Evaluation of the Roots of Empathy program in Switzerland, years 2015 to 2017

David Cyrill Lätsch
Zurich University of Applied Sciences
30 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Madlaina Stauffer
Bern University of Applied Sciences
7 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Schweizerischer Fonds für Kinderschutzprojekte
Fonds Suisse pour des projets de protection de l'enfance
Fondo svizzero per progetti di protezione dell'infanzia

Evaluation of the Roots of Empathy program in Switzerland, years 2015 to 2017

Full Report

David Lätsch, Madlaina Stauffer & Mireille Bollinger

Bern University of Applied Sciences

Correspondence
Correspondence concerning this report should be addressed to David Lätsch, Bern University of Applied Sciences, Department of Social Work, Hallerstrasse 10, CH-3012 Bern, Switzerland. E-mail: david.laetsch@bfh.ch

Recommended citation for this publication
Lätsch, D., Stauffer, M., & Bollinger, M. (2017). Evaluation of the Roots of Empathy program in Switzerland, years 2015 to 2017. Full Report. Bern: Bern University of Applied Sciences.

Funding
This evaluation study was funded by the UBS Optimus Foundation on behalf of the Swiss Project Fund for Child Protection.

Conflict of interest
None of the researchers were in any way involved in the distribution of the Roots of Empathy program. There was no conflict of interest.

Acknowledgments
The authors would like to thank Miriam Stieger and Lisa Bayrami, representatives of Roots of Empathy, and Patricia Lannen and Sofia Milli, representatives of UBS Optimus Foundation, for their continuous support throughout the study, and Alexander Grossniklaus, Manuela Grieb, Rebecca Baumgartner, and Larissa Luchsinger for their invaluable assistance in collecting and organizing the data. Moreover, the authors wish to express their warm gratitude to the school principals, teachers and, most of all, students who participated in the data collection sessions or helped to make them possible.

December 2017

Abstract

Background

Professionals in the Swiss education system increasingly recognize that academic learning needs to be supplemented by incentives for social and emotional learning (SEL). One such program, Roots of Empathy, was recently implemented in several primary schools in the Swiss canton of Zurich.

Aims

The purpose of the study presented was to investigate whether Roots of Empathy was successful in bringing about its primary goals: an increase in empathy and prosocial behavior and a decrease in aggressive behavior for the pupils involved.

Method

A non-randomized matched-controls trial was conducted between fall of 2015 and summer of 2017. The study design included two cohorts consisting of 13 classes in the intervention group and 10 classes in the control group. Classes were matched on sociodemographic characteristics. The final sample included in both pretest and posttest measurements was 403 pupils (192 boys, 211 girls), with one-year follow-up measurements including 107 children. Intervention effects were calculated by analysis-of-variance and other statistical procedures. An additional qualitative inquiry into the implementation quality and impact of Roots of Empathy was carried out based on interviews and focus groups with teachers, instructors and children. Analyses were performed using thematic-content analysis.

Results

Significant effects were found in all three key outcome domains (empathy, aggression, and prosocial behavior) based on composite measures that combined self-reports and peer nominations from pupils and reports from teachers. Effect sizes for these measures ranged from 0.34 (for aggression) to 0.5 (for empathy), constituting an impact somewhat larger than typically observed for successful SEL programs. While teacher reports indicated moderate to large effects, measures collected from pupils did not. An additional analysis regarding several behavioral measures of prosocial behavior and altruism yielded mixed results. Gains in prosocial behavior were largely associated with gains in empathy, while the decrease in aggression was mostly due to other mechanisms unrelated to empathy. Follow-up analysis revealed that the effects for empathy and aggression were retained one year after completion of the program, while the effects for prosocial behavior dropped below the significance threshold. Qualitative analyses showed that teachers, instructors and children thought very highly of the core learning approach in Roots of Empathy (a parent and his/her baby visiting classrooms on repeated occasions). They suggest that more interactive learning approaches be adopted in the pre-visits and post-visits, however.

Conclusions

The Roots of Empathy program is a comparatively effective tool for increasing empathy and prosocial behavior and decreasing aggressive behavior among pupils in Swiss primary schools.

Contents

1 Introduction 9

1.1 Empathy and its Role in Social and Emotional Learning 9

1.2 The Roots of Empathy Program 10

1.3 Previous Research on Roots of Empathy 12

1.3 Objectives of the Current Study 13

2 Methods 14

2.1 Study Design 14

2.2 Quantitative Evaluation 15

2.2.1 Participants 15

2.2.2 Measures 17

2.2.3 Statistical Analysis 20

2.3 Qualitative Evaluation 20

2.3.1 Methods for Data Collection and Analysis.................................... 21

2.3.2 Participants 22

3 Results 24

3.1 Quantitative Evaluation 24

3.1.1 Implementation Fidelity 24

3.1.2 Empathy, Aggression, and Prosocial Behavior 25

3.1.3 Emotion Recognition 33

3.1.4 Altruism 35

3.2 Qualitative Evaluation 39

3.2.1 Implementation 39

3.2.2 Impact 48

4 Discussion 52

5 Conclusion 58

References 59

1 Introduction

In the current report, we present our evaluation of a three-year pilot implementation of Roots of Empathy, a social and emotional learning program designed for primary school. The program was carried out in the canton of Zurich, Switzerland, between autumn 2015 and summer 2017. The report begins with a brief introduction to the concept of empathy and its relationship with children's social behavior (in particular, prosocial behavior and aggression) in middle childhood. We will then give a closer of description of the Roots of Empathy program in Switzerland and the objectives of the present study, followed by an introduction to the methododology of the subsequent chapters.

1.1 Empathy and its Role in Social and Emotional Learning

Empathy may be described as a cognitive or emotional state that an individual experiences when he or she ascribes a cognitive or emotional state to another person. Importantly, while being empathic, the individual is aware that this experience or apprehension of the other person's emotion (or thought or intention) is a vicarious experience, i.e., that the emotion (or thought or intention) of the other person really "belongs to" or "originates in" the other person. In this regard, empathy is markedly different from mere emotional contagion; it is emotional contagion plus awareness of its origin in the other (for similar takes on the definition of empathy, see e.g. Batson, 2008; Bloom, 2016; Eisenberg, Shea, Carlo, & Knight, 1991; Eisenberg, Spinrad, & Morris, 2014). A good example of a full-blown empathic response is a child that perceives another child, notices that the other child is happy, and feels happy along with that other child while knowing that the pleasant experience at the origin of this happiness has happened to the other child, not to the individual him- or herself. It is important to distinguish between two—interrelated, yet distinctive—processes of empathy: a cognitive process, which consists of the intellectual recognition of another person's emotion (or thought or intention), and an affective process, which lies in the vicarious experience of the other person's emotional state. Cognitive empathy, in this sense, does not necessarily imply feeling what the other person feels, but simply recognizing what that feeling might be.

Research on "these things called empathy" (Batson, 2008), a choice of words that gives credit to empathy's multilayered nature, has soared over the last three decades (for milestones of empathy research, see e.g. Baron-Cohen, 2011; Batson, 1991; Eisenberg & Fabes, 1990; Hoffman, 2000). One important reason for this development may be that an individuals' capacities for empathy has been shown to predict socially and morally desirable behaviors such as spontaneously and voluntarily helping other people or sharing resources with them. Not everybody is enthusiastic about empathy's role in human morality: arguments against considering empathy as the only or even the most important driving force of human moral development have been brought forth repeatedly (e.g., Pinker, 2011). One author has argued that empathy, with its focus on needy individuals and its tendency to overlook the larger number of people, might even hinder sound moral judgement and fair decision-making (Bloom, 2016, 2017). But while these are intriguing arguments, the assumption that empathy makes a strong positive contribution to socially and morally desirable actions is not merely a flight of theoretical fancy—it is firmly grounded in empirical research. Studies suggest that more empathic children, adolescents and adults tend to be more prosocial and less aggressive or violent towards others (Batson, 1991; Hoffmann, 2000; Miller & Eisenberg, 1988), are often more sympathetic (Eisenberg et al., 2014) and altruistic (Van Lange, 2008), have a stronger sense of justice (Eisenberg & Morris, 2001; Hoffmann, 1994) and show more elaborate moral reasoning (Hoffman, 1994) than less empathic individuals. Also, aggressive behavior like bulling has been found to be associated with a lack of empathy (Jolliffe & Farrington, 2006). From a theoretical point of view, these findings are explained by the assumption that empathy enables individuals to take the perspective of others (thus allowing them to anticipate the consequences of one's actions on other's emotional states) and to make them feel the inherent value of other's emotional lives (thus motivating individuals to avoid acts that harm other persons) (e.g., Segal, 2011). In the latter sense, empathy is seen as a prerequiste for sympathy (sometimes also called "empathic concern"), which is in turn assumed to motivate individuals (in conjunction with other phenomena, such as a an intellectual sense of a moral self) towards acting prosocially.

Over the course of the last few decades, practitioners in the fields of child care and education have begun to draw from this idea. Legislators, school principals, teachers, and other professionals in the education system increasingly recognize that academic learning must not be the only goal of educational programming but needs to be supplemented by incentives for social and emotional learning (SEL) as well. Promoting emotional and social learning in children is not only considered as a means to foster children's emotional self-regulation, well-being, and social responsibility in the long run, but also as a prerequisite for positive attitudes towards school and thus for learning motivation and academic success (cf. Elias et al., 1997). In a recent meta-analysis covering over 270,000 children who had participated in an SEL program, Durlak and colleagues (2011) found that SEL activities were often fairly successful in achieving such results, with average effect sizes in the 0.2 to 0.3 range. With regard to the Swiss context, methodologically rigorous work on the impact of SEL programs has been rare, and the few high-quality studies that do exist suggest that the evaluated programs are mostly ineffective (e.g., Eisner & Ribeaud, 2005; Averdijk, Zirk-Sadowski, Ribeaud, & Eisner, 2016; Malti, Ribeaud, & Eisner, 2011).

1.2 The Roots of Empathy Program

Within the broader movement of social and emotional learning, the concept of empathy has been at the forefront from the very beginning (e.g., see Feshbach, 1979; for a recent overview, Feshbach & Feshbach, 2011). One approach that posits particularly strong links between empathy and social and emotional learning is Roots of Empathy. The program originated in Canada in 1996 and has since been exported to more than ten countries on three continents (for introductions and overviews, see Bayrami, 2017; Gordon, 2001, 2003, 2007; Gordon & Green, 2008). In the year 2012, Roots of Empathy was for the first time brought to a country in continental Europe, Germany.

Roots of Empathy is an evidence-based social and emotional learning program for kindergarten up to the 8th grade. Its main objectives are to increase prosocial behavior and to reduce aggressive behavior of pupils, while fostering the development of emotional literacy and empathy and raising emotional and social competence. Additionally, the program aims at imparting knowledge of human development and dealing with infants. To use a catch phrase, Roots of Empathy is based on the idea that empathy is "the best peace pill we have" (Gordon & Green, 2008), an assumption derived from the personal experiences of its creator, Mary Gordon, as well as from empirical evidence, some of which was summarized in the preceding section.

Roots of Empathy uses a broad concept of empathy, which may be divided into four different facets: i.) understanding one's own emotions (emotional literacy), ii.) understanding other's emotions (cognitive empathy), iii.) being emotionally responsive to others (affective empathy), and iv.) caring about other's emotions (strive for inclusion, kindness). According to assumptions made in Roots of Empathy literature (Gordon, 2007), the fundaments of empathy exist already at birth and develop to a large degree during the first years of life, driven by experiences in interaction with the environment (especially interaction with parents). But while this time span is thought to be of critical importance, experience may modify a person's empathic responses and capabilities at any time across the whole lifespan. This is where Roots of Empathy comes in: The program aims at imparting experiences to children that will positively stimulate their empathy and, in this way, increase their prosocial behavior and reduce their aggression. Roots of Empathy presumes six strands of human connection that are fundamental to understanding the "social cohabitation" (from understanding our self to understand our society). These strands are the building blocks of the program and may be summarized as follows (based on Gordon, 2007):

- Strand 1, Neuroscience: In the first years of life, human brain development is driven by (social) experiences that influence social interaction for the whole life.
- Strand 2, Temperament: People are different in the way they feel, act and express themselves, depending on their temperament. The ability to accept this and to see things from the point of view of others forms a basis for empathy.
- Strand 3, Attachment: The strength and reliability of early relationships influence emotional development.
- Strand 4, Emotional Literacy: To form a healthy sense of self and intimate relationships, awareness of one's own and others' emotions and the ability to understand and express emotions are needed.
- Strand 5, Authentic Communication: All people have (negative) emotions. An authentic communication helps to deal with them and makes it possible to understand each other better.
- Strand 6: Social Inclusion: Notwithstanding the differences between people, everybody has a need and a right to be an integrate part of society.

Roots of Empathy is intended to impart an understanding of the importance of these six strands to the participating children and thereby to open a way for fostering empathy (Gordon, 2007). The program's curriculum is highly structured. For a period of 9 months, starting in fall, the pupils take part in 27 lessons grouped in 9 themes. These themes are labelled as follows: 1. Meeting the Baby, 2. Crying, 3. Caring and Planning for the Baby, 4. Emotions, 5. Sleep, 6. Safety, 7. Communication, 8. Who am I? 9. Goodbye and good wishes. Each theme consists of a pre-visit, a family visit and a post-visit. The core of the program is the family visit, in which a parent and his/her baby attend the class and thereby enable children to learn experientially. The fact that babies express their feelings and demand immediate responses is supposed to give children the opportunity to learn to recognize emotions in others (Strand 4: Emotional Literacy) and observe empathic reactions directly (Strand 1: Neuroscience; Strand 3: Attachment). The children are encouraged to verbalize their observations during the family visits, which is in turn aimed at fostering their emotional literacy (Strand 4: Emotional Literacy). Additionally, in these lessons the pupils learn about infant development and effective parenting practices (Strand 1: Neuroscience). The experiences during the family visits serve as basis for discussions about one's own emotions (Strand 4: Emotional Literacy), perspective-taking (Strand 2: Temperament), and caring for others. According to the curriculum, the pre-visits and post-visits include preparing and post-processing the family visits, discussions, and various classroom activities (e.g., reading books together, journal writing or artwork such as designing a collage). The objectives of these lessons are to reflect on feelings and the feelings of others, to foster emotional understanding, social skills (like consensus-building or negotiating) and problem-solving skills. During all the lessons the instructor is required to guarantee an open-minded atmosphere, enabling the pupils to feel welcome and value the way they are (Strand 6: Social Inclusion). Additionally, the instructor and the parent speak about their own emotions, which is intended to model authentic communication (Strand 5: Authentic Communication). With all of this, Roots of Empathy is intended to put the affective dimension center-stage, something that is seen as overlooked and neglected in traditional education (Gordan, 2007).

Taking into account the developmental age and the interest of the pupils, the extensive and detailed Roots of Empathy curriculum is available in four different variations (kindergarten, 1st to 3rd grade, 4th to 6th grade, and 7th and 8th grade). All lessons are given by a trained and certified Roots of Empathy instructor, who works closely together with the participating family and the teacher. The instructors have different professional backgrounds, such as social work, early childhood education or school education.

The implementation of ROE in Switzerland began in 2014, with the outlook of a three-year pilot phase that would last until 2017. The implementation was limited to schools in the German-speaking canton of Zurich, which has the largest population of all Swiss cantons and contains Switzerland's largest municipality, the city of Zurich. In Switzerland, the implementation of ROE was managed by a local representative, who was working closely with Roots of Empathy International Office in Toronto, Canada. The Swiss instructors had been trained for the implementation in Switzerland. During the implementation, they all had access to a Mentor in Canada. This Mentor visited Switzerland once a year, providing on-site supervision and feedback to every instructor. Whenever they experienced challenges, instructors had the opportunity to contact the Mentor in writing or via telephone. They could also consult with the Roots of Empathy representative in Switzerland.

1.3 Previous Research on Roots of Empathy

Previous research on the effectiveness of the Roots of Empathy (henceforth abbreviated as: ROE) program has provided evidence that the intervention works, although the patterns of results have not been entirely consistent. For example, Schonert-Reichl, Smith, Zaidman-Zait, and Hertzman (2012) found that, after completion of the program, children who had participated in ROE behaved more prosocially, as rated by their peers, than children in the control group condition. Moreover, based on teachers' reports, ROE participants showed decreased levels of proactive and relational aggression and a better understanding of the causes of infants' crying, with effect sizes ranging from small all the way up to a Cohen's *d* of 0.79. Somewhat surprisingly, however, there were no significant differences in this study between ROE and control groups in terms of cognitive and affective empathy. This suggests that ROE has the potential to promote prosocial behavior and to inhibit aggressive behavior by means that are not necessarily connected to increasing empathy. In another study, evaluating a ROE program carried out in Manitoba, Canada, Santos, Chartier, Whalen, Chateau, and Boyd (2011) found that ROE was successful in significantly reducing children's levels of physical and indirect aggression and in increasing their prosocial behavior, although effect sizes were generally small (ranging from 0.08 to 0.26 for significant between-group differences) and more pronounced for teachers' ratings than for pupils' self-reports. In a Scottish evaluation of ROE, MacDonald and colleagues (2014) found positive effects of the program in almost all of the outcome variables they looked into: In comparison with pupils from control groups, ROE participants showed increased cognitive and emotional empathy, increased prosocial behavior, better anger management, enhanced emotional well-being, and reduced rates of conduct problems, problems with peers, and hyperactivity. Also, ROE participants showed a better understanding of infant development than their peers in the control group. Increases in prosocial behavior were more pronounced for boys than for girls. Finally, in a more recent study from Scotland, Wrigley, Makara and Elliot (2015) found confirming evidence that the program's impact in terms of prosocial behavior was mediated by gains made in empathy.

1.4 Objectives of the Current Study

Taken together, previous studies suggest that ROE has a history of achieving many of its goals. It is as yet unclear, however, whether these effects transfer to different environments, particularly those of Western Europe, where school cultures, educational systems and the sociodemographic compositions of classrooms, among other things, differ from the environments mentioned above. The purpose of the currenty study was to evaluate the pilot implementation of ROE in Switzerland in an objective, scientifically robust research design. More specifically, the objectives of the study were (i) to investigate whether the Roots of Empathy program in Switzerland would bring about significant increases in empathy and prosocial behavior and significant decreases in aggressive behaviors for the pupils involved in the program, (ii), if such effects exist, to investigate the causal pathways that mediate the program's effectiveness, and (iii) to find out in what ways the program might be further improved and adapted to the Swiss context in the future, taking into consideration the perceptions and recommendation of pupils, teachers and ROE instructors involved in the implementation.

2 Methods

2.1 Study Design

A cluster non-randomized matched-controls trial was conducted between autumn of 2015 and summer of 2017. The study design included an intervention group comprised of 13 classes which each participated in the year-long Roots of Empathy program and 10 control classes which did not participate. All classes were drawn from public primary schools in the canton of Zurich, ranging from 3rd to 6th grade. They were grouped in two cohorts: 4 of the intervention classes and 4 of the control classes were followed in years 2015 to 2016, and 9 ROE classes and 6 control classes in years 2016 to 2017. ROE classes were only eligible for participation in the study if the instructor carrying out the program had a minimal working experience with ROE of one year. Data collection began in the fall 2015, with the beginning of the new school-year in Switzerland and one year after the first classes had enrolled and participated in ROE in the canton of Zurich. This one year waiting period was deliberately put in place because it was expected that the first year of the implementation would be subject to certain initial learning challenges or “teething problems” which are usually associated with the start-up of a new social and emotional learning program. It was assumed that these, should they exist, would not be representative of future implementations and therefore should be exempt from the evaluation.

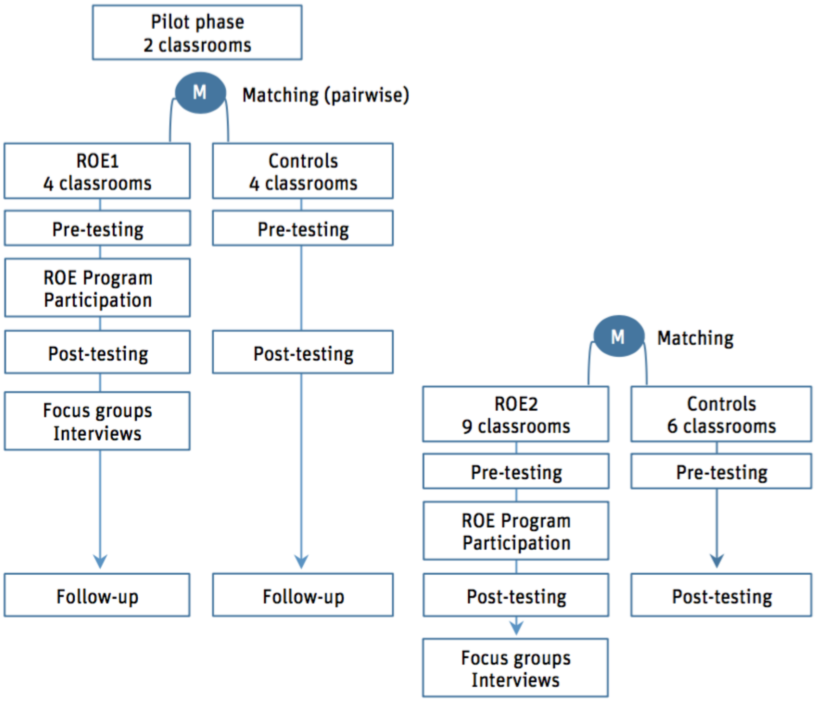
Both cohorts were tested at the beginning of the school-year before the ROE program began and once again at the end of the school-year when the program had been completed (Figure 1). In cohort 1 (years 2015 to 2016), there was an additional one-year follow-up measurement in the summer of 2017 to test for enduring effects. A methodologically preferrable randomized allocation of classes to the two study groups was not possible for practical reasons: there wasn’t a sufficient number of school principals interested in implementing the ROE program who would also have been willing to partake in the randomization procedure. To compensate for the absence of randomization, a matching procedure was applied: For each class enlisted in the ROE program, we sought to recruit a control class that was matched on key socioedemographic variables (such as geographical location, school size, socioeconomic resources in the community, composition of students in terms of sociodemographic characteristics) as closely as possible, resulting in a cluster matched-pairs design. We also considered teacher characteristics (such as age, gender, working experience, attitudes towards social and emotional learning programs, attitudes towards the importance of empathy) in the matching procedure. Control classes were only included in the study if they did not participate in an alternative SEL program during the course of the study. To prevent contamination or spill-over effects, control classes were recruited from different schools than ROE classes. Two to four weeks before pre-testing, parents were informed about the oncoming data collection in a letter by mail and were given the opportunity to refuse their consent if they did not want their children to participate in the study.

Pretest, posttest and follow-up data were collected in classrooms during a three-hour session each. Children were informed about the study by the teacher in advance and by the research team on occasion of the data collection session on site. Children were then explicitly asked whether they wished to participate. If they declined, they were given a surrogate task prepared by the teacher in advance. Measurement tools included questionnaires for pupils and teachers, several video clips to be watched and rated by pupils, and a game where pupils could make a small amount of money (approximately USD 5 at most) by playing with (or against) each other. Two members of the research team instructed the children throughout the session and answered their questions.

The study design also included several focus groups and interviews with teachers, ROE instructors, and a subsample of pupils. These were conducted at the end of the school-year after the program had been recently completed, on occasions separate from the classroom sessions. Their purpose was to provide more in-depth information about the program and its effects and about possible avenues for improvement in the future.

All data collection tools were piloted with two classes several weeks before the first pre-testing sessions, and several revisions to the collection tools were made subsequently. These pilot classes were not included in the study.

Figure 1: Flow of cohorts through the study design



2.2 Quantitative Evaluation

2.2.1 Participants

Based on an assessment of extant research on SEL programs in classroom environments (see chapter 1), we expected that the effects of the intervention, if such exist, would likely be in the small to moderate range. Accordingly, in order to be able to detect a small effect of 0.2 standard deviations with 80% power at the significance level of 5%, we calculated that the appropriate N would be 392 participants, or 146 in each study group. Classes in primary schools in the canton of Zurich typically consist of approximately 20 pupils. Because the total number of ROE classes that met eligibility criteria in years to 2015 to 2017 was only 13—and taking into consideration that not all of the pupils could be expected to participate in the complete set of data collection sessions—we decided to include (if possible) all eligible ROE classes in the study. Power calculations suggested that the control group needed to include at least 8 classes. The total sample at pre-testing (13 ROE classes, 10 control classes) surpassed this target with an N of 471 pupils (*n* ROE = 267, *n* control = 193). The number of pupils per class was equally distributed across the two study groups (*m* ROE = 21.0, *SD* = 2.52; *m* control = 22.0, *SD* = 1.41; p = 0.295). The distribution of other sociodemographic characteristics in the two groups is shown in Table 1, along with the p values indicating that none of these differences were statistically significant. Although the matching procedure did not result in perfectly equal compositions, it was altogether successful in balancing the sociodemographic characteristics between the two groups.

Table 1: Child, family, teacher and classroom characteristics, by study group

	Control	ROE	p value
Child characteristics			
Age, mean (s.d.)	10.35 (1.11)	10.56 (1.00)	.051^{ns}
Female gender, %	47.1	54.1	.179^{ns}
Subjective SES, mean (s.d.)	3.19 (0.43)	3.22 (0.38)	.411^{ns}
Nationality, %			.162^{ns}
Swiss	76.7	68.0	
non-Swiss, European	17.6	20.6	
non-European	1.9	2.6	
don't know	3.8	8.8	
Child migrated to Switzerland, %	10.1	16.7	.065^{ns}
Religious denomination, %			.630^{ns}
Christian	57.1	50.6	
Muslim	19.9	25.5	
other	7.1	5.6	
none	11.5	13.4	
don't know	4.5	4.8	
Family characteristics			
Family type			.148^{ns}
Child lives with both parents	87.3	85.0	
Child lives with single parent (with/without partner)	11.4	15.0	
other	1.3	0.0	
Nationality of mother, %			.633^{ns}
Swiss	52.5	50.3	
non-Swiss, European	32.9	37.8	
non-European	13.9	11.9	
don't know	0.6	0.0	
Classroom and teacher (N=22) characteristics			
Age of teacher, mean (s.d.)	37.29 (9.12)	34.50 (9.80)	.530^{ns}
Female teacher, %	88.9	75.0	.422^{ns}
Professional experience as teacher in years, mean (s.d.)	7.78 (5.92)	7.00 (6.91)	.790^{ns}

P values are from independent samples T tests (means) or Chi Square tests (percentages), both two-sided

Of the 471 pupils in the original target sample, 38 could not participate in the pretest either because their parents had refused their consent (*n* = 12, 2.5%) or because the pupils were absent on the day of data collection due to illness or other unforeseen reasons (*n* = 26, 5.5%). Of the remaining 433 children, only one opted not to participate. This cooperation rate of 99.8% resulted in a final sample size of 432 pupils at pretest (T0). Of these 432 children, 402 could be tested at the end of the school-year (T1) again. 30 children were absent because of illness, for other temporary reasons or because they had moved away from the school. Pupils who had not been able to participate at pretest but were present at posttest participated in the data collection at T1, but their data were excluded from the analysis of intervention effects (no missing values were imputed). Attrition was low, with 93.1% of the sample at T0 remaining in the study at T1. The sample size for the follow-up measurement one-year after the completion of ROE in cohort 1 will be reported alongside the results further below.

2.2.2 Measures

The items of all measurement tools were given to the respondents in the German language only. In some cases, authorized German translations of the measurement tools could be used. In others, we applied a careful translation procedure where several members of the research team translated the items independently from each other, including back-translations from German into English. Translations were compared and revised until a consensus was reached.

Sociodemographics

The sociodemographic data we collected from children included month and year of birth, gender, family type (based on people with whom the children lived in the same household), nationality, religion, migration experience, nationality of parents, parents' profession, number, age and gender of siblings, and the children's subjective assessments of their family's socioeconomic status. Data from teachers included month and year of birth, gender, professional experience and prior involvement in any social and emotional learning program. We also asked teachers whether their class or individual students had ever been (or were presently) involved in any such program.

Self and other reports

The key outcome variables collected in our study broadly fall into four domains: (i) children's capacity for affective and cognitive empathy, (ii) children's aggressive behavior and the extent to which they are subjected to the aggression of their classmates, (iii) children's prosocial behaviors such as helping or sharing with another, and (iv) children's social and emotional well-being. These variables were drawn from three sources: children's self-reports, children's peer-nominations and teachers' reports on the children.

Empathy. Empathy scales for students' self-reports were taken from the "Basic Empathy Scale" by Jolliffe and Farrington (2006). The questionnaire consists of 24 items with two subscales (cognitive and affective empathy). Items are rated on a 5-point Likert format ranging from "strongly agree" to "strongly disagree". Examples are "After being with a friend who is sad about something, I usually feel sad" (affective empathy) or "I can usually work out when people are cheerful" (cognitive empathy). Teachers rated their students' level for empathy on the 5 items of the "Teachers' Reports of Children's Empathy/Sympathy" questionnaire developed by Zhou, Valiente and Eisenberg (2003). We transformed the response format from the original scale into a 5-point Likert format ranging from “not at all” to “very much.” Items include “This child often feels sorry for others who are less fortunare” or “This child gets upset when she/he sees another child being hurt.” To obtain children’s peer nominations on empathy, they were given a complete list with the names of their classmates and were then asked to circle the names of every child that “often shows empathy with other children in your class.” The German word that was used here was “Mitgefühl,” which is a more colloquial, everyday expression than the English term “empathy” and is generally well understood at age 8 and above (this had been established in pilot testing).

Aggression. Students’ self-reports on aggressive behavior were based on the “Reactive and Proactive Aggression Measure” by Little, Henrich, Jones, & Hawley (2003). This 24-item questionnaire contains a 2x2 subscale matrix: overt vs. relational aggression and proactive (instrumental) vs. reactive aggression, resulting in four subscales altogether. Items were rated on a 5-point Likert format ranging from “never” to “very often”. Examples include “If others have threatened me, I say mean things about them” (“relational reactive aggression” subscale) or “I hit, kick, or punch others to get what I want” (“overt proactive aggression”). In addition to rating their own aggressive behaviors, students also assessed the extent to which they were the victims of others’ aggressions, responding to the “Bullying Victim” scale of the “Peer Interactions Primary School Questionnaire” developed by Tarshis and Huffman (2007). Items were rated on a 5-point format ranging from “never” to “very often”. Examples include “Other students take things from me that I do not want to give them” or “Other students look at me in a mean way”. Teachers rated their pupils’ aggressive behavior on the “Reactive/Proactive Aggression—Fasttrack Teacher Checklist” (Dodge & Coie, 1987). The 6 items of this tool were assessed on a 5-point format; we added three items of our own to include more subtle distinctions in the domain of relational and emotional aggression. Finally, children’s peer nominations were collected in way similar to the ones on empathy, with three different items discriminating between physical, emotional and relational forms of aggression. These three peer nomination items could be aggregated into a single peer-nomination aggression index.

Prosocial behavior. To assess children’s prosocial behavior, they were given the 4 items of the “Prosocial Behavior” subscale from the Children’s Social Behavior Scale—Self Report” developed by Crick and Grotpeter (1995). Again, items were rated on a 5-point format ranging from “never” to “very often”. Examples for this scale are “Some kids try to cheer up other kids who feel upset or sad. How often do you do this?” or “Some kids help out other kids when they need it. How often do you do this?” Teachers assessed their students’ prosocial behavior using the “Prosocial Behavior” subscale of the “Strength and Difficulties Questionnaire” (Goodman, 1997). The 5 items of this scale could be rated on a 3-point format choosing between “not true,” “somewhat true” and “certainly true”. Examples are “This child often volunteers to help others” or “This child readily shares with other children.” The peer-nomination scale, constructed in the same way as described above, contained two items, one pertaining to “helping”, the other to “sharing.”

With regard to the outcome domains of empathy, aggression, and prosocial behavior, the data collected from pupils’ self-reports, peer-nominations and teachers’ reports lent themselves to the construction of composite (aggregate) measures. Such measures combine ratings from different sources into a single scale. The composition of such scales is often recommended in the literature because, if constructed well, they have the potential to substantially reduce the error variance associated with single data sources such as self-reports only (e.g., Kagan, 2013; van Dulmen & Egeland, 2011; Van der Ende, 1999). In the present study, we constructed aggregate measures by standardizing each individual scale so that each would have a minimal value of 0 and a maximal value of 10, using the proportion of maximum scaling (POMS) method (cf. Moeller, 2015). Composite scales were then calculated as the mean of these three standardized measures.

Social and emotional functioning. Self-report items on pupils' social and emotional wellbeing were taken from the distress subscale of the "Weinberger Adjusmtent Scale" (Weinberger & Schwartz, 1990). The subscale consists of 12 items which may be rated from "false" to "true" or from "never" to "always" on a 5-point Likert format. Examples include "I really don't like myself very much" (inversely coded) or "I usually think of myself as a happy person." In addition, self-esteem was assessed using the "Hare Area-Specific Self-Esteem Scale" (Shoemaker, 1980). The 10 items of this questionnaire are rated on a 4-point Likert format ranging from "strongly disagree" to "strongly agree." Example include "I am not as popular as other people my age" (inversely coded) or "My parents are proud of the kind of person I am". To assess children's level of social and emotional functioning from the teachers' perspective, the teacher version of the "Strengths and Difficulties Questionnaire" (Goodman, 1997) was used. The 20 items of this measure are rated on 3-point format and are divided equally between four scales, covering self-reported emotional symptoms, peer relationship problems, conduct problems and hyperactivity/inattention. The former two subscales may be combined into an "internalizing" score, the latter two into an "externalizing" score. All four subscale may be aggregated into a sum score that indicates the overall level of adjustment. Examples for items are "This child has many worries or often seems worried" (emotional symptoms), "steals from home, school or elsewhere" (conduct problems), "is constantly fidgeting or squirming" (hyperactivity/inattention) or "is generally liked by other youth" (peer problems, inverse coding).

Social desirability

To detect children's tendencies towards distorting their responses in the direction of social desirability, we included a subset of four items from the "Children's Social Desirability Short Scale". These could be rated as either "yes" or "no". Items included "Have you ever felt like saying unkind things to a person?" or "Sometimes, do you do things you've been told not to do?". Each "no" answer was added to a total desirability bias score ranging between 0 and 4, with higher scores indicating stronger tendencies toward distorted responses.

Implementation fidelity

To measure the quality of the implementation of the ROE program, we used instructor questionnaires that had been developed for previous research on the program. Instructors responded to a short questionnaire for each of the 27 lessons that are part of ROE, indicating, among other things, the date of the lesson, the materials provided by the program that they used or did not use in implementing the lesson, and how engaged they had perceived students and teachers to be during the lesson.

Behavorial measures

Apart from the tools described in the preceding paragraphs, we incorporated three behavioral measures in the study: recognition of facial expressions of emotions, altruism in a decision-making task called the Trust Game, and willingness to make a small charitable donation to others in need. These behavioral measures will be introduced in detail in the corresponding sections that report on the results (chapter 3).

2.2.3 Statistical analysis

All analyses were pre-specified. Before beginning our analyses, we inspected the dataset for patterns of inconsistent responses at the individual case level at pre-testing, post-testing or follow-up, using pairs of items that showed particularly high correlations and calculating difference scores between these items. We then set a threshold for these difference scores; values at or above this threshold were taken to suggest that children had either not understood the questions or had answered them erratically for other reasons. These cases (n = 15, 3.2%) were excluded from all subsequent analyses, as were those of children whose reponses suggested a strong tendency towards social desirability (n = 24, 5.1%). Following this, baseline differences in outcome variables at pretesting were investigated using independent sample *t*-tests after normality and equality of variances had been established, additionally checking the results against those of Mann-Whitney-Wilcoxon *u*-tests in the few cases where the normality assumption was violated (cf. De Winter & Dodou, 2010). To test for significant between-group differences in outcome variables, we considered both multi-level analyses (which take into account that students' individual data are nested within classes) and more conventional analysis of variance (ANOVA) and analysis of covariance (ANCOVA) procedures. The decision between these strategies was pre-specified as dependent on the extent of intracluster correlations (ICCs) at the classroom level. After examining ICCs and finding that the portion of variance bound at the classroom level was not substantial enough to require a multilevel approach, we decided to analyze the data performing traditional ANOVAs on the difference scores between posttest and pretest values, using age and gender (and, in some cases, additional variables) as covariates. Analysis of covariance using posttest scores as dependents and pretest scores as covariates were additionally performed to check whether this would yield meaningful differences in terms of statistical significance and effect sizes (cf. Thomas & Zumbo, 2011). Multifactorial analyses were performed to test for significant interaction between treatment condition and several variables such as gender, grade and baselines values in dependent variables. Because of the relatively large number of outcome variables considered in our analysis, the likelihood of a type I error (i.e., the false rejection of the null hypothesis) rises considerably. This means that finding any effect only slightly below the 5% significance level, particularly if it is seen as inconsistent with the general pattern of results, should be treated with caution. We confront this risk not by statistical correction but by reporting the exact p values for each individual model below, taking up potentially ambiguous cases in the discussion.

2.3 Qualitative Evaluation

The main objective of the qualitative inquiry was to gain information on impacts of ROE beyond the scope of those constructs and measures that were anticipated and predefined in the quantitative part of the investigation. In addition, we sought to shed light on how the different groups of people involved in the program experienced and evaluated its implementation, what they identified as facilitating or obstructive factors in the implementation, and what ideas they had for improving the program—particularly with regard to the Swiss context—in the future. To achieve these objectives, we conducted focus groups with instructors, focus groups with children, and expert interviews with teachers, in a multi-perspective, mult-informant approach. Across groups, the foci of the analysis were similar, albeit with slight modifications in method and interview guidelines (see below).

2.3.1 Methods for Data Collection and Analysis

Focus groups may be characterized as moderated group discussions about a particular topic, usually bringing together people with different backgrounds and attitudes (Morgan, 1996). The group moderator follows a carefully prepared guideline with a set of central questions. In contrast to a setting where a group of experts discusses a particular topic in order to achieve a consensus or is expected to present some kind of product at the end of the session, focus groups are aimed at capturing a broad spectrum of individual opinions. Thus, one particular advantage of focus groups is the exploration and close description of various opinions and attitudes concerning a topic.

In the present study, the focus group with instructors was conducted several weeks after completion of the ROE program in a face-to-face session which lasted two hours. The instructors were posed a series of questions, prepared in the guidelines, that covered the impact of the program as well as the implementation and its underlying conditions, mirroring the research questions introduced above (chapter 1). Additionally they were asked about the "backstage" organization of the program (e.g., the support they received from Mentors). To help stimulate the discussion, instructors were in some parts of the interview confronted with statements from the teacher interviews (see below) that had been previously conducted. After participants were informed about the study and gave their consent to partipicate, the group interviews were recorded on audio tapes and subsequently transcribed.

The three focus groups with children from ROE classes all took place in a room separate from the classroom, while class-members not participating in the group were attending regular lessons. Following guidelines in the literature (e.g., Przyborski & Wohlrab-Sahr, 2010), the method as well as the content and precise wording of the questions were adapted to the children's age, and the duration was restricted to 45 minutes in order to avoid putting too much strain on children's attention spans. The children were asked questions that, in content, largely mirrored those posed to the teachers and instructors, covering the topics of implementation, impact, and ideas for improvement. Children were informed about the data collection and data-use in an age-appropriate manner and gave their oral consent to participate. The parents' consent was covered in the letter sent to them (see chapter 2.1). The group discussions were recorded on audio tapes which were subsequently transcribed.

We originally intended to use focus groups as well for teachers. This turned out to be unfeasible for practical reasons, however, because the teachers' work schedules made it impossible for them to be brought together in the same time at the same place. Therefore, we decided to work with individual expert interviews instead. Expert interviews are a special form of semi-structured interviews. The respondent or interviewee is considered as a carrier of expert knowledge about the research subject and, as such, is an integrate part and representative of a group of experts. Using expert interviews as a data source involves posing open questions that allow the experts to present their point of view concerning the issue under study (Flick, 2006, p. 139).

The expert interviews with teachers were conducted either by telephone or face-to face approximately one month after completion of the ROE program. The interviews took about 30 minutes on average and followed interview guidelines prepared in advance, with most questions paralleling those that were posed to the instructors. After being informed about the purpose of the study, the data collection procedure and the use of the data in the future, teachers gave their consent. Interviews were recorded on audio tapes and subsequently transcribed.

The resulting transcripts and protocols from all focus groups and interviews were subjected to a qualitative content-analysis as described by Mayring (2008). The analysis was driven by analytical questions developed from the research objectives, and the resulting categories were grouped in a format according to target groups (instructors, teachers, and pupils) and evaluative focus (implementation quality and impact) before comparing and integrating them. To achieve a reader-friendly mode of presentation, the results from the categorical analysis were summarized in a running text, using verbatim quotes from the interviews and group discussions in indented paragraphs for illustration (see chapter 3.2).

2.3.2 Participants

All teachers (with the exception of those who did not teach the same class anymore) from cohort 1 and 2 were contacted by email or by other means of personal communication and were asked for their participation in an interview or focus group. Six teachers from five

Table 2: Characteristics of participants in expert interviews and focus groups

	Participants	Cohort	Grade (during ROE)	Location of school
Teacher interviews				
Teacher 1	|	2	4	Agglomeration
Teacher 2	|	2	4	Agglomeration
Teacher 3	|	2	5	City
Teacher 4	|	2	6	City
Teacher 5	|	2	4	City
Teacher 6	|	2	4	City
Focus groups with instructors				
Group 1	3	1 and 2	3 to 6	Agglomeration/City
Focus groups with children				
Group 1	7	1	4	Agglomeration
Group 2	7	1	4	City
Group 3	7	2	5	City

different schools agreed to participate. Four of these teachers had taught a class that had participated in ROE in 4th grade, the other two had participated with their classes being in 5th and 6th grade, respectively (see Table 2). Two of the six teachers taught in schools in the agglomeration of the city of Zurich, three in the city of Zurich itself, and one in another city in the canton of Zurich. All teachers had participated with their classes in the second cohort of the study. The number of focus groups to be carried out with pupils was set at three groups because we anticipated that this would enable us to garner a variety of perspectives without overstraining the resources of the teachers we needed for cooperation in putting together the groups. Of the three focus groups with children, two groups came from classes in cohort 1, the third was from cohort 2. The two groups from the first cohort had participated in ROE in 4th grade, and the third group had participated in 5th grade. One of the groups from 4th grade was from a school in the agglomeration of Zurich, and the two remaining came from the city. The selection of participants for the three groups took place on recommendation by the teachers, who were instructed to build a heterogeneous group (in terms of gender and other sociodemographic characteristics) of seven students who were willing to participate. Among teachers and students, both genders were represented almost equally. The group of instructors was exclusively female.

3 Results

We begin our report on the results of the study with the quantitative portion of the investigation. The findings from the qualitative inquiry will follow in chapter 3.2.

3.1 Quantitative Evaluation

3.1.1 Implementation Fidelity

Complete data from instructors questionnaires for all 27 lessons of the program were available for 10 out of the 14 ROE classes. In the four cases for which data was missing, instructors had failed to submit the questionnaires, which was noticed by the research team only after some time had passed. Because of a risk for distortion, no retrospective evaluations were sought in these cases.

As outlined in chapter 1, ROE consists of 27 lessons, divided into 9 lessons devoted to pre-visits, family visits and post-visits, respectively. With only one exception in the entire ROE program (a post-visit lesson in one class), all lessons in all classes were held by their instructors as scheduled (99.6%, see Table 3). There was very little variation in the mean duration of the lessons implemented; the average duration for all lessons combined had a

Table 3: Characteristics of implementation fidelity

	Mean	s.d.	Range
Lessons held, %			
Pre-visits	100.0	|	|
Family visits	100.0	|	|
Post-visits	98.9	|	88.9–100.0[a]
Overall	99.6	|	96.3–100.0
Lessons using all material provided, %			
Pre-visits	66.7	11.7	44.4–77.8
Family visits	86.7	14.6	55.6–100.0
Post-visits	73.3	13.0	44.4–88.9
Overall	75.6	9.8	55.6–88.9
Duration of lessons, minutes			
Pre-visits	45.4	1.3	43.3–48.3
Family visits	44.1	1.5	41.7–45.6
Post-visits	45.2	1.9	42.2–48.3
Overall	45.0	1.1	43.2–46.9
Students' engagement, instructor ratings			
Pre-visits	4.38	0.26	3.89–4.78
Family visits	4.26	0.26	3.89–4.56
Post-visits	4.34	0.15	4.22–4.67
Overall	4.34	0.15	4.00–4.47
Teachers' engagement, instructor ratings			
Pre-visits	4.24	0.33	3.89–4.78
Family visits	4.19	0.37	3.33–4.63
Post-visits	4.11	0.29	3.67–4.67
Overall	4.19	0.29	3.63–4.69

a. Ranges and standard deviations relate to the class level. This means here that among all classes included in the sample, instructors held between 88.9 and 100% of the post-visit lessons that are part of ROE program. Values should be interpreted accordingly for the whole table.

minimum of 43.2 minutes and a maximum of 46.9 minutes across all classes. This indicates the lessons were neatly fitted into the schedule of Swiss primary schools, where each lesson has a supposed length of exactly 45 minutes. In the questionnaires, instructors could also report whether they used all of the material provided by the ROE program for the corresponding lesson; and if not, which material was not used. We found that across all visits, instructors used all the material in approximately three forth of the visits or, put differently, did not use at least one of the materials in one forth. This latter quota was highest in pre-visits (33.3%) and lowest in family visits (13.3%). Reasons given for not using material were divided broadly into three different categories: extraordinary circumstances (such as parent-visiting day), a lack of time due to high student engagement in other parts of the lesson, and, least frequently, the impression of the instructor that the material did not work properly with the class (such as a book being too complicated for oral use).

Finally, all instructors rated their students' engagement as high or very high on average (using a scale from 1 or "not at all engaged" to 5 or "totally engaged"), but there was some variation here, with the lowest and the highest mean engagement across all lessons differing between classes by approximately half a scale-unit (4.00 vs. 4.47, see Table 3). Interestingly, ratings were not higher for family visits with the participation of the parent and the baby than for pre- and post-visits without them—they were even lower, albeit by a very small margin. According to the instructors' ratings, teachers were on average slightly less engaged than the pupils were, but their engagement on average was still also seen to be in the high to very high range, with approximately one scale-unit separating the lowest from the highest (3.63 vs. 4.69).

To sum up, we did not find any stark differences between classes in terms of implementation fidelity, nor did we find any indication that there was an obvious breach with fidelity in any individual class. Therefore, based on the observed data, we decided to include all 13 ROE classes in the subsequent analyses on the effects of the intervention.

3.1.2 Empathy, Aggression, and Prosocial Behavior

At the outset of our analysis on the effects of the ROE program, we compared baseline differences between the two study groups in all outcome variables (Table 4). In the self-report measures, children in the ROE classes rated their own behavior as significantly less aggressive than those in the control group did, and they perceived themselves as significantly less victimized by others. None of the other self-report measures showed significant differences, and neither did the peer nomination measures. Regarding the teacher reports, teachers in the ROE classes perceived their students as more aggressive, less empathic, and less prosocially inclined than teachers in the control classes did. No differences were found in children's social and emotional functioning. Taken together, these data reveal a striking contrast between pupils' self-reports with regard to empathy, prosocial behavior and, most notably, aggression, and teachers' perceptions of their pupils. We will turn to the implications of these findings in the discussion in chapter 4. The detected baseline differences stress the importance of controlling for pretest scores in the analysis of intervention effects, which the design of the current study allowed us to do.

Table 4: Baseline differences between study groups in key outcome variables

	Control	ROE	p value
Composite Measures, mean (s.d.)			
Empathy	5.70 (1.31)	5.36 (1.24)	.010*[a]
Aggression	1.57 (1.25)	1.71 (1.35)	.312
Prosocial Behavior	5.85 (1.52)	5.42 (1.51)	.005**
Students' self-reports, mean (s.d.)			
Affective empathy	2.84 (0.59)	2.83 (0.55)	.967
Cognitive empathy	2.23 (0.63)	2.29 (0.56)	.331
Overt aggression	0.82 (0.63)	0.65 (0.60)	.008**
Relational aggression	0.77 (0.52)	0.60 (0.46)	.003**
Victimization (by others)	0.70 (0.66)	0.52 (0.51)	.010**
Prosocial behavior	2.88 (0.70)	2.80 (0.73)	.296
Well-being	2.88 (0.51)	2.84 (0.54)	.509
Self-esteem	3.24 (0.60)	3.24 (0.55)	.934
Teacher reports, mean (s.d.)			
Empathy	2.90 (0.83)	2.46 (0.84)	<.001***
Aggression	0.54 (0.66)	0.99 (0.93)	<.001***
Prosocial behavior	6.66 (2.14)	5.52 (2.52)	<.001***
Internalizing problems	2.96 (3.35)	3.17 (3.55)	.550
Externalizing problems	4.68 (4.03)	4.81 (4.69)	.771
Psychosocial functioning, sum score	7.64 (6.33)	7.99 (7.15)	.619
Peer nominations, mean % (s.d.)			
Shows empathy	35.54 (17.98)	37.37 (18.13)	.323
Is a friend	42.57 (14.48)	41.96 (14.22)	.678
Is disliked	23.88 (16.42)	21.51 (15.50)	.148
Is aggressive	13.55 (16.38)	11.42 (17.70)	.227
Helps others	39.54 (18.06)	38.53 (18.78)	.593

a. P values are from independent samples T tests (two-tailed). * p < .05. ** p < .01. *** p < .001.

Table 5 provides information on the construct validity of the composite measures we constructed for empathy, aggression, and prosocial behavior (presented for baseline measurements at T0). An important condition for the validity of such an aggregation of data is that the assessments collected from multiple informants do indeed tap into the same underlying phenomenon, which is supposed to be recognized and assessed from different angles. In statistical terms, this means that individual (single-informant) assessments should covary. As Table 5 illustrates, this condition is met without exception for all single-informant measures within any of the three domains; for example, all three single-informant measures on empathy are significantly correlated with each other. A second criterion that adds to the plausibility of a composite measure is if it allows for powerful predictions concerning external constructs that it is theoretically expected to predict. In this case, this was confirmed: The composite score for empathy turned out to be a moderately strong predictor for the composite score in aggression ($r = -.48$, $p < .001$) and a very strong predictor for the composite score in prosocial behavior ($r = .80$, $p < .001$), which were in turn moderately correlated with each other ($r = -.48$, $p < .001$). Moreover, the predictive power of the composite measures was generally stronger than that of the single-informant measures within or across the three domains (see Table 5 for closer inspection). Taken together, these results indicate the usefulness and validity of the composite measures that were constructed.

The main results of our pretest-posttest analyses are presented in Table 6 on page 28. Findings suggest that the ROE program brought about significant change in the desired direction in all three key outcome domains: empathy, aggression and prosocial behavior, as

Table 5: Zero-order correlations for individual and composite measures in key outcome domains at baseline (T0)

	1	2	3	4	5	6	7	8	9	10	11
1. Empathy Composite Score	1										
2. Empathy self-report	.55***	1									
3. Empathy teacher report	.79***	.10*	1								
4. Empathy peer nominations	.81***	.33***	.40***	1							
5. Aggression Composite Score	−.48***	−.12*	−.53***	−.35***	1						
6. Aggression self-report	−.29***	−.20***	−.21***	−.27***	.62***	1					
7. Aggression teacher report	−.35***	.02	−.55***	−.09	.80***	.17**	1				
8. Aggression peer nominations	−.45***	−.14**	−.36***	−.45***	.81***	.38***	.42***	1			
9. Prosocial Behavior Composite Score	.80***	.43***	.67***	.62***	−.45***	−.29***	−.37***	−.36***	1		
10. Prosocial Behavior self-report	.41***	.53***	.19***	.28***	−.16**	−.22***	−.06	−.11*	.65***	1	
11. Prosocial Behavior teacher report	.70***	.16**	.80***	.42***	−.49***	−.26***	−.49***	−.30***	.81***	.23***	1
12. Prosocial Behavior peer nominations	.65***	.32***	.39***	.69***	−.35***	−.18***	−.20***	−.37***	.75***	.27***	.43***

* p < .05. ** p < .01. *** p < .001 (two-tailed).

indicated by the composite measures in the top rows of the table. P values are far below the 1% level, indicating that the effects are robust against type I errors. Effect sizes (calculated as Cohen's *d*'s) range from 0.34 (aggression) to 0.5 (empathy). A Cohen's *d* of 0.5 signifies that the means of the two groups differ by half a standard deviation in the relevant outcome measure; here, the score of the difference between posttest and pretest.

Table 6 further reveals that significant effects are found in all of the teacher report measures, extending into the domains of students emotional and peer problems (internalizing) and conduct and hyperactivity problems (externalizing) as well. However, none of the differences in the students' self-reports reach statistical significance. Significance aside, there is a trend toward more positive change in the ROE group, but this trend is small. The same observation applies to the peer nomination scores. In one case, the trend towards a small positive impact of ROE is statistically significant: Children in the ROE group showed slightly higher gains in the likelihood to be nominated as a friend by their fellow pupils.

Do some groups of pupils respond differently to ROE than others?

Beyond these findings, which apply to the intervention group as a whole, we were interested in subgroup effects: Do some groups of pupils within the intervention group respond differently to the ROE program than others—perhaps in the sense that some respond positively whereas others do not respond or respond negatively? We looked at such between-group differences in four dimensions: gender, grade level, baseline differences in the outcome variables, and the first vs. the second cohort of the ROE program.

The results of our two-ways ANOVAs investigating possible moderation by gender found a significant main effect in the dimension of prosocial behavior, with girls generally showing more positive developments in prosocial behavior during the course of the school-year than boys ($F(1,377) = 3.95$, $p = .048$). Put differently, the gender gap in prosocial behavior became wider between one year and another. Main effects for gender with regard to difference scores in empathy and aggression approached significance, but did not drop below the 5% level ($p = 0.061$ and 0.194, respectively). Interaction between gender and treatment was

Table 6: Unadjusted (observed) raw scores ouf outcome variables, difference scores and between-group effects

	Control (n = 249)			ROE (n = 187)			Group effects	
	Pretest	Posttest T1	Diff Score	Pretest	Posttest T1	Diff score	*p* value	Cohen's *d*
Composite Measures, mean (s.d.)								
Empathy Score	5.70 (1.31)	5.80 (1.44)	0.09 (0.85)	5.36 (1.24)	5.98 (1.43)	0.59[a] (1.06)	<.001***	0.50[b]
Aggression Score	1.57 (1.25)	1.68 (1.33)	0.12 (0.87)	1.71 (1.35)	1.55 (1.32)	−0.17 (0.83)	.001**	0.34
Prosocial Behavior Score	5.85 (1.52)	5.99 (1.55)	0.13 (1.14)	5.42 (1.51)	6.03 (1.52)	0.59 (1.10)	<.001***	0.41
Students' self-reports, mean (s.d.)								
Affective empathy	2.84 (0.59)	2.89 (0.62)	0.05 (0.61)	2.83 (0.55)	2.97 (0.55)	0.13 (0.54)	.198^{ns}	
Cognitive empathy	2.23 (0.63)	2.35 (0.69)	0.12 (0.55)	2.29 (0.56)	2.44 (0.59)	0.14 (0.54)	.715^{ns}	
Overt aggression	0.82 (0.63)	0.79 (0.61)	−0.01 (0.51)	0.65 (0.60)	0.66 (0.55)	0.01 (0.51)	.679^{ns}	
Relational aggression	0.77 (0.52)	0.73 (0.54)	−0.02 (0.46)	0.60 (0.46)	0.60 (0.47)	−0.03 (0.43)	.864^{ns}	
Victimization (by others)	0.70 (0.66)	0.66 (0.60)	−0.02 (0.49)	0.52 (0.51)	0.54 (0.57)	0.03 (0.48)	.655^{ns}	
Prosocial behavior	2.88 (0.70)	2.84 (0.72)	−0.02 (0.71)	2.80 (0.73)	2.88 (0.66)	0.08 (0.68)	.179^{ns}	
Well-being	2.88 (0.51)	2.76 (0.56)	−0.11 (0.55)	2.84 (0.54)	2.77 (0.60)	−0.07 (0.52)	.481^{ns}	
Self-esteem	3.24 (0.60)	3.31 (0.56)	0.07 (0.55)	3.24 (0.55)	3.32 (0.59)	0.08 (0.49)	.846^{ns}	
Teachers' reports, mean (s.d.)								
Empathy	2.90 (0.83)	2.82 (0.85)	−0.09 (0.64)	2.46 (0.84)	2.84 (0.87)	0.29 (0.85)	<.001***	0.49
Aggression	0.54 (0.66)	0.71 (0.70)	0.18 (0.64)	0.99 (0.93)	0.69 (0.80)	−0.28 (0.67)	<.001***	0.75
Prosocial behavior	6.66 (2.14)	6.68 (2.25)	0.21 (1.86)	5.52 (2.52)	6.96 (2.48)	1.25 (2.16)	<.001***	0.51
Internalizing problems	2.96 (3.35)	2.72 (2.62)	−0.15 (2.26)	3.17 (3.55)	2.49 (3.20)	−0.66 (2.43)	.042*	0.21
Externalizing problems	4.68 (4.03)	4.62 (3.69)	−0.03 (2.76)	4.81 (4.69)	3.74 (4.22)	−0.96 (2.81)	.002**	0.33
Psychosocial functioning, sum score	7.64 (6.33)	7.35 (5.43)	−0.18 (4.12)	7.99 (7.15)	6.23 (6.44)	−1.62 (4.11)	.001**	0.35
Peer nominations, mean % (s.d.)								
Shows empathy	35.54 (17.98)	38.67 (19.85)	2.99 (12.61)	37.37 (18.13)	41.52 (21.46)	3.78 (14.21)	.577^{ns}	
Is a friend	42.57 (14.48)	46.13 (17.68)	3.40 (12.80)	41.96 (14.22)	48.42 (18.72)	6.48 (14.16)	.030*	0.22
Is disliked	23.88 (16.42)	22.89 (16.92)	−0.77 (13.05)	21.51 (15.50)	19.24 (16.12)	−2.26 (12.65)	.263^{ns}	
Is aggressive	13.55 (16.38)	13.17 (18.48)	−0.31 (11.50)	11.42 (17.70)	13.25 (19.26)	1.78 (10.47)	.067^{ns}	
Helps others	39.54 (18.06)	40.70 (20.32)	1.02 (14.36)	38.53 (18.78)	39.55 (21.27)	0.74 (15.73)	.857^{ns}	

Analyses were performed controlling for age and gender as covariates.

a. Means of difference scores may differ from the differences between posttest means and pretest means; this is due to missings at either pretest or posttest.

b. Effect sizes are only reported if between-group differences were statistically significant ($p < .05$).

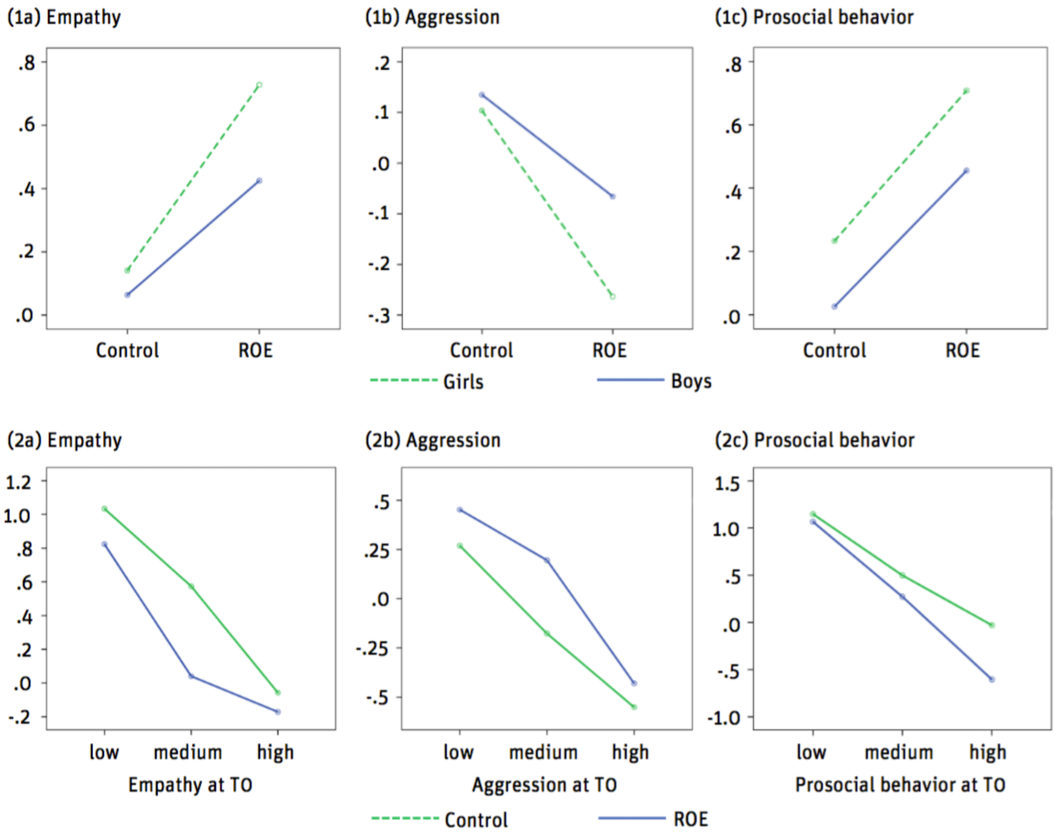
* $p < .05$. ** $p < .01$. *** $p < .001$. ns = non-significant.

not significant for all three outcome variables. This means there is no robust evidence that girls responded differently to the ROE program than boys did (*p*'s ranging from 0.267 for empathy to 0.844 for prosocial behavior). As can be seen in Figure 2, which shows the results for our multifactorial analysis, there is a tendency towards more pronounced effects for girls than boys in empathy and aggression, one that might have been detected in a larger sample, but this must remain speculative.

Next, we looked at possible moderation by baseline differences. In everyday terms, this question may be posed as follows: Do children who show low empathy (or aggression or prosocial behavior) at the outset of the program respond differently than children startingout at an already medium or high level? Does the intervention make already empathic children more empathic—while those who lack empathy at the beginning are left where they are? Or does the program allow those lacking empathy to catch up with the others, closing the gap between them?

In order to investigate this matter, we first divided students into three groups for each outcome domain, based on their values at baseline in the composite measures introduced in

Figure 2: Multifactorial ANOVAs considering moderation by gender and baselines levels



All models were calculated controlling for age and grade. Models for moderation by baseline differences also included gender as a covariate.

chapter 2: that is, the 25% with the lowest values for empathy/aggression/prosocial behavior, those with medium values (interquartile range, 26th to 75th percentile) and the remai-ing 25% with the highest values. We then analysed whether belonging to any of these groups made a difference in terms of intervention effects (Figure 2, bottom half).

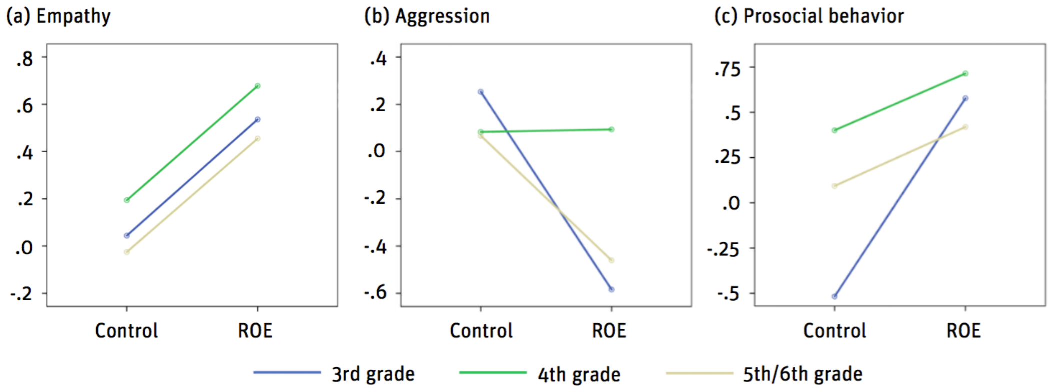
The results show that no such interaction is at play, at least not in a pronounced fashion. Matching baseline differences at T0 in the three domains empathy, aggression and prosociality to the corresponding outcomes (difference scores between T1 and T0) in the same domains, we found significant (and fairly large) main effects for group membership in all cases (for empathy: $F(2, 374) = 22.175$, $p < .001$, partial $\eta^2 = .106$; for aggression: $F(2, 374) = 24.455$, $p < .001$, partial $\eta^2 = .116$; for prosocial behavior: $F(2, 374) = 36.174$, $p < .001$, partial $\eta^2 = .162$). Post-hoc analyses revealed that the differences were significant between each pair, i.e., that students with low levels differed signifcantly from those with both medium and high levels, who in turn differed from each other (Bonferroni-corrected *p*'s ranging from below 0.001 to 0.024). The direction of these effects may be gleaned from the bottom half of Figure 2. Interestingly, in all three outcome dimensions, those who started out with the lowest scores developed most favorably, and the change was always in the desired direction (i.e., towards more empathy, more prosocial behavior, less aggression) both in the ROE and the control classes. For children who started out in the medium range, the intervention seems to have tipped the scales with regard to empathy and aggression: on average, the children from this group in the control classes grew neither more nor less empathic, and they became slightly more aggressive, whereas the corresponding group in the ROE classes made considerable gains in empathy and became slightly less aggressive. In terms of prosocial behavior, pupils in the medium range showed favorable developments both in ROE and controls, albeit not as strongly as those with low baselines scores did. Finally, on average, those children who started out with the highest scores of empathy more or less stayed at that same level in both groups, and those who started out with the lowest scores in aggression became somewhat more aggressive, again in both groups. With regard to prosocial behavior, however, the program again made a difference: While the most prosocial children at T0 grew less prosocial with time in the control group, those in the ROE classes remained roughly at the same level.

Importantly, the interaction between treatment and group membership was not significant (for empathy: $F(2, 374) = 1.940$, $p = .145$; for aggression: $F(2, 374) = .919$, $p = .400$; for prosocial behavior: $F(2, 374) = 1.430$, $p = .241$). This means that the program seems to have produced effects of roughly the same sizes across the whole spectrum of pupils (where this spectrum is defined by baseline values in the outcome dimensions). Considering that the lines for the two groups do not run neatly in parallel, it is possible that a small interaction was in fact at play; visual inspection of the data presented in Figure 2 suggests a tendency towards more pronounced effects (for empathy and aggression) in the medium range. In everyday terms, this would mean that children who start out with moderate levels in these two dimensions might profit slightly more, on average, from the ROE program than those at either the low or the high end of the spectrum do. This remains speculative, however.

Next, we looked at the question whether the effects of the intervention differed depending on which grade the pupils were in (Figure 3). Because of the small subsample for grade 6 students (only one class each in the ROE and the control group), we merged pupils from both 5th and 6th grade into a single group, which resulted in three groups for comparative analysis: 3rd, 4th and 5th/6th grade. The results reveal no significant main effect in the dimension of empathy ($F(2, 374) = 1.880$, $p = .154$) and no interaction with treatment ($F(2, 374) = .001$, $p = .999$). For prosocial behavior, there was a significant main effect for grade ($F(2, 374) = 6.016$, $p = .003$), with Bonferroni-corrected post-hoc analyses showing that pupils in grade 3 made significantly larger gains than those in grade 4 ($p < .001$) and grades 5 to 6 ($p = .030$), both in the ROE and the control classes. There was no such difference between grades 4 and 5/6. A visual inspection of the data suggests that ROE fostered children's prosocial behavior more strongly in 3rd graders than in older children (see Figure 3), but the interaction missed significance ($F(2, 376) = 2.512$, $p = .083$). Concerning aggression, there was both a significant main effect ($F(2, 374) = 0.4896$, $p = .008$) and an interaction effect ($F(2, 374) = 6.994$, $p = .001$). Post-hoc analyses confirmed what the visualization of the data in Figure 3 strongly suggests: While ROE seems to have reduced children's aggression to a similar degree in 3rd and 5th/6th grade students, it was apparently not successful in this regard with 4th graders.

Finally, we also tested for a possible interaction between treatment and cohort, working from the hypothesis that ROE, with experience among instructors increasing, might have brought about larger effects in its third than its second year of implementation. However, no significance differences were detected in any of the three key outcome domains.

Figure 3: Multifactorial ANOVAs considering moderation by grade



Models were calculated controlling for age and gender.

Are the effects of the intervention sustained beyond completion of the program?

Table 7 shows findings on the one-year follow-up measurements. Follow-up data could only be obtained for 3 out of the 4 ROE classes in the first cohort, because in the fourth classroom pupils had left primary school and entered secondary school at follow-up, where they were no longer available for data collection. To keep the matching procedure intact, three corresponding control classes were included in the follow-up measurments, resulting in a subsample of 107 pupils. For these, data were available on all three time-points, pretest, posttest and follow-up. In the top half of Table 7, results pertain to the question whether the effects of the ROE program were still detectable one-year after the program had ended.

The answer is yes for empathy and aggression: the data show lasting effects of moderate proportions (*d* = 0.47 and 0.46, respectively). The between-group differences for prosocial behavior on the other hand were not significant, although this might be due to the considerably reduced power because of the limited sample size. The results shown in the bottom half

Table 7: Unadjusted (observed) raw scores ouf outcome variables, difference scores and between-group effects for follow-up measurements

	Control (N = 55)			ROE (N = 52)			Group effects	
	Pretest T0	Follow-up T2	Diff Score	Pretest T0	Follow-up T2	Diff score	*p* value	Cohen's *d*
Composite Measures, mean (s.d.)								
Empathy Score	5.68 (1.08)	5.73 (1.52)	0.04 (1.00)	5.28 (1.24)	5.92 (1.13)	0.64 (1.14)	0.024*	0.47
Aggression Score	1.20 (1.04)	1.28 (1.22)	0.07 (1.08)	1.69 (1.36)	1.38 (1.32)	−0.32 (1.08)	0.027**	0.46
Prosocial Behavior Score	5.38 (1.48)	6.15 (1.38)	0.77 (1.07)	5.21 (1.53)	6.24 (1.55)	1.02 (1.29)	0.135$^{\text{ns a}}$	
	Posttest T1	Follow-up T2	Diff Score	Posttest T1	Follow-up T2	Diff Score	*p* value	Cohen's *d*
Composite Measures, mean (s.d.)								
Empathy Score	6.10 (1.43)	5.73 (1.52)	−0.38 (1.00)	5.86 (1.33)	5.92 (1.13)	0.05 (1.10)	0.077$^{\text{ns}}$	
Aggression Score	1.34 (1.04)	1.28 (1.22)	−0.06 (0.67)	1.65 (1.49)	1.38 (1.32)	−0.28 (0.90)	0.069$^{\text{ns}}$	
Prosocial Behavior Score	6.27 (1.48)	6.15 (1.38)	−0.11 (0.76)	6.01 (1.67)	6.24 (1.55)	0.22 (0.87)	0.086$^{\text{ns}}$	

Analyses were performed controlling for age and gender as covariates.
a. Effect sizes are only reported if the between-group differences were statistically significant (p<.05).

of Table 7 are relevant to the question whether there were any additional gains in the ROE group setting in after the program had ended. The observed differences in change scores indeed point in that direction for all three outcome domains, but they miss the 5% level of statistical significance.

Does change in empathy explain changes in aggression and prosocial behavior?

In chapter 1, we briefly introduced a theoretical model drawn from the literature, one that assumes an important and straightforward causal connection between empathy and social behavior: Empathy is supposed to be a driving force behind the formation and extent both of prosocial behavior (which it promotes) and aggressive behavior (which it hinders). The reasoning behind this assumption roughly goes as follows: Empathy is required to take another person's perspective, and taking another person's perspective is required to value that perspective, to care for the other person and for his or her emotional well-being. This concern then leads to a desire a) to help the other person (prosocial behavior) and b) to refrain from harming that person (aggression). Translated to the area of ROE, the model predicts that changes observed in the domains of prosocial behavior and aggression should be closely related to changes in empathy. In the current study, we intended to test this model using the longitudinal data available. A straightforward way to do this presented itself in the form of examing the relationship between change scores observed between post-testing and pretesting. We assumed that if changes in empathy are causally related to changes in prosocial behavior and aggression, than this would show up in the correlations between change scores in these outcome domains. Correlations by themselves do not indicate a causal relationship, of course; from a statistical point of view, a correlation between measures A and B may mean that A influences B, that B influences A, or that the common variance is attributable to any of several other possibilities, for example, to a third-factor C influencing A and B at the same time. In the current case, however, the assumption that empathy causally influences prosocial behavior and aggression seemed more plausible than positing the relationship the other way around. In any case, finding no or only a weak relationship between change scores in these measures would invalidate (though not decisively falsify) the assumption that there is a substantial causal relationship between them.

In addition, even if there is a strong relationship between empathy and social behavior, empathy is hardly the only contributing factor. One other factor that has been shown in empirical research to influence social behavior is self-control or, put differently, a person's impulsivity, that is, his or her ability to resist impulses for action out of consideration for the consequences of that action (e.g., Baumeister, Heatherton, Tice, 1994; Finkenauer, Buyukcan-Tetik, Baumeister, Schoemaker, Bartels, & Vohs, 2015). The role of self-control is considered to be particularly important for aggression: the more self-controlled someone is, the more often will he or she be able to resist the temptation to act aggressively. In the current study, information on the level of children's self-control was available from a single source: the hyperactivity scale of the "Strengths and Difficulties Questionnaire" that teachers filled in for every pupil (cf. chapter 2).

The inclusion of this measure, which was taken to serve as a proxy for self-control (cf. Aguilar-Cárceles & Farrington, 2017), afforded us with a straightforward opportunity to test for the influence of empathy in the formation of prosocial behavior and aggression, considering the concurrent influence of self-control. For both outcome domains, we calculated multiple linear regression models with changes in prosocial behavior and aggression between posttest and pretest, respectively, serving as dependents, and age, gender, changes in empathy and changes in hyperactivity (between posttest and pretest) as independents. The results are shown in Table 8.

Table 8: Multiple linear regression models predicting change scores for prosocial behavior and aggression

	Prosocial behavior		Aggression	
	ΔR^2	β	ΔR^2	β
Model	.307		.086	
Gender (female)		.056		-.047
Age (in years)		.041		-.153**
Empathy (change score, composite)		.541***		-.230***
Hyperactivity (change score, teacher report)		-.032		.062

** p < .01. *** p < .001 (two-tailed)

As theoretically predicted, the change scores in empathy correlated significantly with change scores in prosocial behavior. None of the other independents had a detectable independent association with this change. The changes in empathy explained roughly 30% (r = .541, $p < 0.001$) of the variance found in the composite measure for prosocial behavior. With regard to aggression, empathy was a much weaker but still a significant predictor (explaining roughly 5% of the variance; r = –.23, p < .001). Changes in hyperacitivity showed no independent association with changes in aggression. However, age did, with older children on average developing more favorably between post-testing and pre-testing than younger children. Taken together, the models confirm the assumption that changes in prosocial behaviors are to a large extent influenced by changes in empathy, but there is no such strong relationship between empathy and aggression.

3.1.3 Emotion Recognition

The measures in the present study included a test of emotion recognition: Pupils watched 15 video clips which all showed one child or adult expressing a certain emotion through movements of their heads and facial muscles while sound was suppressed. The clips were taken from the "Cambridge Mindreading Face-Voice Battery for Children" (for more information, see Golan, Baron-Cohen & Hill, 2006). After seeing each video, pupils had to choose from a list of four adjectives, picking the one that best identified the emotion. In a first step, we performed a reliability analysis on this original 15-item scale, calculating Kuder-Richardson (KR-20) coefficients to check for internal consistency issues and sorting out individual items with unacceptably low item-total correlations (< 0.3) that reduced the overall consistency of the scale. This procedure resulted in the exclusion of six items, so that the final scale included 9 items. In a second step, we calculated zero-order correlations for this new scale in relation to the major empathy measures included elsewhere in our study, in order to explore the scale's convergent validity. It was expected that the ability to recognize emotions from facial expressions would contribute a small to moderate portion of variance to children's overall empathic capacities.

Table 9: Zero-order correlations for the Emotion Recognition Scale and other measures of empathy

	Composite	Self-reports		Peer ratings	Teacher reports	
	Empathy	Affective empathy	Cognitive empathy	Empathy	Empathy	German skills
Facial Emotion Recognition Scale	.273***	.323***	.204***	.256***	.117*	.300***

* p < .05. *** p < .001 (two-tailed)

The results of this exploration are shown in Table 9. As can be seen, they are in line with expectations. The correlation of the scale was a bit higher for children's self-reported affective empathy (sensing what another feels) than for cognitive empathy (taking another's perspective), with the coefficient for peer-rated empathy being in the middle of these two. The lowest correlation was found between children's emotion recognition performance and the teacher rating for their empathic capacities. Because the test measure used in this study relies on children's ability to pick the right word from a list of alternatives, we expected that students' proficiency in the German language might be an important confounder. The correlation between emotion recognition and language skills as rated by the teacher confirms this assumption (r = .30, p < 0.001, see Table 9). To confront this matter, we also calculated partial correlations between the emotion recognition scale and other empathy measures, controlling for language skills, with the possibility in mind of using only the residuals from a regression of emotion recognition on language for subsequent analysis. Interestingly, however, these partial correlations were not larger overall but smaller than the zero-order correlations. A plausible explanation is that children's knowledge of the German language not only affects the specific performance of picking the right word from a list but also contributes to children's empathy in general, beyond emotion-to-word matching tasks. Sharing a language is an important enabling factor both in understanding others' emotions (which are in part verbally expressed) and for displaying this understanding to others (which is also in part verbally expressed). Therefore, German language skills plausibly explain some of the common variance between performance in a facial emotion recognition task (one drawing on language) and more encompassing measures of empathy. As a consequence of this argument, we decided to use the unadjusted 9-item scale in the examination of intervention effects.

The results of this analysis suggest that ROE made no difference: Participation in the program did not influence children's capacity to recognize and identify emotions from facial expressions alone. A unifactorial model controlling for gender and grade found no effect for treatment (*F*(1, 227) = 0.126, *p* = .723), and there weren't any detectable late gains at follow-up, either (*F*(1, 97) = 1.668, *p* = .200).

3.1.4 Altruism

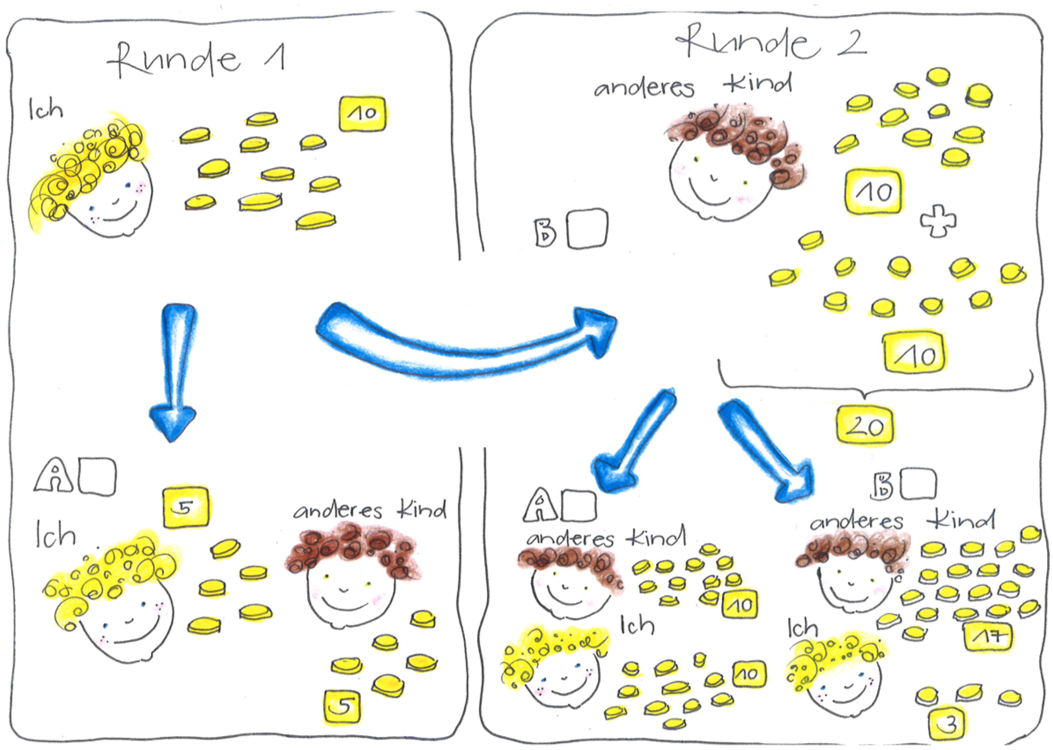
To measure children's tendencies to act fairly and altruistically, the study included two measures behavioral measures briefly pointed to above (see chapter 2): They participated in a decision-making task called the Trust Game and they were given the opportunity to donate all or some of the money they had earned in the game to a charity organization working with disadvantaged children. We will start with our findings on altruism in the Trust Game.

Altruism in the Trust Game

The Trust Game is a decision-making paradigm that was developed within the tradition of behavioral economics. Together with other variants such as the Ultimatum Game or the Dictator Game, it serves as a behavioral measure to capture people's tendencies towards trust, fairness, altruism and other such constructs (and their respective opposites) in scenarios that usually involve one actor playing with or against one other (for a general introduction, see e.g. Fehr, Fischbacher, Von Rosenbladt, Schupp, & Wagner, 2003; for an introduction on the Trust Game with children: Sutter & Kocher, 2007).

In Round 1 of the Trust Game, the children in our study were given (symbolically, in a drawing) 10 coins. They then had to make a choice between two options: A) They could divide the 10 coins equally between themselves and another (anonymous) child they were playing with. This would secure both children 5 coins and would end the game. B) They could pass the 10 coins on to the other child, who would later, in Round 2, receive 10 more coins, resulting in 20 coins altogether. The other child would then have to choose to either divide the coins equally—which would secure both children 10 coins each—or to keep 17 coins for him/herself and give only 3 coins to the child who had passed on the decision in Round 1. Therefore, choice A) meant that children would receive a modest amount of 5 coins with certainty, while choice B) might either reward them with the large amount of 10 coins or leave them with the small amount of 3 coins, depending on the decision the other child was going to make. The rules of the game were explained to the children in an age-adequate manner, using drawings and reserving enough time for questions and answers (see Figure 4 for illustration). It was also explained that the coins would eventually be exchanged for real money when the researchers would come back to meet the class again (for post-testing or follow-up), and that the maximum amount of money children could make would be approximately 5 Swiss Francs (roughly USD 5). In Round 1, which presented children with the choice between options A) and B), they repeated the game six times, three times with an anonymous child from their own class and an equal number of times with an anonymous child from a different class in a different school. Children were told that while the researchers knew who they were playing with, none of the children would ever be given this information, neither now nor later, in order to prevent quarrels or other vexations that might result from such knowledge. This, however, was a cover story: In truth, the other players were made up by the researchers, a necessary condition to establish equal conditions for all students in Round 2 of the game. The cover story was eventually resolved and explained to the children after the post-testing session.

Figure 4: Visual illustration of the Trust Game



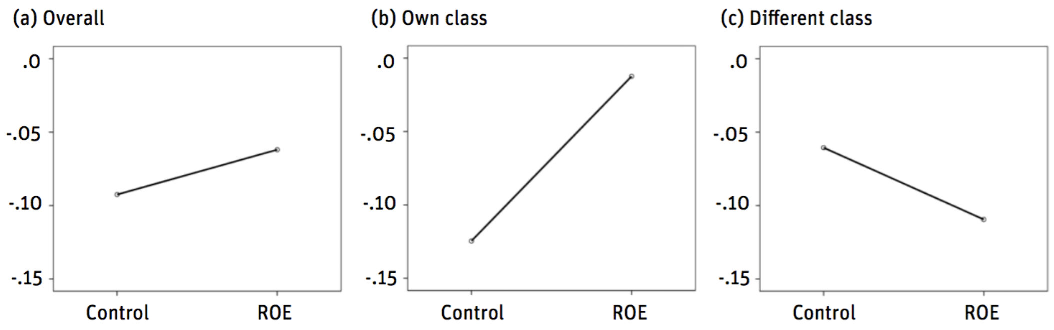
Translation: Runde 1 (2) = Round 1 (2). Ich = Me. anderes Kind = other child.

In Round 2, children were placed in the position of the second child, the one who had been trusted with the choice between an equal/altruistic division (10 coins for both children) and an unequal/egotistical division (17 for themselves vs. 3 for the other child). We prepared materials so that each child believed the other two children he/she was playing with, one from their own class and one from another, had trusted him/her with the decision in Round 2 two out of three times. As a result, each child could make the choice between an equal and an unequal division four times altogether.

In the current study, we were interested in children's decisions in Round 2 of the game. The equal division of the 20 coins in Round 2 may be termed an altruistic (and not merely a prosocial) act because, by performing this act, children choose to benefit another (giving him/her 10 coins instead of 3 coins) at a cost to themselves (receiving 10 coins instead of 17), which meets the textbook definition of altruism. As children were informed beforehand that Round 2 would end the game and that nobody would ever learn what decisions they took, an egotistical explanation of the equal division (such as a need to present oneself in favorable light or an attempt to get the other child to cooperate in the future) is not plausible. By examining children's choices in Round 2, we investigated whether participation in the ROE program made them more likely to act altruistically and, if so, whether it would make a difference which class the other child supposedly was from, their own or another from a different school.

Prior to the analysis, we checked whether children had fully understood the procedures of the Trust Game. Misunderstanding the game inevitably led to certain errors in filling in

Figure 5: Change scores for the two study groups with regard to altruism shown in the Trust Game



Models were calculated controlling for age, grade and gender.

the response sheets, and we used these errors to exclude erring pupils from the analysis, appyling conservative criteria that did not tolerate a single mistake. This resulted in a considerable proportion of students at pretest (21.9%) and posttest (9.4%) being excluded. The dependent variable used in our analysis was, again, a difference score: the difference between the number of altruistic choices (possible values: from 0 to 4) at post-test and pretest. We divided this original difference score by the number of choices that the children could take (4 altogether), so that it became proportional to a single choice: a difference score of 0.2, for example, would mean that the likelihood of an altruistic choice rose by 20 percent.

The results reveal an intriguing pattern (Figure 5). First, mean difference scores were negative in both groups, meaning that the frequency of altruistic choices had decreased overall, by a small margin, in both ROE classes and controls. Second, there was no main effect for treatment (ROE vs. control) on the difference score for all choices combined (F(1, 259) = .424, p = .516). However, there was a significant effect when only the choices regarding the children's own class were considered (F(1, 269) = 4.301, p = 0.039, Cohen's d = 0.26): Pupils in the ROE group were significantly more likely to behave altruistically towards an anonymous member of their own class than those in the control group. Taking a closer look, the likelihood of altruism towards a member of one's own class decreased by 11% in the control group between pretest and posttest, while it did not decrease in the ROE group. Concerning choices made with regard to an anonymous child from another school, the two groups did not differ in their altruism (F(1, 262) = .766, p = .382).

Altruism in making a donation

The second measure of altruism in our study relates to children's readiness to donate money to a charity organization, one that works with disadvantaged children in Southeast Asia. The donation sequence of the study was part of only post-test and follow-up sessions. The money that the children could choose to donate was directly linked to the Trust Game. Children were informed at posttest and follow-up that they would all receive the maximum amount of 5 Swiss Francs from the Trust Game they had played several months before (at either pretest or posttest), and we explained that this was because we did not wish to create any inequalities between them (an announcement that was invariably greeted with cheers). Then, all children were given envelopes containing 5 coins in the amount of 1 Swiss Franc each. Thereafter, children saw a short film of approximately five minutes which presented the work of the charity organization. Children were then told that they could make a donation to the charity organization if they wanted to by putting any number of coins back in the envelope. The envelopes were collected by the researchers in a way that ensured anonymity, which was considered critical in avoiding social desirability bias. The money collected in this procedure was in fact donated to the charity organization.

The donation measure of altruism was included in the study because it provided an opportunity to test children's "real-world" behavior, something that went beyond the game-like character of the Trust Game, and also because it extended the kind of altruism being addressed: here, *the other* in the altruistic act was not a peer going to the same class or at least to a school in the same part of the same country (as in the Trust Game), but an obviously disadvantaged group of children living in a remote part of the world. In this sense, the donations could be used as a tool to validate the choices taken in the Trust Game and, beyond that, as a behavioral correlate of the outcome measures of empathy and prosocial behavior. Table 10 shows the zero-order correlations (Spearman's ρ) for the amount of money donated and several other constructs at posttest (T1).

Table 10: Zero-order correlation matrix for donations and related measures of altruism, empathy, prosociality

	1	2	3	4	5	6
1. Donation	1					
2. Altruistic decisions (Trust Game, overall)	.254***	1				
3. Empathy (composite score)	.324***	.307***	1			
4. Empathy (self-report)	.273***	.222***	.618***	1		
5. Prosocial behavior (composite score)	.296***	.202***	.788***	.457***	1	
6. Prosocial behavior (self-report)	.188***	.102	.410***	.472***	.591***	1
7. Self-esteem (self-report)	.185***	.030	.312***	.345***	.382***	.481***

*** p < .001 (two-tailed)

As can be seen, the donations were signifantly correlated in the expected directions with all other constructs considered. Interestingly, the correlation was stronger for empathy (composite score and self-reports) than for altruism in the Trust Game or prosocial behavior (composite score and self-reports). The correlation with self-esteem disappeared when empathy was partialed out, indicating that empathy's third-factor contribution both to self-esteem and to altruism (as expressed in donations) explained the link between the two.

Did participation in ROE have an effect on children's willingness to give money to a charitable cause? Comparing post-test values in a unifactorial ANOVA controlling for gender, grade and subjective socio-economic status of children[1], we found that the estimated means differed between the two groups: average donations amounted to CHF 1.90 in the ROE group and CHF 1.68 in the control group, but this difference barely missed significance. At follow-up, the gap had widened, with estimated means now being at CHF 2.84 for the ROE and CHF

[1] We suspected that children's willingness to donate money might depend to some degree on how much material resources they already have, and therefore included a composite measure of children's subjective socio-economic status (sSES) in the analysis. There was no significant bivariate correlation, however, and the inclusion of sSES in the model did not substantially alter results.

1.88 for the control classes ($p < 0.05$). As we could not control for pretest scores in this case, a rigorous test of intervention effects was not possible. Judging only from the posttest scores, there is a suggestion that ROE substantially increased children's willingness to donate.

3.2 Qualitative Evaluation

In the following chapter, we will give an account of our qualitative inquiry into the implementation quality and impact of the ROE program in Switzerland. As outlined in chapter 2, the inquiry involved individual interviews with six teachers, one focus group with instructors and three focus groups with children. The chapter is organized into two sections: i) responses to questions about the implementation quality of ROE in Switzerland, and ii) responses about the program's impact. The implementation section covers a variety of topics divided into subsections: what the participants perceive as ROE's key messages, how they describe and evaluate the educational approaches used in the program, whether they think the program will be more or less suitable and effective depending on specific children and class characteristics, what they think were facilitating or obstructive conditions underlying the implementation, and, finally, what ideas they have that might help to improve the program, particularly with regard to Swiss context, in the future.

All groups responded to questions in Swiss German or standard German. In presenting the results, we translated quotes from the interviews and groups into English, trying to mirror the colloquial mode of expression (including grammatical irregularities in German where there was an English equivalent) wherever possible.

3.2.1 Implementation

Teachers and instructors were first asked what constituted to them the key message of the ROE program. The teachers' answers forefront the two topics of infant development and empathy. The following statement by a teacher exemplifies this:

> For sure, a big part was devoted to rearing infants or babies and to knowledge about pregnancy, but also a lot of it was about empathy and how children can try to empathize with somebody. (T5)[2]

Apart from knowledge about early infant development, the majority of teachers also mention the experience of social interaction with babies as a primary topic. Concerning empathy, their focus varies. Half of the teachers mention that paying attention to other people and treating them respectfully was at the heart of the lessons. One teacher considers communicating the importance of empathy as central, and another thinks the program stresses how empathy plays a role in many different domains of everyday life. Some teachers bring

[2] Participants are identified in this chapter by abbreviations, with the letter indicating group membership (T for teachers, I for instructors, C for children) and the number refering to individuals within the group. This numbering is arbitrary. T5, for example, means teacher number 5.

up still other aspects like learning how to find solutions in difficult situations or how certain feelings (such as joy or insecurity) occur in children and adults independent of age or life experience.

For the instructors, ROE's primary focus rests on empathy and the development of emotions, with knowledge about infant development playing only a secondary, assisting role. In their view, the program is supposed to support children in becoming aware of their emotions and to help them in naming and communicating their emotions appropriately. Building on that, children are supposed to learn how to empathize with others and to see interindividual differences between people not as a problem but as something to be appreciated.

Description of the learning approaches used in the program

All participants in the focus groups and interviews were asked to recollect what educational methods and formats had been used in the program and how they had experienced them. The teachers differentiate strongly between the categories of pre-visits, family visits and post-visits. Concerning the family visits, they agree that the most important learning approach was that of observation, as is expressed in the following quote:

> It was like an observational study, really. The baby was in the center and we were sitting around the baby and carefully observing the baby. (T6)

In contrast, most teachers see the predominant method in the pre- and post-visits as teacher-centred instruction, or colloquially called "chalk and talk", which involves the instructor explaining something or asking questions which are subsequently discussed with the whole class. More than half of the teachers consider the development of a theme on the basis of story-telling as another important element. The instruction to work in groups (e.g., by crafting or writing) is only mentioned by one person. The program's educational approach is summed up by a teacher in the following way:

> So, at the center… let's put it this way… the program focuses on specific tasks and skills that the children do and learn, and then the competence for empathy is something that is kind of learned on the side, in passing… So they learned things and then were asked how that felt. It was actually kind of a training in recognizing emotions. (T6)

The instructors, by comparison, do not draw such a strong distinction between pre- and post-visits on the one hand and family visits on the other. In their description, the program communicates its contents and messages primarily in the form of asking children questions and of subsequent discussions, which are led by the instructor and take place in a circle with all children. Beyond that, instructors also mention the use of stories, which are considered as important particularly for knowledge transfer, and autonomous tasks (e.g., worksheets, group works, crafting) as important elements.

When the children are asked about the predominant educational approach of the program, all three groups emphasize that they were frequently observing the baby and were given the opportunity to perform experiments, such as giving the baby a doll or other toys.

They also agree that the instructor was doing “a lot of talking” (as one child puts it) in the pre- and post-visit lessons and that they mainly listened to the instructor, while there was sometimes also room for discussion. All three groups mention additional elements such as telling stories, doing handicrafts (such as collages, door signs, posters, rhymes) and worksheets as well as the repetition of learnt skills.

Evaluation of the learning approaches

Having described the main learning methods they recollected, all groups of participants were asked how they evaluated these methods: whether (and where) they found them appropriate and effective. In their response, the teachers unanimously agree that they highly appreciated the hours with the baby and the parent. In their perception, the children always looked forward to these lessons and were interested and engaged while they lasted. The following statement by a teacher sums it up:

> And the lessons with the baby… they were obviously a big hit with the children and they were waiting all the time for it [the baby] to return. Because for them this was really a highlight in their life in school, I would say. (T2)

One teacher points out that getting involved with the baby made it possible to induce empathy in children in a very direct and natural way:

> In particular I liked the visits of the baby. I think this is really fantastic, because through the baby’s presence they got like a direct connection. This way they can relate, they can directly observe and also comprehend those things… and I think that empathy is really stimulated much more in this way than if just children of the same age were present… or if we just told them about the topic. (T5)

In contrast, several teachers note that pre- and post-visit lessons tended to exhaust children. They say this was due to the large share of teacher-centred methods. Some teachers mention, however, that children were usually focused during these lessons and didn’t make noise, which is interpreted as a sign that they were genuinely interested. One of them puts it this way:

> But it must be said that all were really very engaged. The spectrum ranging from disinterested to very interested was smaller than in regular lessons. (T6)

Opinions differ with regard to the story-telling elements of ROE. While several teachers think that the contents of the stories were generally appropriate, others explicitly criticize these contents. Some of the stories, they say, were not age-adequate or not adapted closely enough to the school context. As one teacher expresses it:

> I think these should be stories about things that really happen to them, stories about peers and maybe also stories without pictures. It is hard for them to accept the picture books, because at their age they feel that they’re supposed not to be that child-

> like anymore… you know, just to admit that they like the book or the story is challenging for them. When it was a story from their everyday life, they got involved much more. (T5)

One teacher goes beyond questioning the stories' fit with the children's experiences and extends this assessment to pre-visits and post-visits in general. In this teacher's opinion, the heavy reliance on teacher-centered approaches results in too little attention being paid to personal experiences and individual questions from the children. This way, the teacher goes on, the children's potential to connect new information and skills with their everyday lives is not sufficiently activated, and the newly acquired knowledge tends to remain abstract. Other teachers do not mention that the connection to everyday life was missing in general, but several agree that children's interest had its ups and downs and that it usually surged as soon as some element of the lesson closely related to their everyday life.

The majority of teachers are convinced that children generally understood the contents of the program, above all those with regard to knowledge about babies and early infant development. They also say it was important that key messages were sometimes demonstrated through actions such as experiments. Interestingly, several teachers note that children were probably not aware of the fact that the ROE lessons were primarily about empathy. But they suspect that children understood this part "subconsciously," as one teacher puts it, and that it would take some time until they would develop a deeper sense of the program's core messages. One of the teachers expresses this as follows:

> If it all really clicks in ten years and they [the pupils] will realize in ten years' time why we did it back then, then we will have reached our goal. Because then, they will be mature enough. (T3)

In the group with instructors, participants report that children and teachers were looking forward to the ROE lessons. When the instructors came into the classroom, they experienced the atmosphere as pleasant and relaxed. Like the teachers, instructors agree that the lessons with the baby and the parent were particularly precious and important. They unanimously believe that the baby's participation is of great help in stimulating an intended "change in perspective." In line with the perception of teachers, a majority of instructors note that the teacher-centred method is demanding for children and that some children struggle to keep their attention focused during sequences that rely heavily on this "chalk and talk" approach. Their appraisal of the story-telling element is controversial. One instructor notes that children showed positive reactions to the stories, despite the fact that they sometimes seemed to be addressing younger children:

> Those stories really succeed in getting a discussion started… They can really delve into it, particularly fourth, fifth and sixth graders, they have also, yes… it is apparent that they really enjoy it when somebody comes to tell them a story, they are really able to engage, to get involved, I think. This works really well, also to take it up later when a topic is discussed. (I1)

Another instructor, however, reports a different experience:

> Books generally are a great tool for working with children, I think, to communicate things. But in this case it's books for smaller children, and when you take them to children in puberty, it can happen that they feel a little, how should I put it, like not taken seriously. (I2)

All instructors agree that stories are suitable for practicing changes in perspective and that they offer a solid base for in-depth discussions of topics. The stories as well as the worksheets are seen as providing a door into the realm of children's lived experience:

> I also experienced many situations in my class where we started with a worksheet, but this worksheet turned up situations that really happen in this class, and the discussion switched from the worksheet to the specific situation to what's currently going on in this class. And this was used as an admission ticket and we worked on this actual topic. Also with books (…) they helped facilitate this kind of process. (I3)

On a more general note, instructors say that the contents of the program are easy to teach and that they work with most of the children. Some perceive difficulties, however, in the transfer of theoretical notions into everyday life. This view is summarized in the following quote:

> On the other hand, connecting this knowledge to one's own experience, this is a different matter altogether… you know, for example, when you have a fight at recess with a friend, to remember in that situation, in the heat of that moment, what you learned moments before in the classroom, this is much harder to teach… I think it requires several additional steps until this actually gets to the level where it affects behavior… It's much easier to deal with this on a purely reflective level, with the baby, with the book, with the worksheet, where I am kind of standing outside of the heat of the moment and can keep a clear head about it all. (I3)

When the children are asked to assess the ROE lessons and their educational approaches, they immediately start to enthuse about the lessons with the baby:

> When the baby was here and it behaved in a funny was, I just really liked that. (C11)
>
> I just liked it when [baby's name] was here and above all, yes, when we were allowed to give him toys and he threw it back. (C20)
>
> And I was really looking forward to see [baby's name] and when we were allowed to give toys to him, he tried it and we wanted to see what he was doing with it. (C13)
>
> Because we observed [baby's name] quite closely and when we did not talk, we were just looking at him and he was so cute, he was so cute and now I am happy when I meet babies, because now I think they are cuter to me than before in a way. (C4)

Children agree that it was interesting and informative to be closely observing the baby's development. As this child puts it:

> So I think that it was fine that the baby came, because when a baby is here, then it is easier to learn about it… because when you don't know so much about a baby and you can see it directly how it feels and what it is doing, it is better than just to imagine it. (C5)

In contrast, several students consider the pre- and post-visits as rather boring at times, pointing out that they typically had to listen and fill in worksheets. The following criticism is expressed by one child:

> Because it was just boring when we had to be there and this and that and most of us we were just looking forward to see [baby's name] and not really cared for the other things, because it is just not so very interesting when you have to listen for a whole lesson. There were children for sure who liked that, but most of them were not really happy just to listen. (C13)

When asked about more engaging elements in the pre- and post-visits, children agree that the lessons were more interesting whenever they were given the opportunity to do something on their own (e.g., making a doorplate, collage), as the following quote shows:

> I liked it when we created a collage, newspapers, cut out scraps and wrote about our favorite animal and stuff like that. (C7)

The focus groups had different opinions about the stories. In two groups, there is a near-consensus: The children say that the stories were usually not age-appropriate, and some children make fun of the pictures in the books. Interestingly, however, most of the children in the third focus group had a positive view of the stories. This is expressed in the following statement:

> The stories were more for smaller children, but I liked that they were easy to understand (…) I think it was good that this story about bullying shows, it showed what it really can be like and this book showed how he [the character in the book] was feeling while he was bullied. This can make some children think, those who bully others. And I also liked that the girl helped and it just showed what reality is like. (C18)

This difference in opinion between the groups cannot be explained in terms of grade or age: The children in the group who generally appreciated the stories were, on average, neither younger nor older.

Suitability of the program depending on child and class characteristics

When asked whether they think that the suitability of ROE varies with grade level, approximately half of the teachers agree to 4th grade being an ideal level. At that age, teachers point out, children are able to reflect and to express themselves well while still inclined to childish behavior and quarrels. Also, the primary school system in the canton of Zurich is organized in such a way that children remain in the same class with the same teacher from grades 1 to 3 and are then reassembled into a newly composed class and assigned to a different teacher in grade 4, whereafter they remain together with the same class and teacher until grade 6. The formation of this new class in grade 4 is seen by some teachers as a perfect opportunity for ROE to step in because the program may shape class climate and culture in the state of its nascency. According to one teacher, implementing ROE in 5th grade has its own advantage, because children tend to become more interested in pregnancy at that age, often lacking even basic knowledge about it.

With regard to other child and/or class characteristics, the majority of teachers are convinced that the program is suitable both for "easy" and "difficult" classes and that a positive impact is achievable, in principle, in both cases. One teacher, however, thinks that the program might not be adequate for classes with acute problems, ones that have to be dealt with by a specific targeted proceeding. Reflecting on the question what characteristics of children ROE may be particularly suitable for, the teachers point out different aspects. Two of them hypothesize, for example, that children who are intellectually less gifted than others, different from others in their social behaviour, or generally less responsive to regular schooling, respond particularly well to the program. Two others disagree. According to them, the teacher-centred approach in the pre- and post-visits makes it easier for intellectually gifted children to follow. There is disagreement, too, on whether children with or without the experience of having a younger sibling benefit more from the program. In terms of gender, teachers unanimously think that girls and boys are, in principle, equally responsive.

The instructors unanimously stress that ROE is equally suitable for all children independent of age, other individual characteristics of composition of the class. The program, they point out, provides an opportunity for all children to see and experience themselves and each other from a different perspective than usual.

One instructor explains the general difficulty of assessing which children might be more responsive than others in the following way:

> I think it is incredibly difficult to tell which child you have been able to make a connection to and which you haven't. I made the experience so often that at first I thought, with regard to certain children, that we weren't really warming up to each other or that they felt uncomfortable having a class with me (…) and then it was these same children that came running across the schoolyard toward me (…) and they embraced me and I was thinking, wow, I would have thought that I had not made a connection with them (…) And sometimes I have children in school who never say a word, but they're always around me when I'm there, clinging to me. Apparently, I am able to give them something, something they don't normally get. And I think this is a substantial impact, even though of course I can't put my finger on what exactly is happening to this child. But my being around is good for them. (T3)

Appraisal of facilitating and obstructive factors

After discussing the details of the implementation, focus groups turned their attention to the topic of conditional factors that had either been beneficial or obstructive to the successful deliverance of the program. In their response, approximately half of the teachers identify a good collaboration between parent, instructor, teacher and the pupils as one of the central factors for a successful outcome. They also point to the importance of a positive atmosphere in the classroom. The latter is seen as dependent on a teacher who is able to lead the class and to provide support to the instructor. In addition, it is seen as a strong advantage if the teacher strongly backs the project and knows how to integrate its contents into his or her own lessons. One teacher specifies the requirements for the parent and lists several characteristics that are seen as disadvantageous: speaking dialect but not standard German (which is conflict with the fact that many children understand only standard German), finding it difficult to cope with the pupils, or being unreliable (e.g., inclined to arrive late). One teacher thinks it is an advantage if a participating parent already knows the children or has a relation to the school:

> For sure, I thought it was great that we had a parent who had a previous connection to the school [having worked as a school social worker there], because this influenced the children's relationship with the school and the baby on so many levels. I feel this had a much greater effect than could have been the case if it had been somebody else, somebody you don't know and who has no connection to the class. (T5)

Several teachers point to a flexible and reliable instructor as a crucial factor. In addition, there is agreement that, whenever possible, ROE lessons should be scheduled in the morning, when children find it easier to focus.

When asked about facilitating and obstructive conditions of the program's implementation, instructors, like teachers, emphasize the importance of a working collaboration between all the adults involved. They attach particular importance to the collaboration with the teacher. In this context, clarifying the division of roles and responsibilities between teacher and instructor is seen as imperative. One example for this division is that, according to instructors, teachers and not instructors should be responsible to keep up the discipline in class during ROE lessons. Instructors agree that the impact of the program depends substantially on the value and importance the teacher attaches to it and the support the teacher provides in transferring the new knowledge into children's everyday lives. With regard to the parent, one instructor makes the point that the person should be authentic and well connected to their own feelings, but at the same time not too sensitive. Instructors agree that if the parent has an extravert personality and/or is used to working with children, the lessons may be easier for the instructor, but working with an inexperienced or more guarded parent is seen as very well manageable, too. Concerning the baby, one instructor points out that an active and expressive baby will encourage the class to be particularly engaged in interacting with each other and the baby, whereas a more calm and reluctant baby requires a more active role from the instructor. As another factor beneficial to the quality of the program, several instructors point to the possibility of contacting a Mentor in Canada for support. Instructors made use of this offer several times and, in hindsight, consider it to have been very useful. They also found it helpful to be able to contact the person responsible for ROE in Switzerland anytime, either by phone or email.

Ideas for improving the implementation

When teachers were asked what opportunities they saw for improving the implementation quality of the ROE program, none of them came up with suggestions concerning the family visits. With regard to the pre- and post-visits, a majority of teachers said that the approach might be less teacher-centred at times, with more variation promoting a stronger involvement and engagement of the children. Some teachers make specific proposals in this regard, like the ones in the following quotes:

> For example, a topic could be introduced with a question that everybody has to answer, writing it down, and then this could be discussed in small groups of five children, and then back to discussion with the whole class. To conclude, a writing or listening task or a scenic task could follow, yes, something like this. (T6)
>
> Role-playing is something that the children like to do, this way they can feel their way into certain situations. Or they might perform or present something, so that they have to deliver a little more. (T4)
>
> Or in the sense of… that children engage with the topic on their own before they are told something about it (…) That they get the instruction, for example, in the case of a baby who has trouble sleeping, that they go and ask their parents what they would have done about it, back in the days (…) In preparing for the discussion, they could go on their own and do research (…) This way, their interest is sparked. (T5)

This last quote presages another, more general suggestion several teachers make when talking about pre- and post-visits: that the contents should connect more closely to children's reality, their lived experience. To achieve this, some propose a change in the choice of books, particularly with regard to the age-group the books are targeting.

Like the teachers, the instructors make no specific suggestions for improving the lessons with the baby and the parent. There is disagreement about what changes in pre- and post-visits would benefit the program. Confronted with the suggestions the teachers had made, not everyone agrees that more variation in the learning approaches or a different selection of stories would lead to a better fit. One instructor, for example, thinks that sometimes using a book that portraits a somewhat different reality from the one the children are experiencing poses a challenge and is therefore a suitable learning field. One instructor would prefer to have fewer tasks in the form of worksheets, in exchange for more interactive elements such as role plays or interviews in the streets. Further, one instructor suggests more time for open discussions and an increased support of pupils in transferring theoretical insights into everyday life.

The children are the only group that comes up with suggestions regarding the family visits as well. In one focus group, several agree that the program would be enhanced if students were allowed to do have more interaction with the baby. This applies to physical contact with the baby such as carrying the baby around (instead of "merely touching its feet," as one child puts it) and to ideas such as that children should be allowed to bring the baby their own toys or to come up with their own experiments (the proposals here range from ball games to online gaming on smartphones). Concerning the pre- and post-visits, the children contribute quite specific proposals: such as not to fill in the poster of what the baby is able to do every time, including more games, going outdoors when the weather is good, or not having lessons extend into recess. The common thread in most of these proposals is a demand for more interactive and more flexible forms of teaching and learning. Further, in two of the groups, children voice ideas that resemble the teachers' suggestions of bringing the contents closer to the children's own reality. They suggest that conversations should more frequently be focused not on the baby alone, but also on the opinions and experiences of the pupils. This, many say, could be facilitated through a different choice of books:

> There are books that would have been more exciting. But they will never read these to us. For example, Gregs Tagebuch [Diary of a Wimpy Kid]. Everybody here liked to read it, and everybody has at least read three of them. (C15)

Finally, some children suggest that the lessons should be more responsive to what's bothering the class at the time. The following quote captures this view:

> Or like when she [the instructor] asked if we were having problems in our class, and we said that there were some, that a child is being bullied by others… then it would have been nice to talk about it in the whole class, about the reason why. (C2)

3.2.2 Impact

The teachers agree in unison that their students learned a great deal about babies and about infant development through ROE. Beyond that, they think their pupils learned how to interact appropriately with babies. As a consequence, according to one teacher, they gradually ceased being shy about babies and became more understanding of them and their families. The majority of teachers are convinced that ROE was successful in making children reflect more thoroughly on their own emotions. One teacher notes that this extended well beyond the scope of the program's lessons and into regular class as well, and points out that the topics of prosocial behavior and empathy were regularly talked about in her class. Other teachers agree. Because of this "spill over" effect, more than half of the teachers are not sure, however, whether changes in empathy or prosocial behavior in their class may really be directly attributable to ROE. The following quote illustrates this:

> But I can't say it is because of Roots of Empathy, I may just say that Roots of Empathy contributed to it… But how much with regard to any individual child, that is impossible to say and not measurable. (T6)

Most teachers assume that, if the program has led to positive and long-lasting change in empathy and prosocial behaviour, this change will be mostly on a subconscious level and hardly observable. Three of the teachers say they have noticed a heightened sensitivity among their students for their own feelings and those of others, and they speculate this might have led to improvements in recognizing, naming and talking about feelings. The following quote exemplifies this:

> Yes, I think most of all it is sustainable, this matter about feelings, talking about feelings and naming them, what is it exactly that I feel, is it anxiety or timidity or whatever, those nuances. And generally, getting to know about the whole range of emotions that exist. (T1)

One of the teachers believes to have noticed that there have been fewer arguments, fewer fights and complaints in her classroom since ROE. Another thinks that the program is supportive in settling disputes among students:

> When there had been an argument, we discussed it and the way we approached it was like, if you would never treat [baby's name] that way, why then do you do this to your classmate? In such ways, the children could at times be made aware of something from the program, in regular classes. (T2)

Two teachers have their doubts about how much the program really was able to foster children's empathy and prosocial behavior. Three of them say that the program exerted a positive influence on their class as a whole. ROE is perceived to have created a positive mood and to foster the team spirit in the class. The other half, however, did not observe such an influence, with two of the teachers suspecting that their classes already had a good team spirit in the beginning, which didn't leave much to improve on.

Finally, all teachers say they are sure their pupils will remember the baby's visits for a long time and will be able to retain much of their new knowledge on infant development. Or as one teacher expresses it:

> It can be said that there was a good basic mood and that the kids liked to come to school and they were really looking forward to seeing the baby, and this fact leaves in its wake good thoughts in their memories. This is something very precious in the development of any child. (T6)

According to the instructors, the program does not aim at a measurable and directly visible effect, but is focused on the children's growth and development resulting out of this experience. It is supposed to generate an open atmosphere in which children are able to show their feelings, communicate them, and feel appreciated for who they are. In the ROE lessons, as instructors see them, children are able to lose their emotional burden and learn that other people undergo similar situations and feelings, which produces a noticeable relief. Getting in touch with their own feelings has, instructors note, a calming and relaxing effect on the children, and this may be directly experienced during the lessons. The program is perceived to enable children to gain positive experiences, including positive relationship experiences. One instructor adds that, in the course of the year, children learn to listen to their inner voice, to open up, to embrace their own positions and feelings and to accept the positions of others. This is observed, for example, in how ordinary differences between children are dealt with and reconciled. The instructors are in agreement, however, that in many regards the changes initiated by the program will hardly be visible to an observer because the program aims at fostering children in their own very personal and individual development. Therefore, children and teachers will often not be aware of this change even while it is happening. The program is believed to initiate a variety of processes which, instructors suspect, will unfold only in the long run. As one of them elaborates:

> It's so important that we don't judge but invite the children to decide for themselves whether something is good for them. For example, when a child hands in a drawing or something, and most children ask is it right the way I did it? and then we ask back is it good for you? That sets something off, right. That starts an internal process in this child, like getting away from external judgements to something like, how is it actually for me? I also think that all those impulses will probably come to fruition much later but, like I said before, it doesn't take a lot to plant seeds that will eventually start to grow some day. I think that all these elements, they matter as much for teachers as for the students (…) Because there have been many feedbacks over the years now, where teachers have asked me, how do I do this? (T3)

As anticipated in the last part of this quote, several instructors point out that the encouragement to engage with one's feelings can be educational for teachers, too. Besides, instructors note, observing their pupils in a different context can encourage teachers to see pupils in a fresh light and thereby, in some cases, help to improve the relationship between teacher and pupils.

Finally, when the children are asked what they had learned through ROE, the first topic they raise is their knowledge about infant development. They recall (and readily lecture the group moderator on), for example, how newborns develop vision in the early months of their lives, how they are not able to hold up their heads by themselves, that they have a soft spot in their skulls, or that they feel emotions from the first moments of their lives. They also recall do's and dont's concerning physical interaction with babies, such as never to shake a baby, have it sleep on its back, or change its diapers, and emotional interaction, such as the importance of love, of responding to feelings, reacting appropriately when the baby starts crying, or encouraging and supporting it in discovering the world. According to children's statements in two groups, the family visits inspire them to face up to their personal experiences and the experiences of others. The following quote gives an example of this view:

> In a way [baby's name] just like showed us that we are able to change, that, for example, you think I don't like this and then suddenly you start to like it nevertheless. He showed how humans develop, showed us how development unfolds, that some love broccoli, others don't. This was shown to all of us and I learned that humans can always change. That was also the topic of the story. (C13)

When it comes to the domains of emotion and social behavior, the children have divergent opinions about the program's impact. In two groups, there is some agreement that nothing much has changed—or if so, just a little. This applies to the individual level as well as to the class as a whole. The following quote by a child captures this appraisal:

> Well, I think not really… not a lot has changed. Well, in fourth grade we were still good, but in fifth it was not good in the first half of the year until Christmas, because we were a little older and we did a lot of mischief, and I think that it did not really

> help much, just the thing with the feelings and with the other people. But concerning the class spirit, it didn't really… most of us just focused on the baby. (C9)

Despite this skeptical appraisal, there is evidence in all three groups, from the children's own perspective, that some change has in fact taken place. For example, in both of the focus groups where a negative appraisal prevails overall, at least one child disagrees and reports that he or she has become more sensitive toward his/her own and others' feelings through the program. Beyond that, when the question "Do you think the program has changed something about how you deal with each other in your class?"—which the children tend to answer in the negative—is rephrased as "What do you think you learned about emotion and social behavior in the program?", the participants in all three focus are not at a loss to answer; rather, their responses are immediate and plentiful. In all the groups, there are several participants who report on some change concerning awareness and understanding of emotions, and they go on to say that this understanding influences their social behavior in different areas as well. The following statements exemplify this:

> You know, I learned that, when you know how the person feels at that moment, you are able to communicate better, because if this person might be angry, then you may try to calm her down or… when she is sad, like that. (C11)

> A friend of mine (…) is quite a happy person and when she suddenly turns more quiet and walks around sadly and in a way doesn't talk much, then I do ask her if something happened and that she can tell me. If she doesn't say anything, I ask if she is tired or so. You can notice when somebody is tired, because when somebody is tired it is similar to being sad, in that case people don't say that much, either. (C14)

> For example, when I say something to somebody, then I think sometimes, I should perhaps not have said that, because it makes you feel sad. It was stupid that I said this, I could just have ignored that situation because it didn't do me any good if I offend somebody else, except that this person feels bad and I actually don't want this to happen, either. (C3)

In addition, some of the children point out changes in their social behaviour that relate to the domains of emotion regulation and impulse control:

> That you don't respond so aggressively when somebody says something wrong, that you don't freak out and want to hit that person, because this [saying something wrong or insulting] can happen to anybody. (C13)

> To be nice to other people, to not instantly snap. To listen to them, what they say… I've become more kind. (C11)

4 Discussion

The quantitative data and analyses presented in this report suggest that the Roots of Empathy program, as carried out in the years 2015 to 2017 in the canton of Zurich, was successful in bringing about desired effects in all three domains that were central to the program. The effect sizes, calculated as Cohen's *d*'s, range from 0.34 for empathy to 0.5 for aggression. In two of these realms, empathy and aggression, effects were retained one year after the program had been completed, and the effects sizes for these domains remained the same or even slightly increased (0.47 and 0.46, respectively). According to a statistical convention going back to Cohen (1988), such effect sizes are considered to be in the small to moderate (0.2—0.5) range. When compared to effects usually found in the research literature on social and emotional learning programs, the ones found in the present study fare well in statistical terms. When more specifically compared to previous findings on the ROE program in other countries (Santos et al., 2010; Schönert-Reichl et al., 2012; see chapter 1), they are in a similar range.

This assessment of the results is based mainly on the composite measures on all three outcome variables, measures that integrated children's perceptions of themselves, children's perceptions of their peers, and the teachers' perceptions of their pupils. As mentioned in the Methods section above (see chapter 2), such an aggregation of data into composite measures is often recommended in the literature, and its absence is frequently critized as a serious flaw in many research designs (e.g., Kagan, 2013; Van der Ende, 1999). Therefore, it seems reasonable to rely on these measures and to foreground them in the interpretation of results.

At the same time, it has to be noted that the strong suggestion of intervention effects in the quantitative portion of this study depends largely on the teachers' assessments. Therefore, a contamination of these assessments—e.g., via a social desirability bias or strong outcome expectations—would seriously distort the results and invalidate the conclusions. How probable is such a contamination? On the one hand, it is well-known that participants in intervention programs often are emotionally invested in the successful outcome of the intervention and may thus, without necessarily being aware of it, distort their perceptions towards a more favorable interpretation of observed phenomena (see Gawronski, 2012, for an interpretation of this mechanism in terms of cognitive dissonance theory). In the case of the present analysis, it is possible that teachers in the ROE classes expected their students to make gains in the relevant aspects of behavior and then went on to observe what they expected to observe. However, there are two factors that limit the strength of this argument. First, it should be pointed out that teachers in the quantitative part of our study were never asked whether they thought their students had become more empathic or less aggressive etc. globally and in hindsight, but rather they were asked to rate each individual student's recent behavior in typically very concrete terms (such as "When this child gets angry at another child, it pushes him or her"). Such pretest-posttest differences in concrete behaviors, although certainly not entirely immune, are better protected against serious distortions than retrospective and global assessments are. Secondly, as we saw in the qualitative analysis in chapter 3.2, when teachers were explicitly asked about whether they had observed any desired effects of the ROE program in their students, many answered that they were not sure, that changes could not be attributed with certainty to ROE or that they thought such effects, if they existed, would only become apparent in the long run and would hardly be measurable. It seems possible, but not highly probable, that teachers systematically distorted their ad hoc ratings but did not distort their retrospective evaluations in an interview-setting.

Assuming that the teachers' assessment can be taken at face value, this leaves the question why the pupils' self-reports and peer-nominations do not indicate strong effects. There are several possibilities. Provided that the teacher and the pupil measures do indeed tap into the same phenomena, it is important to recall that any measure relying on subjective reports in the social sciences is a function not only of the relevant aspect of reality, but of subjects' proclivities towards perceiving reality, recollecting their perceptions, and reporting their recollections. Perceptions, recollections and reporting behavior do not necessarily mirror reality in exactly the same way across different measurement times and environments. Along the lines of such an interpretation, it may be argued that the children in our sample assessed levels of empathic, prosocial and aggressive behavior not as measured against an absolute standard, but in the context of how they perceived their peers to behave. In this case, a child's self-report would reflect the relative rank of that child's behavior, as perceived by herself against the standard set by how she perceives her classmates concurrent behavior, and not some absolute value. If that is true, then one would expect average scores (reflecting perceived rank) in the sample not to change considerably between one point in time and another even if the same behaviors, when measured in absolute terms, actually showed a considerable increase or decrease during the same period of time. Researchers in the social sciences are generally well aware of such a potential bias, and this awareness is the reason why questionnaire items are usually not targeted at some relative assessment of a general behavior (such as, "How aggressive do you think you are when comparing yourself to other students in your class?"), but at an objective assessment of a concrete observable behavior (such as, "How often do you hit other students in your class?"). But while this strategy was adopted in our study to a large degree as well, it may not have completely eliminated the bias.

Another, perhaps more straightforward explanation of the contrast between teachers' and pupils' evaluation is that children in the ROE classes might have become more sensitive towards noticing aggressive behaviors as a result of the program; this may then have led to a stronger tendency to report these behaviors. This interpretation is in line with the fact that students in the ROE classes reported significantly lower levels of aggression than their peers in the control classes at pre-testing, despite the fact that their teachers at the same reported higher levels of aggression. In this way, the increase in the children's proclivities to notice and report aggression may mask an intervention effect; a decrease in aggression and an increase in the proclivity to report may have canceled each other out. Interestingly, such an interpretation would provide a fresh perspective on one of the quotes from the qualitative analysis, which we repeat here:

> Well, I think not really… not a lot has changed. Well, in fourth grade we were still good, but in fifth it was not good in the first half of the year until Christmas, because we were a little older and we did a lot of mischief, and I think that it did not really help much, just the thing with the feelings and with the other people. But concerning the class spirit, it didn't really… most of us just focused on the baby.

The admission of this child that his class had changed from “being good” in forth grade to “doing a lot of mischief” in fifth grade, which corresponds to the time ROE was implemented, may of course track a truly negative development during that time; it may also, however, be the product of shifting perceptions, of a heightened sensitivity towards perceiving certain behaviors and interactions as mischievous or “bad” in the first place. This explanation is speculative, and it does not apply equally to the domains of empathy and prosocial behavior. But the contrast between teachers’ reports and pupils’ self-reports is not as stark in these domains, as can be seen from Table 5 in the Results section.

Another more general potential explanation of the pattern of results found in the current study is the phenomenon often labelled as regression to the mean (e.g., Barnett, Van Der Pols, & Dobson, 2004; Bland & Altman, 1994). This “regression effect” means that baseline differences between groups in an observed variable, where one of these groups was selected with the intention of reducing extreme baseline values in that variable, will tend to level out over time even in the absence of any intervention. Since the two groups compared in this study did indeed (despite the matching procedure) differ significantly in the domains of empathy and prosocial behavior at baseline, the assumption of a regression effect could offer an alternative explanation why the change scores in the ROE group were signifantly more positive than those in the control group: ROE pupils might have regressed to the mean. What limits the scope of this explanation, however, is the fact that an effect size in the same range as for the other two domains was observed for aggression, where no significant differences between the groups were detected at baseline in the composite measure and where children in the ROE group showed lower levels than their peers after completion of the program. Also, we found that the initial differences between the two groups in terms of pupils’ self-reports on aggression (where children in ROE classes had reported *lower* levels at baseline) remained stable over time. The assumption of a regression to the mean, however, would have predicted that these differences will level out.

Beyond answering the question whether ROE works, the current study has also shed some light on how it may work. In line with previous investigations both in the literature on ROE and in the literature on empathy more generally, we have found some support that empathy does indeed play an important causal role in bringing about prosocial behavior, altruism, and the eschewal of aggressive behaviors. In particular, we found that the change observed in prosocial behavior was to a modest degree (explaining roughly 30 % of the variance) correlated with changes in empathy, which is in line with the theoretical assumption that empathy causally contributes to (perhaps via induction of empathic concern or sympathy) prosocial behavior (rather than vice versa, which would be a statistically valid explanation of the data but not a theoretically plausible one). The explorative inquiry into this causal pathway presented in the current study has its limitations, however, and the dataset will surely lend itself to a more sophisticated re-analysis of this matter in the future.

On the other hand, the study has also contributed evidence to the question of how ROE works by showing how it apparently does not work. Although ROE seems to have been successful in reducing children’s aggression to a similar degree as it was successful in increasing empathy, the two mechanisms appear to be largely unrelated, because changes in empathy account for only about 5% of changes in aggression. This leaves the question what other causal paths might be responsible for the change, with emotion regulation and self-control being likely candidates, although the proxy measure for self-control used in this study (teacher-rated hyperactivity) yielded no such results. Beyond that, the study also showed that the recognition of emotion expression in faces did not improve under ROE, suggesting that mere recognition of emotions from facial cues does not play an important role in bringing about the observed gains in empathy overall. A plausible hypothesis here is that ROE does not achieve increases in empathy by teaching children how to recognize emotions they did not recognize before, but by encouraging them to being more attentive to others' emotions or, put differently, to use their capacity of recognizing emotions (and to shape their behavior based on that) more actively. In other words, it might not be the potential for empathy that is improved, but rather the extent to which this potential is used.

With regard to the behavioral measures considered in this study, we found, intriguingly, that ROE slightly increased the likelihood that children would act altruistically towards members of their own class, when compared to children from the control group. It did not, however, influence the likelihood that they would act altruistically towards a stranger, an unknown child from a different class. At first sight, this finding seems to give some support to critics of empathy, such as Bloom (2016, 2017), who point out that empathy, which is considered a limited resource, will lead us to favor certain people (those we feel close to and therefore empathize with) to others (whom we do not feel close to and therefore do not empathize with) in ways that are clearly unfair and immoral. It needs to be pointed out, however, that we found no indication that the pupils from ROE classes behaved less altruistically towards strangers than those from the control classes; the increased altruism towards their classmates was in no way compensated for by a decrease in altruism towards strangers. It would certainly be interesting in future research to create scenarios in which one cannot be had without the other: where an altruistic choice towards one's in-group would inevitably lead to a decrease in altruism towards one's out-groups. It would also be interesting to examine whether there are phenomena that moderate the relationship between a child's empathic abilities and a possible in-group favoritism (such as attitudes towards one's group, which might be negative or neutral as well as positive, depending on the nature of experiences with the in-group). Until such relations are investigated, it would be unsound to conclude that our results from the Trust Game point to any moral flaw in empathy.

A very relevant question that comes up in the discussion of any intervention study is what the estimated effects actually mean in the real world. A statistical effect size like Cohen's *d* tells us how large an effect in one group is relative to the effect in another group. It tells us nothing per se, however, about the practical significance of this difference—whether it literally makes a difference in the real world. One way to assess the practical significance of the effects found in the current study is to look at the actual change between pretest and posttest scores. As was mentioned above, classes in the ROE group started out at baseline with lower average levels of empathy and prosocial behavior and higher average levels of aggression, as indicated by the composite measures integrating all data sources. After the completion of the program, students in the ROE classes were, on average, slightly more empathic, slightly less aggressive and equally prosocial when compared to their peers in the control classes. This means that ROE was altogether successful in either alleviating or even turning around deficits that had previously existed. None of the outcome measures skyrocketed, however; after the completion of ROE, there were students in the ROE classes who were clearly less empathic, for example, than some of their peers in the control classes. This is indicated by the fact that the distributions in the two groups, which may be inferred from the reported means and standard deviations at post-testing (see Table 5), still overlap (with a Cohen's *d* of 0.5 and assuming normal distributions, this overlap is about 80%).

Whether an effect of this magnitude warrants the implementation of the ROE program in Swiss primary schools is not for the researchers to answer—this decision is reserved for practitioners. What can be said from the empiricist's perspective, however, is that Roots of Empathy is one of the very few social and emotional learning programs, particularly in the Swiss context, for which desired effects have now been shown in a scientifically rigorous design. For educators interested in strengthening the empathic and prosocial capacities in primary school students, the program seems therefore very much worthy of serious consideration.

Beyond this quantitative perspective, our qualitative inquiry into the implementation and impact of ROE yielded additional and often supplementary insights. For one thing, all groups mention (and the children impressively demonstrate) an increase in knowledge about infant development as an outcome of the program. With regard to the implementation, the lessons with the baby and the mother are unconditionally appreciated across all three groups of participants, i.e., teachers, instructors, and children. The family visits are not just seen as pleasurable and exciting for the children, but all groups think they are an effective tool for fostering learning, too. In comparison, the opinions about the pre-visits and post-visits are much more controversial, and they vary as much within each group as between the groups. Beyond dispute is the notion that teacher-centred methods are the dominant learning approach in the pre- and post-visits. Most often, the instructor explained a topic to the whole class followed by a discussion, or read a story that was subsequently discussed in plenary. Other approaches, such as individual exercices and tasks, experiments or group work, were not completely lacking, but they were perceived as the exception to the rule. In all groups there are participants who consider the combination of learning approaches used in the program as appropriate. Others, however, disagree, arguing that teacher-centred methods should be complemented more frequently by interactive approaches. This view is emphasized in the groups of children and teachers, but receives some support among instructors as well. The argument is that teacher-centred methods, if they become too dominant, exhaust childrens' attention and focus, and do not encourage children's own initiative and activity. Or, in the words of the children, they can be plainly boring.

Although considered as a teacher-centred method as well, all groups appreciate in principle the approach of reading and discussing stories, with several participants pointing out that children like to be told stories and find it easier to reflect on theoretical topics if they are introduced in the vivid format of narrative. Also, there are participants in all groups who formed a positive opinion about the particular storybooks used in the program. Several others, however, make the point that many of the stories were not age-appropriate (targeting younger children than the ones enrolled in ROE) and not matching closely enough with the daily realities and lived experiences of the children. This last criticism is extended, by some participants both in the teacher and the children groups, to the pre- and post-visits in general: they feel that the visits offer too little opportunity for considering and coming to grips with the actual experiences of the children and the class as a whole.

Summarizing ideas for improving ROE in the Swiss context, it is important to distinguish between suggestions that were made by several participants from different groups and those that were made by only one or very few individuals or within only one group. The suggestions that were brought up most often and received most support across all groups (but particularly in the teacher and the children groups) concern the pre- and post-visits. They fall into three broad categories: i) more interactive and flexible approaches to teaching and learning in general, ii) closer consideration of children's own realities and lived experiences; and iii) more age-approriate choices for the storybooks. Specifically, participants suggest more group work and more role-playing. According to several of them, an increase of such methods would make the visits more attractive for children, and the children's realities would be paid more attention to. Children would be more active, which might provide better opportunities to transfer newly learned knowledge and skills into their everyday lives.

5 Conclusion

The study presented in this report has several strengths. Unlike most previous research on classroom-based social and emotional learning program interventions in Switzerland and elsewewhere in continental Europe, it was based on a quasi-experimental pretest-posttest design including a control group carefully matched on key sociodemographic factors. It included a sample which provided adequate statistical power to detect even small-scale effects. Outcome variables were broadly assessed using different data sources from multiple informants, which constitutes a desirable departure from the reliance on single sources such as teachers’ or pupils’ reports. This provided us with the opportunity to integrate several different perspectives into composite measures. The study also incorporated a follow-up measurement to test for long-term effects, and the implementation quality and impact were investigated through an accompanying qualitative design, which allowed us to elucidate the participants’ subjective perceptions, experiences and evaluations with more detail than would have been possible in a merely quantitative approach.

At the same time, the design of the current study clearly could have been improved by using an appropriate randomization procedure for allocating classrooms to study group. This was not possible for practical reasons. Although our matching procedure was altogether successful in balancing the two study groups in terms of sociodemographic factors at the level of school, classroom, children and teacher, it did not prevent significant differences in key outcome variables at baseline. Also due to practical reasons, it was not possible in this study to mask researchers, teachers or pupils against the treatment condition. Therefore, a contamination of our results due to social desirability biases, favorable outcome expectations or regression effects cannot be completely ruled out.

To conclude, the current study suggests that ROE is an effective tool for fostering empathy and prosocial behavior and reducing aggression among primary school pupils in the Swiss educational system. The calculated effect sizes surpass those typically reported in the literature for successful social and emotional learning programs, and they are retained one year after completion of the program in two of its three key outcome domains. With regard to empathy, the domain that gives the program its name, ROE raised the likelihood that a child would increase his or her empathic capabilities during the school-year. This is a significant and substantial impact. At the same time, our results also show that there was still considerable overlap in terms of empathy, aggression and prosocial behavior between children who had participated in the program and those who had not. This makes clear that the development of children’s emotional and social skills depends on other factors as well, ones that no single intervention of limited duration can affect all at once. The task of supporting and encouraging children in their social and emotional development rests on individuals, families, professionals, and society as a whole.

References

Aguilar-Cárceles, M. M., & Farrington, D. P. (2017). Attention deficit hyperactivity disorder, impulsivity, and low self-control: which is most useful in understanding and preventing offending? *Crime Psychology Review, 3*(1), 1-22.

Arsenio, W. F., & Lemerise, E. A. (2004). Aggression and moral development: Integrating social information processing and moral domain models. *Child development, 75*(4), 987-1002.

Averdijk, M., Zirk-Sadowski, J., Ribeaud, D., & Eisner, M. (2016). Long-term effects of two childhood psychosocial interventions on adolescent delinquency, substance use, and antisocial behavior: a cluster randomized controlled trial. *Journal of Experimental Criminology, 12*(1), 21-47.

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J. (2004). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology, 34*(1), 215-220.

Baron-Cohen, S. (2011). *Zero degrees of empathy: A new theory of human cruelty*. London: Allen Lane.

Barth, J.M., Dunlap, S.T., Dane, H., Lockman, J.E. & Wells, K.C. (2004). Classroom environment influences on aggression, peer relations, and academic focus. *Journal of School Psychology, 42*, 115-133.

Batson, C. (1991). *The altruism question: towards a social-psychological answer*. Hillsdale, NJ: Erlbaum.

Batson, C. (2008). These things called empathy: Eight related but distinct phenomena. In J. Decety, & W. Ickes (Eds.), *The social neuroscience of empathy* (pp. 3–16). Cambridge, MA: MIT Press.

Batson, C. D., & Ahmad, N. Y. (2009). Using empathy to improve intergroup attitudes and relations. *Social Issues and Policy Review, 3*(1), 141-177.

Baumeister, R. F., Heatherton, T. F., & Tice, D. M. (1994). *Losing control: How and why people fail at self-regulation*. San Diego, CA: Academic Press.

Bayrami, L. (2017). Roots of Empathy program: Cultivating sensitive and responsive parenting. *International Journal of Birth and Parent Education, 4*(4), 16-18.

Baxter, S. D., Smith, A. F., Litaker, M. S., Baglio, M. L., Guinn, C. H., & Shaffer, N. M. (2004). Children's social desirability and dietary reports. *Journal of Nutrition Education and Behavior, 36*, 84-89.

Bland, J. M. & Altman, D. G. (1994). Statistic notes: Some examples of regression to the mean. *British Medical Journal, 308*(6942), 780.

Bloom, P. (2016). *Against empathy: The case for rational compassion*. New York, NY: Random House.

Bloom, P. (2017). Empathy and its discontents. *Trends in Cognitive Sciences, 21*(1), 24-31.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Crick, N. R. (1996). The role of overt aggression, relational aggression, and prosocial behavior in the prediction of children's future social adjustment. *Child Development, 67*(5), 2317-2327.

Crick, N. R. & Grotpeter, J. K. (1995). Relational aggression, gender, and social psychological adjustment. *Child Development, 66*, 710-722.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10*, 85-90.

De Winter, J. C. & Dodou, D. (2010). Five-point Likert items: t-test versus Mann-Whitney-Wilcoxon. Practical Assessment, *Research & Evaluation, 15*(11), 1-12.

Dodge, K. A. & Coie, J. D. (1987). Social-information-processing factors in reactive and proactive aggression in children's peer groups. Special issue: Integrating personality and social psychology. *Journal of Personality and Social Psychology, 53*(6), 1146-1158.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*(1), 405-432.

Eisenberg, N. & Fabes, R. A. (1990). Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion, 14*(2), 131-149.

Eisenberg, N. & Morris, A. S. (2001). The origins and social significance of empathy-related responding. A review of empathy and moral development: implications for caring and justice by ML Hoffman. *Social Justice Research, 14*(1), 95-120.

Eisenberg, N., Shea, C. L., Carlo, G., & Knight, G. (1991). Empathy-related responding and cognition: A "chicken and the egg" dilemma. In W. Kurtines & J. Gewirtz (Eds.), *Handbook of moral behavior and development: Vol. 2. Research* (pp. 63-88). Hillsdale, NJ: Erlbaum.

Eisenberg, N., Spinrad, T. L., & Morris, A. (2014). Empathy-related responding in children. In M. Killen, & J. G. Smetana (Eds.), *Handbook of Moral Development* (2nd ed.) (pp. 184-207). New York: Taylor & Francis.

Eisner, M. & Ribeaud, D. (2005). A Randomized Field Experiment to Prevent Violence. The Zurich Intervention and Prevention Project at Schools, ZIPPS. *European Journal of Crime, Criminal Law and Criminal Justice, 13*(1), 27-43.

Elias, M. J., Zins, J. E., Weissberg, R. P., Frey, K. S., Greenberg, M. T., Haynes, N. M., et al. (1997). *Promoting social and emotional learning: Guidelines for educators.* Alexandria, VA: Association for Supervision and Curriculum Development.

Fehr, E., Fischbacher, U., Von Rosenbladt, B., Schupp, J., & Wagner, G. G. (2003). A nation-wide laboratory: examining trust and trustworthiness by integrating behavioral experiments into representative survey. *IZA Discussion paper series, No. 715*. Available at: https://www.econstor.eu/bitstream/10419/20527/1/dp715.pdf

Feshbach, N. D. (1979). Empathy Training: A Field Study In Affective Education. In S. Feshbach & A. Franczek (Eds.), *Aggression und behavior change: Biological und social processes* (pp. 234-249). New York, NY: Praeger.

Feshbach, N. D. & Feshbach, S. (2011). Empathy and education. In J. Decety & W. Ickes (Eds.), *The social neuroscience of empathy* (pp. 85-98). Cambridge, MA: MIT Press.

Finkenauer, C., Buyukcan-Tetik, A., Baumeister, R. F., Schoemaker, K., Bartels, M., & Vohs, K. D. (2015). Out of control: identifying the role of self-control strength in family violence. *Current Directions in Psychological Science, 24*(4), 261-266.

Flick, U. (2006). *Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen.* Reinbek bei Hamburg: Rowohlt.

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition, 30*(6), 652-668.

Golan, O., Baron-Cohen, S., & Hill, J. (2006). The Cambridge mindreading (CAM) face-voice battery: Testing complex emotion recognition in adults with and without Asperger syndrome. *Journal of Autism and Developmental Disorders, 36*(2), 169-183.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*(5), 581-586.

Gordon, M. (2001). Roots of Empathy. *Canadian Children, 26*(2), 4-7.

Gordon, M. (2003). Roots of Empathy: Responsive parenting, caring societies. *The Keio Journal of Medicine, 52*(4), 236-243.

Gordon, M. (2007). *Roots of Empathy: Changing the World, Child by Child*. New York: Workman Publishing.

Gordon, M., & Green, J. (2008). Roots of Empathy: Changing the World, Child by Child. *Education Canada, 48*(2), 34-36.

Hoffman, M. L. (1994). The contribution of empathy to justice and moral judgment. *Reaching out: Caring, altruism and prosocial behavior, 7*, 161-194.

Hoffman, M. L. (2000). *Empathy and moral development: Implications for caring and justice*. New York: Cambridge University Press.

Jolliffe, D. & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. *Journal of Adolescence, 29*(4), 589-611.

Kagan, J. (2013). *The Human Spark: The Science of Human Development*. New York, NY: Basic Books.

Little, T.D., Jones, S.M., Heinrich, C.C., & Hawley, P.H. (2003). Disentangling the 'whys' from the 'whats' of aggressive behavior. *International Journal of Behavioral Development, 27*, 122-133.

MacDonald, A., McLafferty, M., Bell, P., McCorkell, L. Walker, I., Smith, V., & Balfour, A. (2014). *Evaluation of the Roots of Empathy Programme by North Lanarkshire Psychological Service* [on-line].

Malti, T., Gummerum, M., Keller, M., & Buchmann, M. (2009). Children's moral motivation, sympathy, and prosocial behavior. *Child Development, 80*(2), 442-460.

Malti, T., Ribeaud, D., & Eisner, M. P. (2011). The effectiveness of two universal preventive interventions in reducing children's externalizing behavior: a cluster randomized controlled trial. *Journal of Clinical Child & Adolescent Psychology, 40*(5), 677-692.

Mayring. P. (2008). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (10. Aufl.). Weinheim: Beltz.

Miller, P. H., Baxter, S. D., Royer, J. A., Hitchcock, D. B., Smith, A. F., Collins, K. L., . . . Vaadi, K. K. (2015). Children's social desirability: Effects of test assessment mode. *Personality and Individual Differences, 83*, 85-90.

Miller, P. A., & Eisenberg, N. (1988). The relation of empathy to aggressive and externalizing/antisocial behavior. *Psychological Bulletin*, 103(3), 324.

Moeller, J. (2015). A word on standardization in longitudinal studies: don't. *Frontiers in Psychology, 6*, 1389.

Morgan, D. L. (1996). *Focus groups as qualitative research*. London: Sage.

Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments. *American Journal of Public Health, 94*(3), 423-432.

Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. New York: Viking.

Przyborski, A. & Wohlrab-Sahr, M. (2010). *Qualitative Sozialforschung. Ein Arbeitsbuch* (3rd ed.). München: Oldenbourg.

Sanchez-Perez, N., Fuentes, L. J., Jolliffe, D., & Gonzalez-Salinas, C. (2014). Assessing children's empathy through a Spanish adaptation of the Basic Empathy Scale: parent's and child's report forms. *Front Psychology, 5*, 1438.

Santos, R. G., Chartier, M. J., Whalen, J. C., Chateau, D., & Boyd, L. (2011). Effectiveness of school-basel violence prevention for children and youth. Cluster randomized controlled field trial of the Roots of Empathy program with replication and three-year follow-up. *Healthcare Quaterly, 14*(Special Issue), 80-91.

Schonert-Reichl, K. A., Smith, V., Zaidman-Zait, A., & Hertzman, C. (2012). Promoting Children's Prosocial Behaviors in School: Impact of the ''Roots of Empathy'' Program on the Social and Emotional Competence of School-Aged Children. *School Mental Health, 4*, 1-21.

Segal, E. A. (2011). Social empathy: A model built on empathy, contextual understanding, and social responsibility that promotes social justice. *Journal of Social Service Research, 37*(3), 266-277.

Shoemaker, A. L. (1980). Construct validity of area specific self-esteem: The Hare Self-Esteem Scale. *Educational and Psychological Measurement, 40*(2), 495-501.

Smetana, J. G., & Killen, M. (2008). Moral cognition, emotions, and neuroscience: An integrative developmental view. *International Journal of Developmental Science, 2*(3), 324-339.

Sutter, M., & Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior, 59*(2), 364-382.

Tarshis, T. P., & Huffman, L. C. (2007). Psychometric properties of the Peer Interactions in Primary School (PIPS) questionnaire. *Journal of Developmental and Behavioral Pediatrics, 28*, 125–132.

Thomas, D. R. & Zumbo, B. D. (2011). Difference scores from the point of view of reliability and repeated measures ANOVA: In defence of difference scores for data analysis. *Educational and Psychological Measurement, 72*(1), 37-43.

Van der Ende, J. (1999). Multiple informants: Multiple views. In H. M. Koot, A. A. M. Crijnen, & R. F. Ferdinand (Eds.), *Child pschiatric epidemiology: Accomplishments and future directions* (pp. 39-52). Assen: Van Gorcum.

van Dulmen, M. H., & Egeland, B. (2011). Analyzing multiple informant data on child and adolescent behavior problems: Predictive validity and comparison of aggregation procedures. *International Journal of Behavioral Development, 35*(1), 84-92.

Van Lange, P. A. (2008). Does empathy trigger only altruistic motivation? How about selflessness or justice? *Emotion, 8*(6), 766-774.

Weinberger, D. A. & Schwartz, G. E. (1990). Distress and restraint as superordinate dimensions of self-reported adjustment: a typological perspective. *Journal of Personality, 58*(2), 381-417.

Wrigley, J., Makara K., & Elliot D. (2015). *Evaluation of Roots of Empathy in Scotland 2014-2015. Final Report for Action for Children*. Unpublished document.

Zhou, Q., Valiente, C., & Eisenberg, N. (2003). Empathy and its measurement. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook of models and measures* (pp. 269-284). Washington, DC: American Psychological Association.