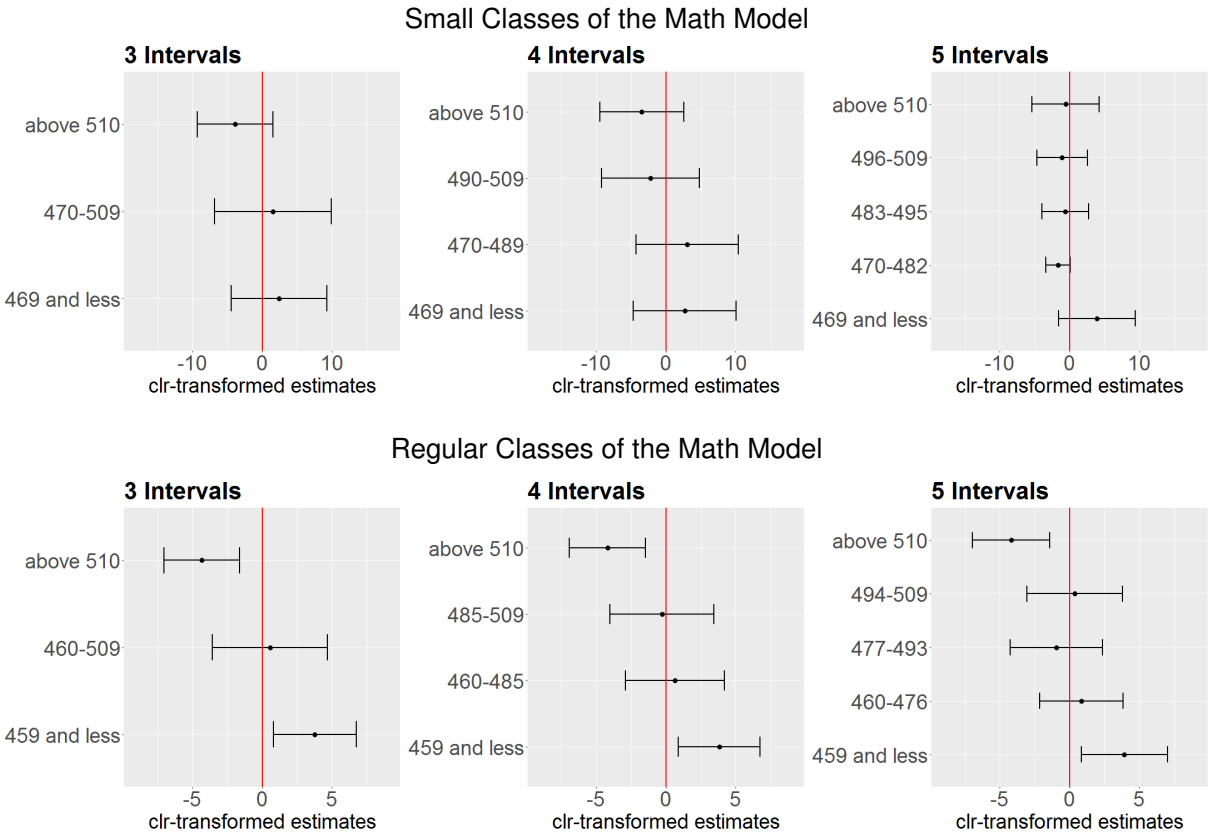

Figure 5.3: clr-transformed coefficients with their confidence bounds per class size for math test scores using three, four or five intervals as compositional terms

the distribution in kindergarten, they have a higher score in fourth grade. This result corresponds to the previous observations of the whole data set. The math model shows a linear behavior across all models over the intervals except for the finer range of the intervals of regular classes. In that case, some of the intervals, including higher scores, show a positive but relatively close to zero impact on the fourth grade scores. The previous observation of section 5.2, the finer the intervals get, the more they get closer to the neutral element, can be observed again in the subset of the small classes using the math data. Starting with five intervals, the compositional terms are leaning more towards the neutral element. In this case, already five intervals are introducing too many zero values in total across the whole distribution and cancel out the influence of all other occurring values. All in all, the results in the compositional terms between the two class sizes do not show considerable differences in the visualization.

In the case of reading scores, the clr-transformed predictors of regular-sized classes show as well that students in classes with a relatively higher share of students scoring in the highest section have a lower score in fourth grade. Conversely, students with a relatively higher share of classmates scoring in a lower interval have a higher score in fourth grade, see Figure 5.4. The predictors of the small classes head in the same direction except for some of the lower-mid intervals. Some intervals also negatively impact the fourth grade score, while in the regular classes, these intervals are close to zero. Even though the number of zeros increases in these subsets, if the amount of intervals increases, the impact of compositional terms stays the same as in the lower number of intervals. Thus the increase in zeros in this subset with compositions up to six intervals does not affect the

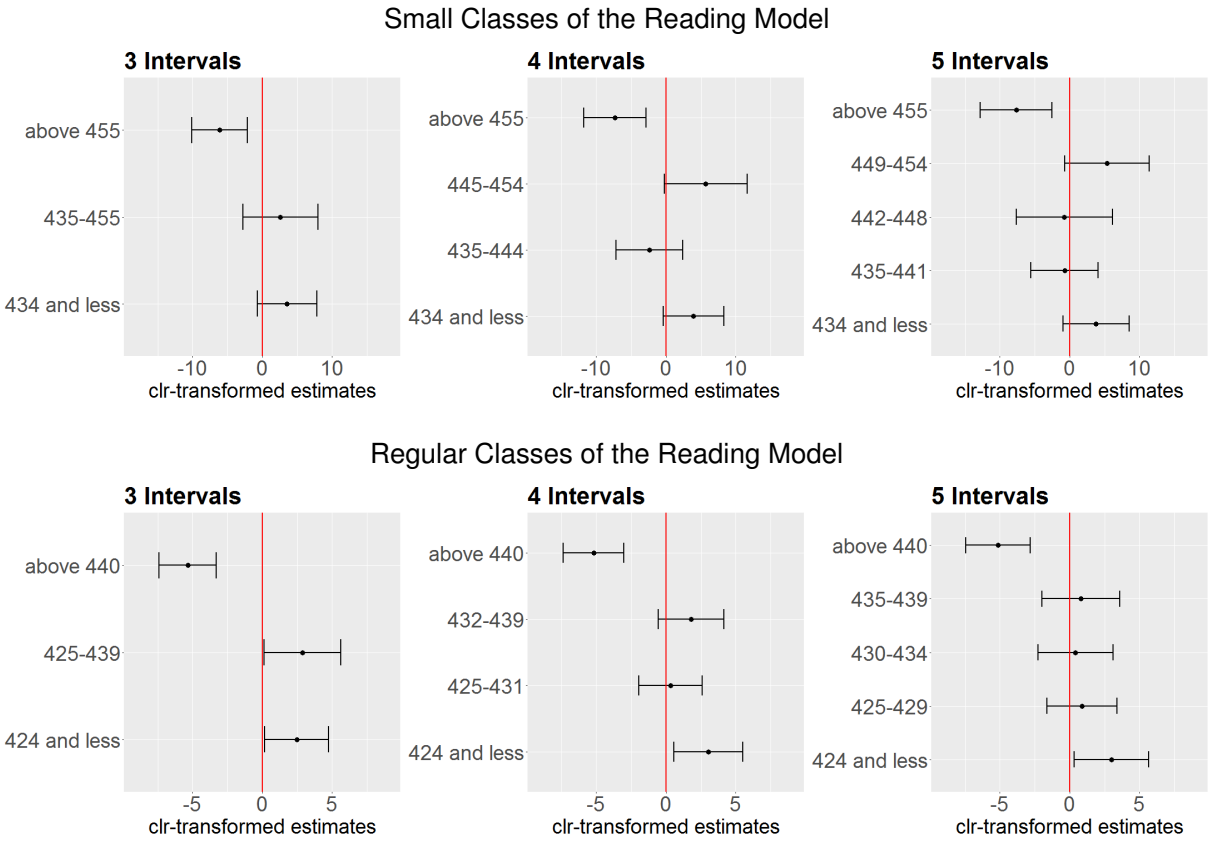

Figure 5.4: clr-transformed coefficients with their confidence bounds per class size for reading test scores using three, four or five intervals as compositional terms

regression results.

Considering these observations, there are differences between the groups depending on the class size. The main change lies in the variables of the student's demographics, as their impact on the response changes. The compositional terms have a similar influence on the response if the relative ratio in the composition changes, no matter which class size is assigned to a student.

5.4 Long-Term Impacts on Education Achievements

The previous analyses focused on the post-project score recorded in the fourth grade. To study the long-term impacts of the compositional peer effect, the influence of the peer effect on the grades up to eighth grade test scores is computed. Thus the CBTS test score of grades four to eight is used as a response, as it was in the previous analysis, while in first to third grade, the SAT test scores are the response. The version of the test in each grade is fitted for the level of the students in the associated grade. Therefore, the range of the score is not the same over all the grades.

This way, the impact of the compositional peer effect of the kindergarten classroom can be tracked throughout the students' entire school career. The four interval compositional terms of the students' math and reading scores at each grade level up to eighth grade are estimated. Since this separation of the class distribution showed constantly good model properties and interpretable results in the previous analyses. In this interval selection, the amount of zeros is reduced and thus does not affect the compositional estimates.

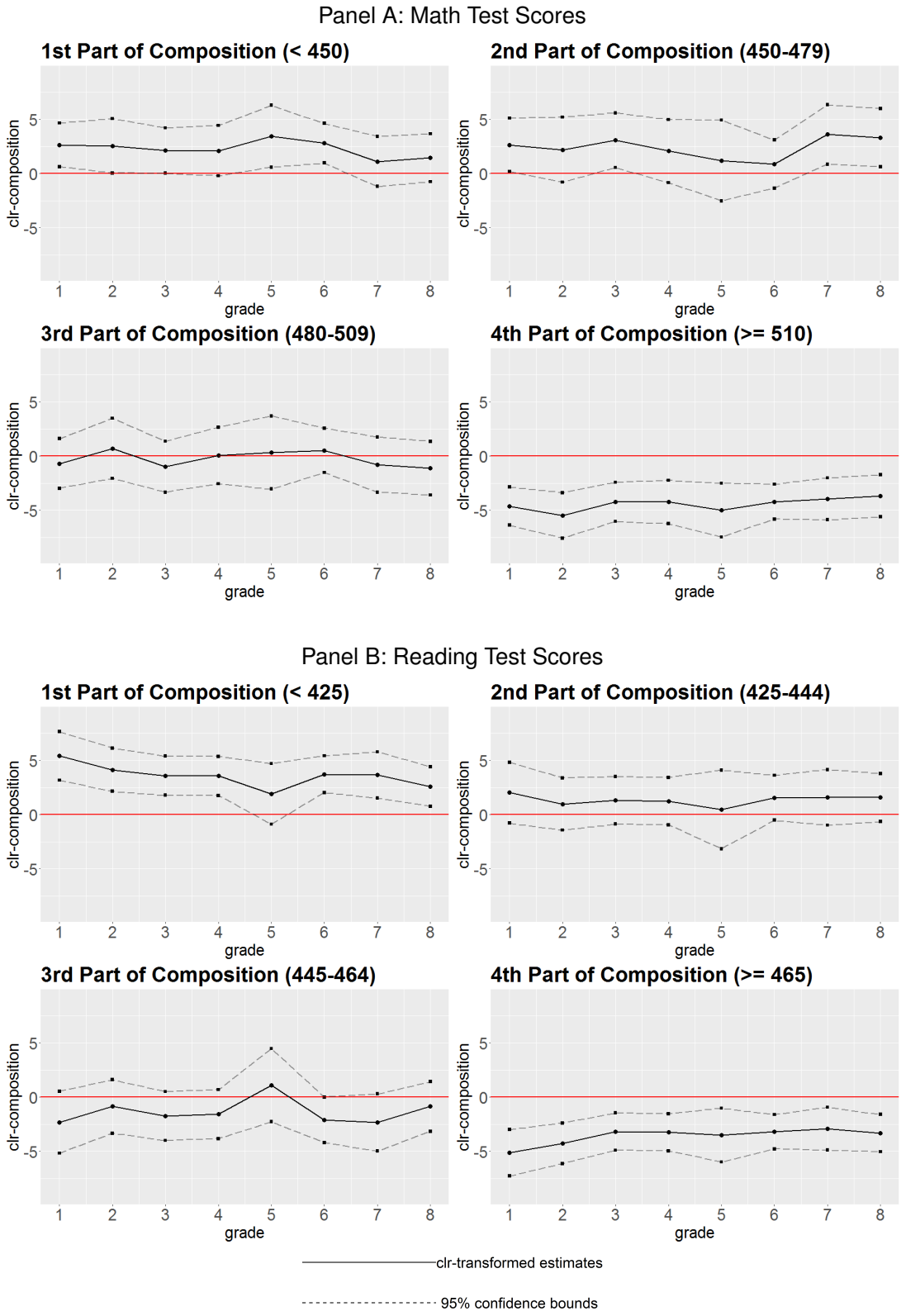

Figure 5.5: clr-transformed estimate (solid line) and their confidence bounds (dashed line) for the kindergarten class composition over each grade on the respective math and reading test scores

Figure 5.5 reports the estimates and confidence bounds of all models. Panel A contains the development of the impact of the four parts of the composition with kindergarten math scores and Panel B the development of the estimators using the reading scores for all later grades. The included estimators are computed by the full model of the form presented in equation 4.7 exchanging the response to each grades' respective test score. The clr-transformed estimator and the corresponding error bounds are computed, as explained in section 4.3 and plotted for each part, i.e. interval, of the composition separately in Figure 5.5. The straight line shows the impact of the clr-transformed composition estimates on later test scores over the students' school career from first grade to eighth grade. The dashed lines are the 95% confidence bounds computed separately for each estimate, respectively.

The direction of influence on the response is constant in each composition, except for the third part of the math composition. This part represents the proportion of students who scored in the upper half but were not under the top students overall schools while attending kindergarten. The impact on the response for the change of the relative ratio of students in this part stays close to zero, which means that there is no effect on the response by this part. At the same time, the estimates are changing signs, even if the impact increases in some of the years. Hence, this variable can be assumed not to have any considerable impact on the response.

In some of the components, a so-called wash-out effect of the impact can be observed. The general trend in each compositional part is toward the neutral element. This can be observed well in the first and fourth part of the composition for the math test scores and the second part up to the sixth class. For the second part containing the proportion of students in a class which scored between 450 and 479, the estimates suddenly increased, but a simple procedure can not find the direct reason. In the reading model, the wash-out effect can be observed as well in the first part, containing the proportion of kindergarten test scores below 425, and the fourth part of the composition, ratio of kindergarten test scores higher than 465. In the other two compositions, the estimates are not decreasing in such a linear manner; instead, they stay relatively constant overall grades.

The observed wash-out effect quantitatively means that if the relative ratio of students within a class increases in this part of the composition, the impact on the response, in both directions, decreases towards the neutral element and, in some cases, disappears completely. Thus, the spill-overs of the compositional peer effects in a class while the students visited the kindergarten are stronger in the short-term and lose their impact on an individual's performance in the long term. However, even though the impact is getting weaker in the long term, the actual direction the response is affected by the composition stays generally the same.

5.5 Post-Kindergarten Compositions and their Long-Term Impacts

Up until now, all models considered the end of kindergarten test scores which were recorded at the beginning of the project as the compositional term. Since the class structure in all four grades, recorded while Project STAR was ongoing, is known, a subsequent analysis on the compositional impacts of the other grades of the project can be performed.

To start the regression analyses on all grades, the interval selection has to be adapted for each of the three compositional terms. As mentioned above, the SAT test was changing across each

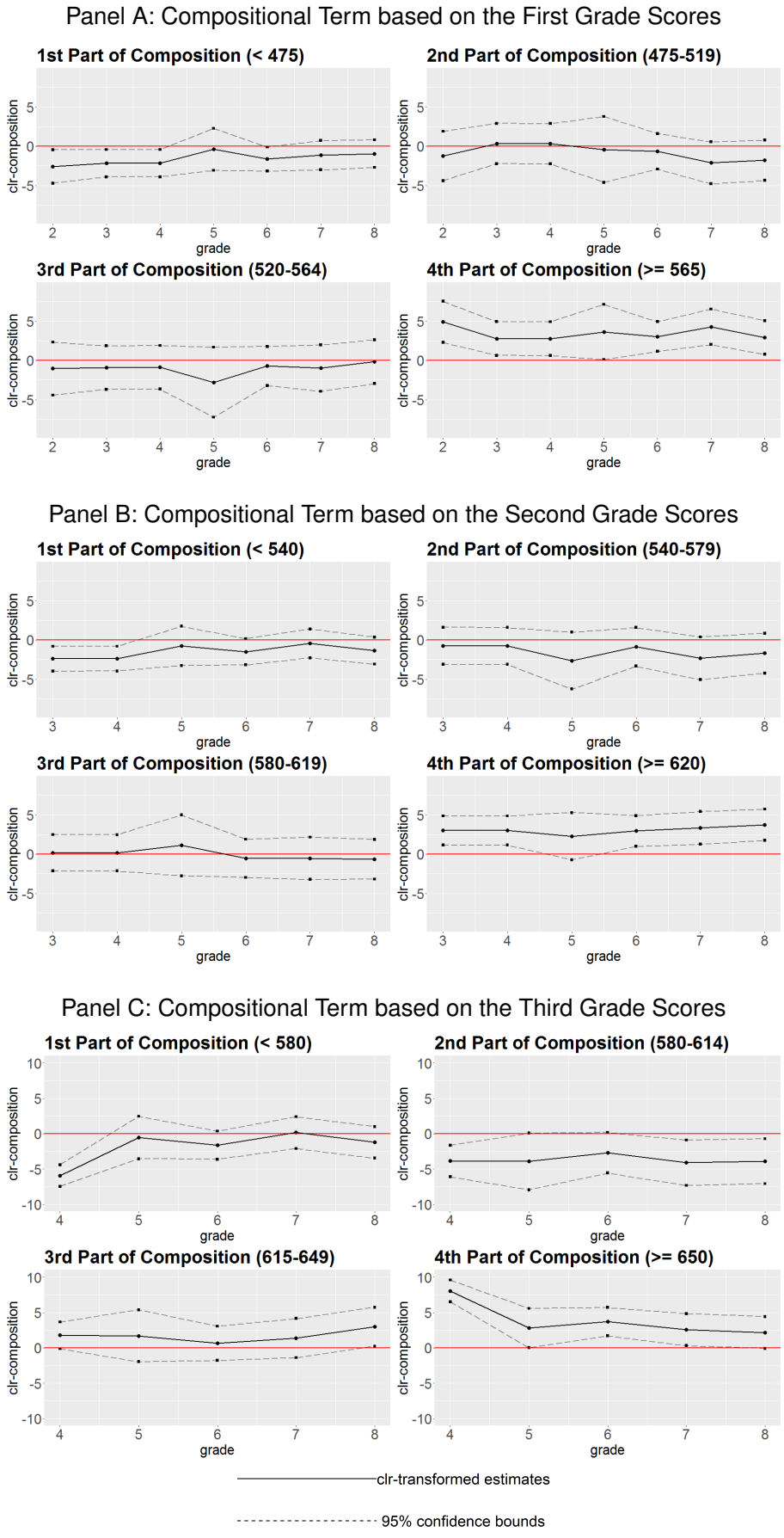

Figure 5.6: clr-transformed estimate and the confidence bounds for the class composition based on grades 1 to 3 in the models with the reading test scores of all following grades as response

grade as well as the range of the score. Since the range across the grades does not coincide, the intervals are chosen separately for each grade and the corresponding compositional term. The chosen interval borders for each grade are written in brackets in the title of each compositional part in the associated plot in Figure 5.6.

The models over all grades are fitted for each compositional term and the estimators belonging to the four parts of the composition are reported in Figure 5.6. In Panel A, the clr-transformed estimates and their confidence bounds for the composition of the test scores in grade 1 are shown. Panel B contains the corresponding figures for the second grade composition and Panel C for the third grade composition. Due to the properties of the models using the reading test scores, they are more suited for the purpose of analytical exploration and shown in Figure 5.6. The figure of the math scores shows corresponding results and is included in the appendix.

The class composition of the first grade shows opposing results compared to the composition of the kindergarten scores in the previous section. The first, second and third part of the composition, which cover the lower intervals, are negatively affecting the responses if the relative ratio of a class would increase in those parts. On the other hand, the fourth component containing the highest interval influences the responses positively if the ratio in this part increases.

Previously, the impact on the score was the opposite using the kindergarten test scores. The first and second parts had a positive and the third and fourth part had a negative impact on the response. Starting with the first grade, the compositional peer effect changed.

This observation continues in the other Panels as well. In Panel B of Figure 5.6, the third part has almost no impact on the responses anymore. The estimates in the other parts coincide with Panel A. Lastly, in Panel C, the influence of the third part has now turned positive. The other parts are still showing the same behaviour as in Panel A. Using the math test scores leads to similar findings and the same result: The third and fourth part of the composition, containing the intervals of the higher test scores, lead to an increase in the response if the ratio in a class scoring in those intervals of the associated grade increases. The first and second part of the composition, containing the lower test scores' intervals, lead to a decrease of the response if the ratio increases in those intervals. These are the expected results based on previous observations using peer effects based on peer means and similar values. To sum it up, the higher the ratio of students receiving higher scores is within a class, the more the individual score increases.

6 Simulation

6.1 Student Sample Specifications

In this section, subsequent to the previous estimation, the results are applied to a set of simulated data to assess the performance of the estimated model. Therefore, the data set is created assuming the response and kindergarten test scores are normally distributed. To keep the data structure according to the structure in Project STAR for the simulation, $N = 80$ schools are generated. For all schools, a random vector containing either a 1 or 2 is generated for each class type, i.e. small-sized class, regular-sized class and regular-sized class with a teacher aide. This vector initializes the number of classes of each type in the school, as there were schools with more than three classes included in the project. The vectors are generated with a probability of 0.2 for twos and 0.8 for ones corresponding to the distribution of the actual data set, which can be expressed as follows

$$c_1 \sim B(N, 0.2, 2), \quad c_2 \sim B(N, 0.2, 2), \quad c_3 \sim B(N, 0.2, 2), \tag{6.1}$$

where c_1 corresponds to the number of classrooms of small size, c_2 is the number of regular-sized classes and c_3 the number of regular-sized with teacher aide classes per school.

Similar to the number of classrooms per class type, the class size is randomly generated. The small classes are based on a uniform distribution across the numbers from 13 to 17 and the regular-sized classes are based on a uniform distribution across the numbers from 22 to 27. As the class size is a discrete number, the randomly generated numbers are rounded to the next whole number. If two classrooms of a class type are generated, then for simplicity reasons, the same class size is used for both classes.

$$n_{\text{small}} \sim U([13, 17]), \quad n_{\text{regular}} \sim U([22, 27]) \tag{6.2}$$

Each class data gets a random assignment of kindergarten scores based on the distribution of the kindergarten scores in the actual data set. Since the math score of the actual data has a mean of $\mu = 485$ and variance of $\sigma^2 = 2275$, the sample is drawn from the distribution $\mathcal{N}(485, 2275)$ for the math score. The reading score had the characteristics $\mu = 437$ and $\sigma^2 = 1005$, thus the distribution $\mathcal{N}(437, 1005)$ is used for the reading score.

The student demographics are then randomly assigned across all generated students. To simulate the gender, the uniform distribution $U([1, 2])$ is used, as there are only two possibilities, male and female, considered. Since in the actual data set, around 50% of the students were female, this is assumed in the simulated data as well. Therefore, the values are drawn from a uniform distribution and the values are rounded to the next whole number because a discrete case is considered.

For the free or reduced-price meal, the random variable is generated with a probability of 0.3 that a student received a free or reduced-price meal and 0.7 that they did not get a free lunch, i.e. a distribution of $B(N_{\text{Tot}}, 0.3, 2)$ with the total number of students N_{Tot}. The same applies to the students' race; the distribution of white and non-white students had a ratio of 80:20 on average. Therefore, this is assumed in the simulation and the data is drawn from the distribution $B(N_{\text{Tot}}, 0.8, 2)$. These variables are as well generated across all students and not based on their class.

This leads to a total number of N_{Tot} students, which varies over each simulation based on the generated number of classes and their size. Lastly, for all students, the fourth grade scores are computed by the generated data of the demographics and classes as defined above and the estimated model coefficients of section 5.2. Additionally, an error term is randomly added to the fourth grade scores, which was drawn from the distribution $\mathcal{N}(0, 2162)$ for the math term and $\mathcal{N}(0, 1490)$ in the reading score. The variance of the error term is based on the variance which was observed in section 5.2 for the scores in fourth grade for each subject, respectively.

6.2 Descriptive Statistics of Simulation Data

The simulation of the data set is repeated 100 times and to compare the structure generated by the random draw in the simulation and the actual data set of Project STAR, an overview of the descriptives statistics for the average of all variables across all simulated data sets are included in Table 6.1. The average, minimum, maximum and standard deviation of the average value computed in each data set are presented in each column. For computational reasons, the data sets for the math and reading score models were generated separately, both 100 times, respectively. Thus the

Table 6.1: Descriptive Statistics of the Simulated Data

	Min	*Mean*	*Max*	*SD*
Math Data Set				
Demographics				
female	0.4832	0.5004	0.5184	0.0073
non free lunch	0.6845	0.7004	0.7152	0.0060
small class	0.2249	0.2418	0.2689	0.0084
white	0.7907	0.8005	0.8175	0.0044
Test Scores				
kindergarten math score	483.257	484.947	486.762	0.6501
4th-grade math score based on 3 intervals	703.835	705.798	708.403	0.8630
4th-grade math score based on 4 intervals	703.023	705.678	707.478	0.8396
4th-grade math score based on 6 intervals	703.870	705.801	708.384	0.8205
Reading Data Set				
Demographics				
female	0.4844	0.5008	0.5216	0.0066
non free lunch	0.6847	0.7001	0.7109	0.0055
small class	0.2211	0.2394	0.2702	0.0081
white	0.7924	0.8006	0.8151	0.0046
Test Scores				
kindergarten reading score	435.943	436.993	438.010	0.3886
4th-grade reading score based on 3 intervals	618.048	619.238	620.779	0.5599
4th-grade reading score based on 4 intervals	617.343	619.031	620.966	0.6044
4th-grade reading score based on 6 intervals	617.467	618.795	620.341	0.6195

Note: This table reports the descriptive statistics for the averages of the simulated data in each variable and their properties. Gender is measured by an indicator taking the value 1 if the student is male and 0 otherwise. Free lunch takes the value 1, if a student is receiving a free or reduced-price lunch and 0 otherwise and small class takes the value 1, if the student was assigned to a small-sized class and 0 otherwise. White has the value 1, if the student is white and 0 otherwise.

student demographics of the simulated data are occurring twice in Table 6.1.

The student demographics of the math and the reading score model are almost the same since the identical distributions and generation process was used. The simulation generated an average of 50% female students across all data sets, 30% of students receiving a free or reduced-price meal and 80% of students who were white. Within the math data sets as well as the reading data set, an average of 24% small classes occurred, which was generally lower than in the actual data set due to the generation of doubled classes.

The scores were drawn from different distributions. Hence they do not have common values. The kindergarten math scores had a mean value of 485 points and the reading scores had a mean of 437 points. The minimum and maximum values, as well as the standard deviation, show that the mean scores did not deviate across the 100 generated data sets. The fourth grade score was simulated using the estimates of the regression in section 5.2, for the three interval sets with three, four and six intervals, and an additional error term which was randomly assigned. Therefore, three values were computed for the fourth grade scores, leading to similar values due to the same assumption.

Comparing the descriptive values of the averages over all simulated data sets with the actual data set, see Table 3.3, shows that the demographics have the same distribution in the data sets as in the data set of Project STAR. The mean values of the scores in kindergarten and fourth grade of the simulation correspond to the actual data set. The math scores in kindergarten had the mean 485 overall students and 500 if the math data set is considered. The simulation was executed assuming the mean overall students as the expected value; therefore, the simulation's mean lies around 485. The same applies to the reading scores in the actual data set. The mean in the reading data set is with 445 higher than the mean overall students, which is 437. In the simulation, an expected value using the latter one of both values was assumed and thus, the mean has a value of 437.

There were no such deviations between the overall data set and the subsets for the math and reading scores in fourth grade. The mean was for math in both data sets 715 and for reading 624. In the simulation, the means of these values are a little lower. In this case, the mean of the fourth grade math score is around 701 and the mean for the reading score is around 617.

6.3 Regression on Simulation Data

After the data generation of the compositional peer effect of interval sets of three, four and six equidistant intervals, as they were used in section 5.2, are computed. Afterwards, they are imputed using the algorithm `impRZilr`. Then one model for each interval set and each subject is fitted. This leads to a total of six models, which are fitted 100 times respectively. The average estimates of all regressions in each model are shown in Table 6.2.

The estimates in each iteration are stored and the compositional estimates are back-transformed into the original space and clr-transformed and stored as well in each iteration. Figure 6.1 shows the mean value of these compositional estimates, $b = \mathrm{ilr}^{-1}(\beta)$, after the clr transformation, i.e. $\bar{b}_i^{\mathrm{clr}} = E(\mathrm{clr}(b))$ for $i = 1, ..., n$ and n is the number of intervals. Additionally, their 95% confidence intervals are shown, which were computed by

$$\bar{b}_i^{\mathrm{clr}} \pm z_{0.975}\sigma(b_i), \tag{6.3}$$

where $z_{0.975}$ is the 97.5 quantile of the standard normal distribution and $\sigma(b_i)$ is the standard deviation yielded by the estimates across the whole simulation.

The distribution of these compositional terms shows the same behaviour as observed in section 5.2. The compositional reading terms are again relatively linear and the higher intervals lead to a decrease in the fourth grade score if the ratio of students increases in these intervals, while it increases if the ratio in the lower intervals is higher. In the math model, only the highest interval shows a negative influence on the response if the ratio in that interval increases, the other intervals have, as in the previous estimation, no influence or a positive influence on the response.

Therefore, it can be assumed that the regressions results found in section 5.2 were not sensitive to outliers. If that would have been the case, the average results observed in the simulation should differ greatly from the observed results of the actual data set. However, the simulation confirmed the previous findings.

Table 6.2: Average estimates of the non-compositional terms over each model

	Math Fourth Grade			Reading Fourth Grade		
	3 Intv.	*4 Intv.*	*6 Intv.*	*3 Intv.*	*4 Intv.*	*6 Intv.*
Intercept	685.13	686.78	689.33	606.29	609.86	608.95
Female	6.1044	5.9041	5.7918	4.8892	4.9391	5.0937
White	-2.2537	-2.1463	-2.3939	5.3503	5.1942	5.5119
Non Free Lunch	10.6668	10.7670	10.7000	4.4355	4.2834	4.3829
Small Class	0.558	0.5853	0.1427	4.5758	4.6466	4.8545
Class RE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
AIC	56326.06	56345.43	56351.01	58196.46	58192.27	58209.88
R^2	0.3460	0.3436	0.3454	0.3551	0.3554	0.3549

Note: The standard errors are added in parenthesis. The stars show the significance based on the p-value from very significant '***' with values below 0.001 to '**' with values below 0.01 and '*' with a p-value below 0.05. The R^2 value is the adjusted R^2 for the model. RE = Random Effects, FE = Fixed Effects

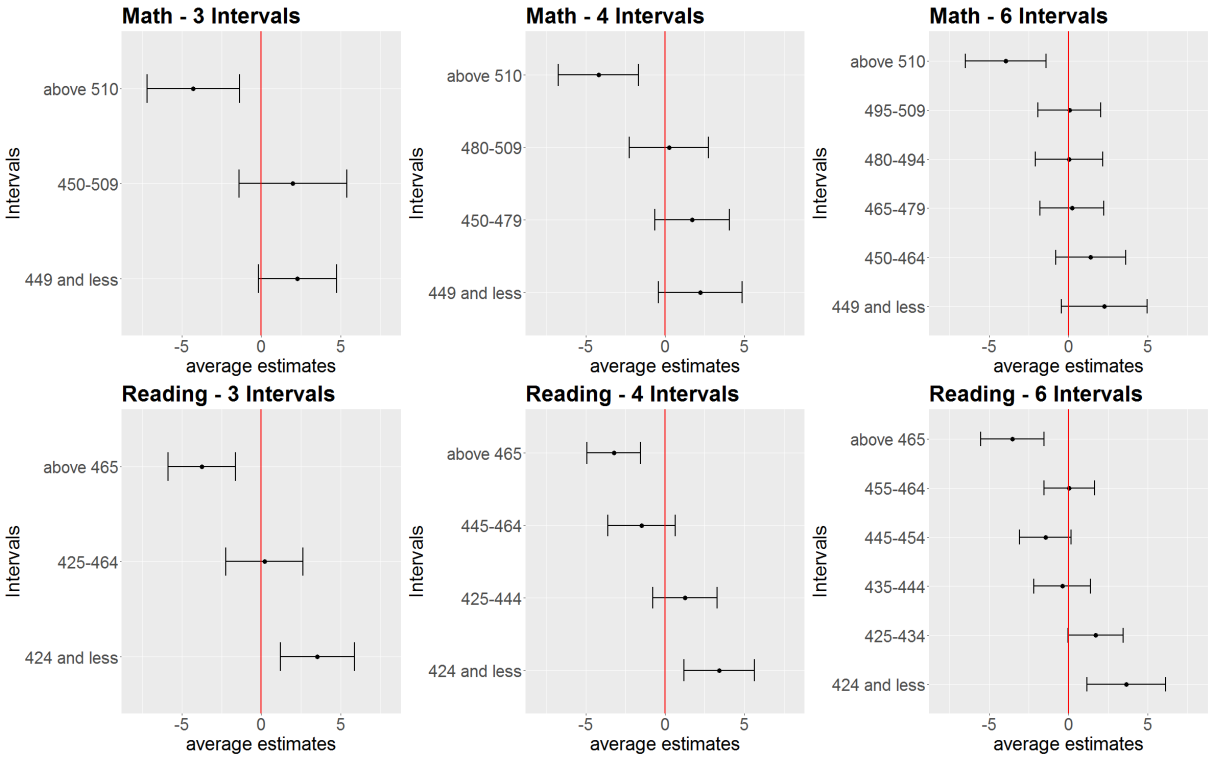

Figure 6.1: clr-transformed average estimate for the class composition with the 95% confidence interval using the standard deviation over the estimates

7 Conclusion

Considering a compositional peer effect offers new possibilities to analyse a class structure and the possible influence on an individual's score in the later school career. For compositional peer effects the ratio of students in a specific test scores is considered and the results are giving the impact of the relative distribution of the class on subsequent test scores of an individual. These estimates could be used as a guide for the future assignment of students into classrooms.

In the data set of Project STAR using a non-compositional peer effect like the mean net the individual's score is showing positive relation as it was also shown in previous studies by Krueger (1999) and Boozer and Cacciola (2001).

The composition based on the kindergarten test score shows a somewhat surprising result that the individual's test score in the later grade improves with a higher ratio of students scoring in the lower test scores and the scores of the individual decline if a higher ratio of high-scoring students is inside their class. In comparison to the non-compositional peer effect using the kindergarten scores, this is rather contradicting.

Using compositions for the peer effects based on a continuous variable showed that the selection does not change the results completely. Depending on the underlying data set, the choice of the intervals of the compositional variable can lead to an under- or overestimation. Therefore, the selection has to be carefully chosen to select the best fitting choice for the data. As a measure to find the best selection, the AIC or similar criteria can be used. Furthermore, it was observed in this data set that the model assessment for different choices of intervals, in general, does lead to the same variable selection of the best model.

The analysis of the dependency of covariates on the response for each class size respectively showed no large differences to the findings of the whole data set together. However, the terms of the non-compositional covariates had some deviations based on the class size, leading to the result that certain kinds of students were differently affected by the class size. The composition itself changed as well. However, it behaved accordingly to the previous whole model.

Applying the model checks provided by the `mgcv` package onto the model fits showed that there are issues with normality present in the math data set while there are no issues in the reading data set. The issues could not be solved with standardization of the scores and therefore, the results of the math data sets are violating the model assumptions. This means it does not fit itself well for the analysis of peer effects.

Taking a look at the impact of the kindergarten class distribution as composition on the entire school career up to eighth grade shows some spill-overs in each part and it is staying constant with a slight decline towards the higher grades, which is a wash-out effect. Thus, the kindergarten composition is leading to an increase of an individual's score in the consecutive year, if a higher ratio of lower-scoring students were in the same kindergarten classrooms and vice versa.

Comparing these long-term results for a composition of each year in Project STAR on the consecutive school years shows that after the first grade, this impact on the response is changed. If a higher ratio of good students is present within a class, then the individual scores higher in the tests of the following years. On the other hand, if a higher ratio of low-scoring students is contained in a class,

the individual's score decreases in the following years. This observation does correspond to the findings from our non-compositional peer effects. Hence, an essential change of the interpretation of the problem at hand occurred. It seems that the students were influenced positively by a higher ratio of peers who were high performing starting from the first grade. As this impact stayed constant for the subsequent compositions, a change could have occurred after the students started with their first grade, which also changed their interaction with peers and how this influenced their individual scores. Because no information regarding this is known from the experiment, it is impossible to find the change's actual cause.

Checking the sensitivity of the model estimation using a simulation showed that the results are close to the observations of the fit on the actual data and no essential effects which could be tracked back to outliers could be detected. Therefore, the simulation confirmed the previous analyses. Then the influence of the individual's classmates' performance in kindergarten positively influenced the subsequent test scores if a higher ratio of low-performing students were in their class. This could be because the students themselves were scoring low in the test and their scores improved much more than students in classes with a higher ratio of higher-performing students. However, in first class, the spill-over effect of the peers started to positively influence the peers if a higher ratio of high-performing peers were in the class.

To sum it up, it made sense to use the context of compositional data analysis to look into the effect of peers' performance based on the distribution of their score relatively across the class. This leads to the possibility to analyse the impact of the relative information on an individual's performance. This means it was possible to observe how a higher ratio of students having high or low scores affected an individual's score. In the next step, the information obtained by a compositional peer effect could be used to make predictions on the score of students in the future based on the current scores of the students and their class.

Further analyses of interest beyond this thesis could include zero-handling methods, which use the known data of the covariates and response together in an algorithm to impute the zero values. Furthermore, more imputation algorithms in the area of count and structural zeros are still needed to find a general approach for their imputation. In case of more methods being available in the future, an extension of the comparison of the impact of the zero handling methods presented in this thesis could be performed.

Analysing interaction in the above settings did not lead to any relevant results, so that regression results were not included. However, if a different setting and data set are considered and compositional peer effects are analysed, the interaction of compositional and non-compositional terms could be of interest. Moreover, an interpretation of such a mixed variable showing the interaction between relative and absolute information could be of interest.

References

Achilles, C.M., Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. *Tennessee's Student Teacher Achievement Ratio (STAR) project*. https://doi.org/10.7910/DVN/SIWH9F. Version V1. Last accessed on 06.10.2020 14:08. 2008. DOI: 10.7910/DVN/SIWH9F.

Aitchison, John. "Principles of Compositional Data Analysis". In: *Lecture Notes-Monograph Series* 24 (1994), pp. 73–81.

Aitchison, John. "The one-hour course in compositional data analysis or compositional data analysis is simple". In: *IAMG '97. Proceedings of the 3rd annual conference of the International Association for Mathematical Geology, Barcelona, Spain, September 22–27, 1997. In 2 vol*. Ed. by Vera Pawlowsky-Glahn. Barcelona: CIMNE, 1997, pp. 3–35.

Aitchison, John. "The Statistical Analysis of Compositional Data". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2 (1982), pp. 139–177.

Aitchison, John. *The Statistical Analysis of Compositional Data*. Chapman and Hall Ltd (reprinted 2003 with additional material by The Blackburn Press), 1986.

Aitchison, John and S.M. Shen. "Logistic-Normal Distributions: Some Properties and Uses". In: *Biometrika* 67.2 (1980), pp. 261–272.

Barceló-Vidal, Carles, Josep A. Martín-Fernández, and Glòria Mateu-Figueras. "Compositional differential calculus on the simplex". In: *Compositional Data Analysis: Theory and Applications*. Ed. by Vera Pawlowsky-Glahn and A. Buccianti. John Wiley and Sons, 2011, pp. 176–190.

Bietenbeck, Jan. "The long-term impacts of low-achieving childhood peers: Evidence from Project STAR". In: *Journal of the European Economic Association* 18.1 (2019), pp. 392–426.

Billheimer, Dean, Peter Guttorp, and William F. Fagan. "Statistical analysis and Interpretation of Discrete Compositional Data". In: *NRCSE Technical Report Series* 11 (1997).

Billheimer, Dean, Peter Guttorp, and William F. Fagan. "Statistical interpretation of species composition". In: *Journal of the American statistical Association* 96.456 (2001), pp. 1205–1214.

Boogaart, K. Gerald van den and Raimon Tolosana-Delgado. *Analyzing Compositional Data with R*. 1st ed. Springer-Verlag Berlin Heidelberg, 2013. DOI: 10.1007/978-3-642-36809-7.

Boogaart, K. Gerald van den, Raimon Tolosana-Delgado, and Matevz Bren. *compositions: Compositional Data Analysis*. R package version 2.0-1. 2020. URL: https://CRAN.R-project.org/package=compositions.

Boogaart, K. Gerald van den, Raimon Tolosana-Delgado, and Matevz Bren. "Concepts for handling zeroes and missing values in compositional data". In: *Proceedings of IAMG'06 — The XI annual conference of the International Association for Mathematical Geology*. Ed. by E. Pirard, A. Dassargues, and H. B. Havenith. University of Liége, Belgium, 2006.

Boozer, Micheal A. and Stephen E. Cacciola. "Inside the 'Black Box' of Project STAR: Estimation of Peer Effects using Experimental Data". In: *Yale University Economic Growth Center - Discussion Papers* 832 (2001).

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR". In: *The Quarterly Journal of the Economics* 126.4 (2011), pp. 1593–1660.

Egozcue, Juan Jose and Vera Pawlowsky-Glahn. “Basic Concepts and Procedures”. In: *Compositional Data Analysis: Theory and Applications*. Ed. by Vera Pawlowsky-Glahn and Antonella Buccianti. 1st ed. John Wiley and Sons, Ltd, 2011, pp. 12–27.

Egozcue, Juan Jose, Vera Pawlowsky-Glahn, Gloria Mateu-Figueras, and Carles Barcelo-Vidal. “Isometric Logratio Transformations for Compositional Data Analysis”. In: *Mathematical Geology* 35.3 (2003), pp. 279–300.

Filzmoser, Peter, Karel Hron, and Matthias Templ. *Applied Compositional Data Analysis*. 1st ed. Springer Nature Switzerland AG, 2018. DOI: 10.1007/978-3-319-96422-5.

Finn, Jeremy D., John Folger, and Deborah Cox. “Measuring Participation among Elementary Grade Students”. In: *Educational and Psychological Measurement* 51.2 (1991), pp. 393–402.

Fry, Jane M, Tim RL Fry, and Keith R McLaren. “Compositional data analysis and zeros in micro data”. In: *Applied Economics* 32.8 (2000), pp. 953–959.

Golsteyn, Bart H.H., Arjan Non, and Ulf Zölitz. “The Impact of Peer Personality on Academic Achievement”. In: *University of Zurich Working Paper* 269 (2017).

Harezlak, Jaroslaw, David Ruppert, and Matt P. Wand. “Generalized Additive Models”. In: *Semiparametric Regression with R*. Springer New York, 2018, pp. 71–128. DOI: 10.1007/978-1-4939-8853-2_3.

Hastie, Trevor and Robert Tibshirani. “Generalized Additive Models”. In: *Statistical Science* 1.3 (1986), pp. 297–310.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer New York, 2009. DOI: 10.1007/978-0-387-84858-7.

Hijazi, Rafiq. “An EM-Algorithm Based Method to Deal with Rounded Zeros in Compositional Data under Dirichlet Models”. In: *Proceedings of the 4th International Workshop on Compositional Data Analysis* (2011). Ed. by J.J. Egozcue, R. Tolosana-Delgado, and M.I Ortego.

Hron, Karel, Peter Filzmoser, and Matthias Templ. “Imputation of missing values for compositional data using classical and robust methods”. In: *Computational Statistics and Data Analysis* 54.12 (2010), pp. 3095–3107.

Krueger, Alan B. “Experimental Estimates of Education Production Functions”. In: *The Quarterly Journal of the Economics* 114.2 (1999), pp. 497–532.

Lavy, Victor, Olmo Silva, and Felix Weinhardt. “The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools”. In: *Journal of Labor Economics* 30.2 (2012), pp. 367–414.

Little, Roderick J.A. and Donald B. Rubin. *Statistical Analysis with Missing Data*. 2nd ed. John Wiley and Sons, Inc., 2002. DOI: 10.1002/9781119013563.

Lubbe, Sugnet, Peter Filzmoser, and Matthias Templ. “Comparison of zero replacement strategies for compositional data with large numbers of zeros”. In: *Chemometrics and Intelligent Laboratory Systems* 210 (2021), p. 104248.

Manski, Charles F. “Identification of Endogenous Social Effects: The Reflection Problem”. In: *The Review of Economic Studies* 60 (1993), pp. 531–542.

Martín-Fernández, Josep A., Carles Barceló-Vidal, and Vera Pawlowsky-Glahn. “Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation”. In: *Mathematical Geology* 35.3 (2003), pp. 253–278.

Martín-Fernández, Josep A., Karel Hron, Matthias Templ, Peter Filzmoser, and Javier Palarea-Albaladejo. “Bayesian-multiplicative treatment of count zeros in compositional data sets”. In: *Statistical Modelling* 15.2 (2015), pp. 134–158.

Martín-Fernández, Josep A., Karel Hron, Matthias Templ, Peter Filzmoser, and Javier Palarea-Albaladejo. “Model-based replacement of rounded zeros in compositional data: Classical and robust approaches”. In: *Computational Statistics and Data Analysis* 56.9 (2012), pp. 2688–2704. DOI: 10.1016/j.csda.2012.02.012.

Martín-Fernández, Josep A., Javier Palarea-Albaladejo, and Juan Gómez García. “Markov chain montecarlo method applied to rounding zeros of compositional data: first approach”. In: *Proceedings of the CODAWORK'03, 1st Compositional Data Analysis Workshop* (2003). Ed. by S Thió-Henestrosa and JA Martín-Fernández.

Martín-Fernández, Josep A., Javier Palarea-Albaladejo, and Ricardo Antonio Olea. “Dealing with Zeros”. In: *Compositional Data Analysis: Theory and Applications*. Ed. by Vera Pawlowsky-Glahn and A. Buccianti. John Wiley and Sons, 2011, pp. 43–57.

Mateu-Figueras, Glòria, Vera Pawlowsky-Glahn, and Juan Jose Egozcue. “The normal distribution in some constrained sample spaces”. In: *SORT (Statistics and Operations Research Transactions)* 37 (2013), pp. 29–56.

Murphy, Richard and Felix Weinhardt. “Top of the Class: The Importance of Ordinal Rank”. In: *The Review of Economic Studies* 87.6 (2020), pp. 2777–2826.

Narayanan, A. “Small sample properties of parameter estimation in the dirichlet distribution”. In: *Communications in Statistics - Simulation and Computation* 20.2-3 (1991), pp. 647–666.

Palarea-Albaladejo, Javier and Josep A. Martín-Fernández. “A modified EM alr-algorithm for replacing rounded zeros in compositional data sets”. In: *Computers and Geosciences* 34.8 (2008), pp. 902–917.

Palarea-Albaladejo, Javier and Josep A. Martín-Fernández. “zCompositions – R package for multivariate imputation of left-censored data under a compositional approach”. In: *Chemometrics and Intelligent Laboratory Systems* 143 (2015), pp. 85–96. URL: http://dx.doi.org/10.1016/j.chemolab.2015.02.019.

Palarea-Albaladejo, Javier, Josep A. Martín-Fernández, and Juan Gómez-García. “A Parametric Approach for Dealing with Compositional Rounded Zeros”. In: *Mathematical Geology* 39 (2007), pp. 625–645.

Pawlowsky-Glahn, Vera and Juan Jose Egozcue. “Geometric approach to statistical analysis on the simplex”. In: *Stochastic Environmental Research and Risk Assessment* 15 (2001), pp. 384–398.

Pawlowsky-Glahn, Vera, Juan Jose Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. 1st ed. John Wiley and Sons, Ltd, 2015.

Sojourner, Aaron. “Identification of Peer Effects with Missing Peer Data: Evidence from Project STAR”. In: *The Economic Journal* 123.569 (2013), pp. 574–605.

Templ, Matthias. “Artificial Neural Networks to Impute Rounded Zeros in Compositional Data”. In: *Advances in Compositional Data Analysis*. Ed. by Peter Filzmoser, Karel Hron, Josep A. Martín-Fernández, and Javier Palarea-Albaladejo. 1st ed. Springer Nature Switzerland AG, 2021, pp. 163–187.

Templ, Matthias, Karel Hron, and Peter Filzmoser. "`robCompositions`: an R-package for robust statistical analysis of compositional data". In: *Compositional Data Analysis: Theory and Applications*. Ed. by Vera Pawlowsky-Glahn and A. Buccianti. John Wiley and Sons, 2011, pp. 341–355.

Templ, Matthias, Karel Hron, Peter Filzmoser, and A. Gardlo. "Imputation of rounded zeros for high-dimensional compositional data". In: *Chemometrics and Intelligent Laboratory Systems* 155 (2016), pp. 183–190.

Verbelen, Roel, Katrien Antonio, and Gerda Claeskens. "Unravelling the predictive power of telematics data in car insurance pricing". In: *Applied Statistics* 67.5 (2018), pp. 1275–1304.

Whitmore, Diane M. "Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment". In: *American Economic Review* 95.2 (2005), pp. 199–203.

Whitmore, Diane M. and Alan B. Krueger. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR". In: *The Economic Journal* 111.468 (2001), pp. 1–28.

Whitmore Schanzenbach, Diane. "What Have Researchers Learned from Project STAR?" In: *Brookings Paper on Education Policy* 9 (2006/2007), pp. 205–228.

Wood, Simon N. "Fast stable direct fitting and smoothness selection for generalized additive models". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 70.3 (2008), pp. 495–518.

Wood, Simon N. *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. R package version 1.8-33. 2017. URL: `https://cran.r-project.org/package=mgcv`.

Wood, Simon N., Natalya Pya, and Benjamin Säfken. "Smoothing Parameter and Model Selection for General Smooth Models". In: *Journal of the American Statistical Association* 111.516 (2016), pp. 1548–1563.

Word, Elizabeth, John Johnston, Helen P. Bain, B. DeWayne Fulton, Jayne B. Zaharias, Charles M. Achilles, Martha N. Lintz, John Folger, and Carolyn Breda. *The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project - Technical Report 1985-1990*. 1990.

Appendix

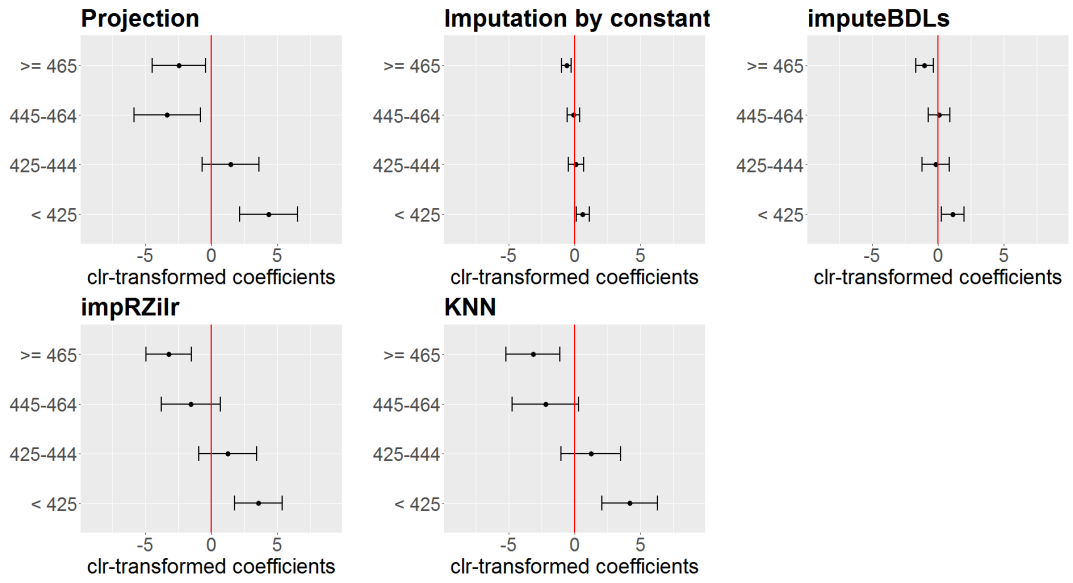

Figure 7.1: clr-transformed compositional estimates and their confidence bounds in the math score model for the composition of the kindergarten reading scores (divided into scores below 425, 425 to 444, 445 to 464 and above 465)

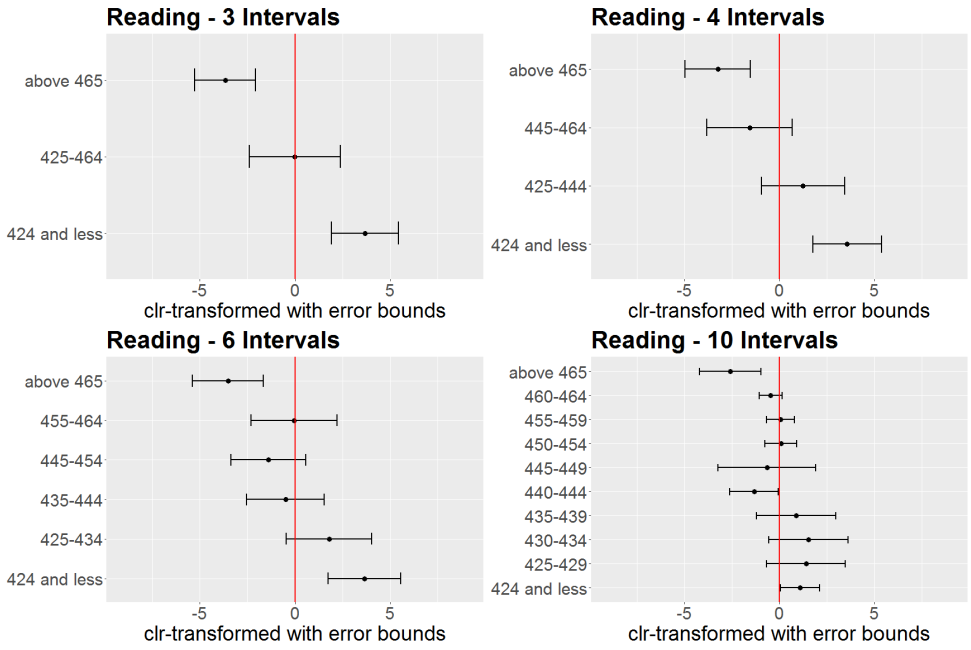

Figure 7.2: clr-transformed compositional coefficients with their confidence bounds for each interval selection of the reading score model

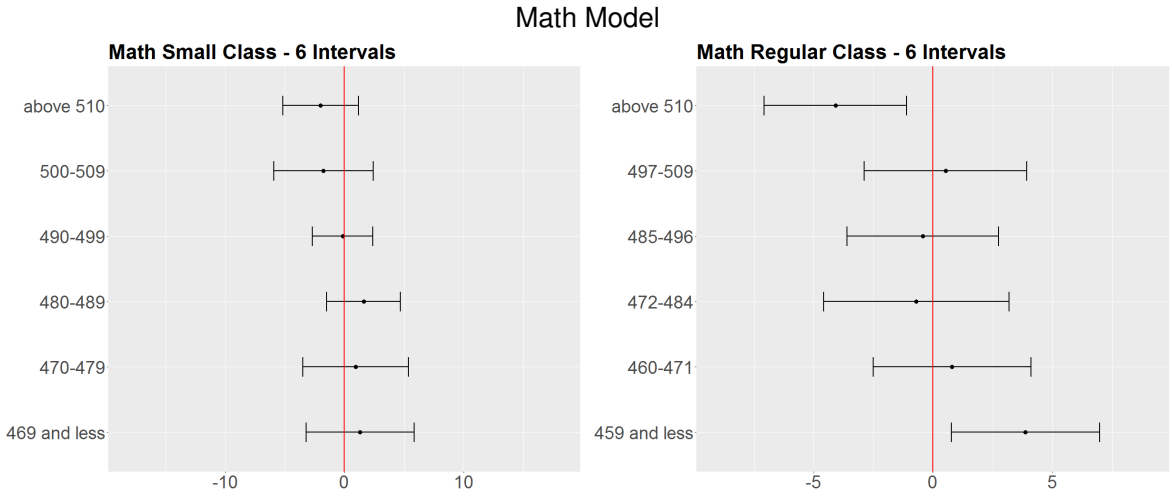

Figure 7.3: clr-transformed coefficients with their confidence bounds per class size for math test scores using six intervals as compositional terms

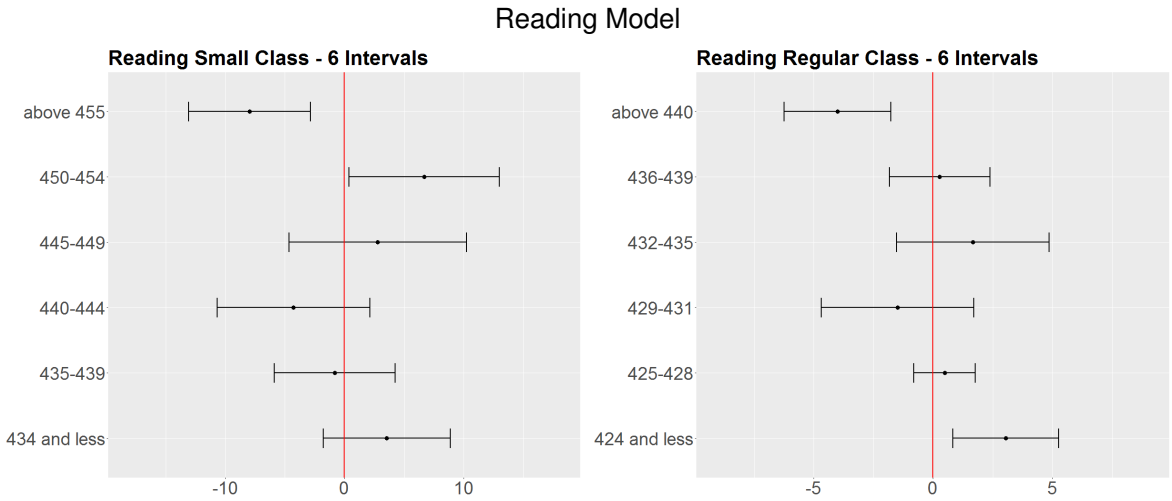

Figure 7.4: clr-transformed coefficients with their confidence bounds per class size for reading test scores using six intervals as compositional terms

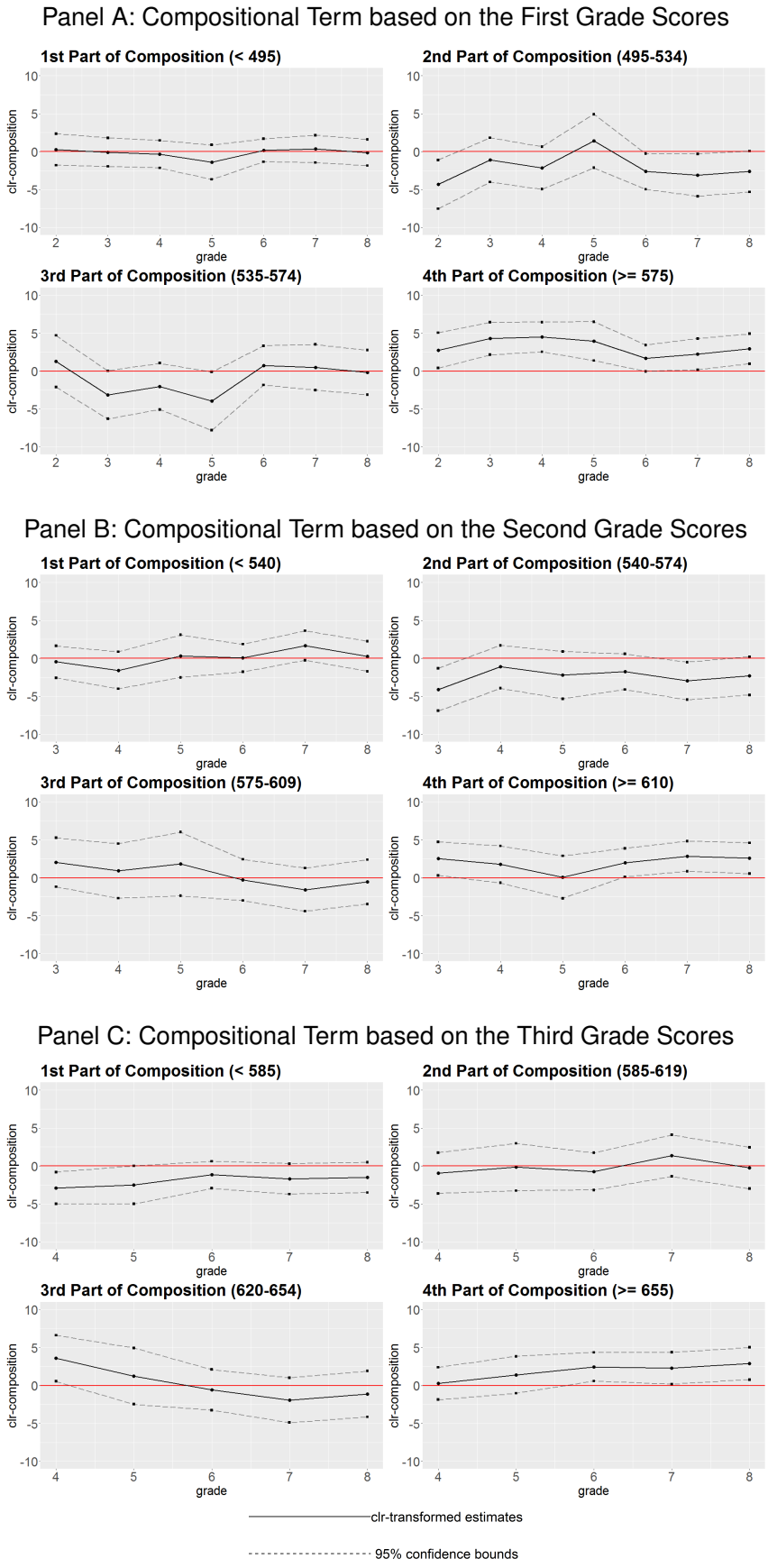

Figure 7.5: clr-transformed estimate and the confidence bounds for the class composition based on grades 1 to 3 data over all following grades on the math test scores